



(86) Date de dépôt PCT/PCT Filing Date: 2014/03/13
(87) Date publication PCT/PCT Publication Date: 2014/09/25
(45) Date de délivrance/Issue Date: 2023/06/27
(85) Entrée phase nationale/National Entry: 2015/09/09
(86) N° demande PCT/PCT Application No.: US 2014/026476
(87) N° publication PCT/PCT Publication No.: 2014/151805
(30) Priorité/Priority: 2013/03/15 (US61/799,336)

(51) Cl.Int./Int.Cl. *C12N 1/16* (2006.01),
C12N 9/42 (2006.01), *C12P 7/10* (2006.01)
(72) Inventeurs/Inventors:
MCBRIDE, JOHN, US;
WISWALL, ERIN, US
(73) Propriétaire/Owner:
LALLEMAND HUNGARY LIQUIDITY MANAGEMENT
LLC, HU
(74) Agent: NORTON ROSE FULBRIGHT CANADA
LLP/S.E.N.C.R.L., S.R.L.

(54) Titre : EXPRESSION DE BETA-GLUCOSIDASES POUR L'HYDROLYSE DE LA LIGNOCELLULOSE ET OLIGOMERES ASSOCIES
(54) Title: EXPRESSION OF BETA-GLUCOSIDASES FOR HYDROLYSIS OF LIGNOCELLULOSE AND ASSOCIATED OLIGOMERS

(57) **Abrégé/Abstract:**

The present invention provides for heterologous expression of beta-glucosidase (BGL) polypeptides encoded by *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* in host cells, such as the yeast *Saccharomyces cerevisiae*. The expression in such host cells of the corresponding genes, and variants and combinations thereof, result in improved specific activity of the expressed BGL. Thus, such genes and expression systems are useful for efficient and cost-effective consolidated bioprocessing systems.

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

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
25 September 2014 (25.09.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/151805 A3

(51) International Patent Classification:

C12N 1/16 (2006.01) *C12P 7/10* (2006.01)
C12N 9/42 (2006.01)

(21) International Application Number:

PCT/US2014/026476

(22) International Filing Date:

13 March 2014 (13.03.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/799,336 15 March 2013 (15.03.2013) US

(71) Applicant: **LALLEMAND HUNGARY LIQUIDITY MANAGEMENT LLC** [HU/HU]; Wesselenyi utca 16, 3rd Floor, Budapest 1077 (HU).

(72) Inventors: **MCBRIDE, John**; 174 Dartmouth College Highway, Lyme, New Hampshire 03768 (US).
WISWALL, Erin; 770 Oak Ridge Road, P.O. Box 219, Danbury, New Hampshire 03230 (US).

(74) Agents: **CAHILL, Ronald, E.** et al.; Nutter McClennen & Fish LLP, 155 Seaport Boulevard, 7th Floor, Boston, Massachusetts 02210 (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, IR, IS, JP, KE, KG, 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, SA, 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, KM, 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))

(88) Date of publication of the international search report:

8 January 2015

(54) Title: EXPRESSION OF BETA-GLUCOSIDASES FOR HYDROLYSIS OF LIGNOCELLULOSE AND ASSOCIATED OLIGOMERS

(57) Abstract: The present invention provides for heterologous expression of beta-glucosidase (BGL) polypeptides encoded by *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* in host cells, such as the yeast *Saccharomyces cerevisiae*. The expression in such host cells of the corresponding genes, and variants and combinations thereof, result in improved specific activity of the expressed BGL. Thus, such genes and expression systems are useful for efficient and cost-effective consolidated bioprocessing systems.



WO 2014/151805 A3

EXPRESSION OF BETA-GLUCOSIDASES FOR HYDROLYSIS OF LIGNOCELLULOSE AND ASSOCIATED OLIGOMERS

BACKGROUND OF THE INVENTION

[0001] Lignocellulosic biomass is widely recognized as a promising source of raw material for production of renewable fuels and chemicals. 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 these materials to conversion into useful fuels. Lignocellulosic biomass contains carbohydrate fractions (*e.g.*, cellulose and hemicellulose) that can be converted into ethanol. In order to convert these fractions, the cellulose and hemicellulose must ultimately be converted or hydrolyzed into monosaccharides; it is the hydrolysis that has historically proven to be problematic.

[0002] Biologically mediated processes are promising for energy conversion, in particular, for the conversion of lignocellulosic biomass into fuels. 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. 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.

[0003] Bakers' yeast (*Saccharomyces cerevisiae* or *S. cerevisiae*) remains the preferred microorganism for the production of ethanol (Van Zyl *et al.*, *Adv. Biochem. Eng. Biotechnol.* 108:205-235, 2007). Attributes in favor of this microbe are (i) high productivity at close to theoretical yields (0.51 gram of ethanol produced / gram glucose used), (ii) high osmo- and ethanol-tolerance, (iii) natural robustness in industrial processes, (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 hydrolyzates

resulting from biomass pretreatment. The major shortcoming of *S. cerevisiae* is its inability to utilize complex polysaccharides such as cellulose, or its break-down products, such as cellobiose and cellodextrins. One strategy for developing CBP-enabling microorganisms such as *S. cerevisiae* is by engineering them to express a heterologous cellulase and/or a hemicellulase system.

[0004] Three major types of enzymatic activities are required for native cellulose degradation. One type is endoglucanases (1,4- β -D-glucan 4-glucanohydrolases; Enzyme Commission (EC) 3.2.1.4). Endoglucanases (Eg or EG) cut at random in the cellulose polysaccharide chain of amorphous cellulose, generating oligosaccharides of varying lengths and consequently new chain ends. Another type is exoglucanases. Exoglucanases include cellodextrinases (1,4- β -D-glucan glucanohydrolases; EC 3.2.1.74) and cellobiohydrolases (1,4- β -D-glucan cellobiohydrolases; EC 3.2.1.91). Exoglucanases act in a processive manner on the reducing or non-reducing ends of cellulose polysaccharide chains, liberating either glucose (glucanohydrolases) or cellobiose (cellobiohydrolase) as major products. Exoglucanases can also act on microcrystalline cellulose, presumably peeling cellulose chains from the microcrystalline structure. Classically, exoglucanases such as the cellobiohydrolases (CBHs) possess tunnel-like active sites, which can only accept a substrate chain via its terminal regions. These exo-acting CBH enzymes act by threading the cellulose chain through the tunnel, where successive cellobiose units are removed in a sequential manner. Sequential hydrolysis of a cellulose chain is termed "processivity."

[0005] Yet another type is beta-glucosidases (beta glucoside glucohydrolases, β -glucosidases or BGLs; EC 3.2.1.21). BGLs play an important role in the hydrolysis of materials containing cellulose or soluble oligomers of glucose. There have been reports of the role and importance of BGLs during hydrolysis (*see, e.g., Viikari et al., Adv. Biochem. Eng. Biotechnol.*, 108:121-145, 2007; and Bhatia *et al., Crit. Rev. Biotechnol.*, 22:375-407, 2002). These enzymes typically act on soluble oligomers of glucose which are linked via beta 1-4 type bonds, including dimers (cellobiose) where they usually have highest activity, as well as longer chain oligomers where they are typically less active. Examples of BGL domains have been described and include, for example, a glycosyl hydrolase family 3 n-terminal domain, a glycosyl hydrolase family 3 c-terminal domain, and a fibronectin type III like domain.

[0006] Structurally, cellulases generally consist of a catalytic domain joined to a cellulose-binding module (CBM) via a linker region that is rich in proline and/or hydroxy-amino

acids. In type I exoglucanases, the CBM domain is found at the C-terminal extremity of these enzyme (this short domain forms a hairpin loop structure stabilized by 2 disulfide bridges). In type 2 CBHs, the CBM is found at the N-terminus. In some cases, however, cellulases do not contain a CBM, and only contain a catalytic domain. Examples of such CBM-lacking cellulases include CBHs from *Humicola grisea*, *Phanerochaete chrysosporium* and *Aspergillus niger*. Grassick *et al.*, *Eur. J. Biochem.*, 271:4495-4506, 2004.

[0007] With the aid of recombinant DNA technology, several of these heterologous cellulases from bacterial and fungal sources have been transferred to *S. cerevisiae*, enabling the degradation of cellulosic derivatives (Van Rensburg *et al.*, *Yeast*, 14:67-76, 1998), or growth on cellobiose (Van Rooyen *et al.*, *J. Biotech.*, 120:284-295, 2005; and McBride *et al.*, *Enzyme Microb. Technol.* 37:93-101, 2005).

[0008] Related work was described by Fujita *et al.*, (*Appl. Environ. Microbiol.*, 70:1207-1212, 2004) where cellulases immobilized on the yeast cell surface had significant limitations. First, Fujita *et al.* were unable to achieve fermentation of amorphous cellulose using yeast expressing only recombinant BglI and EgII. A second limitation of the Fujita *et al.* approach was that cells had to be pre-grown to high cell density on standard carbon sources before the cells were useful for ethanol production using amorphous cellulose (*e.g.*, Fujita *et al.* uses high biomass loadings of ~15 g/L to accomplish ethanol production).

[0009] As noted above, ethanol producing yeast such as *S. cerevisiae* require addition of external cellulases when cultivated on cellulosic substrates, such as pre-treated wood, because this yeast does not produce endogenous cellulases. Expression of fungal cellulases such as *Trichoderma reesei* (*T. reesei*) Cbh1 and Cbh2 in yeast *S. cerevisiae* have been shown to be functional. Den Haan *et al.*, *Enzyme and Microbial Technology*, 40:1291-1299, 2007. However, current levels of expression and specific activity of cellulases heterologously expressed in yeast are still not sufficient to enable growth and ethanol production by yeast on cellulosic substrates without externally added enzymes. While studies have shown that perhaps certain cellulases, such as *T. reesei* Cbh1, have some activity when heterologously expressed, there remains a significant need for improvement in the specific activity of heterologously expressed cellulases in order to attain the goal of achieving a CBP system capable of efficiently and cost-effectively converting cellulosic substrates to ethanol.

[0010] Currently, there is no reliable way to predict which cellulases will be efficiently expressed in heterologous organisms. For example, despite the fact that *T. reesei* Cbh1 and *T.*

emersonii Cbh1 are both endogenously expressed at high levels, heterologous expression of these proteins in yeast yielded disparate results. Also, *Talaromyces emersonii* (*T. emersonii*) Cbh1 expression in yeast was significantly greater in yeast than *T. reesei* Cbh1 under similar conditions. See Int'l Pub. No. WO 2009/138877. Efficient expression may depend, for example, on chaperone proteins that differ in the heterologous organisms and in the cellulase's native organism. Furthermore, even cellulases which are expressed at high levels may not be particularly active in a heterologous organism. For example, a cellulase may be subject to different post-translational modifications in the heterologous host organism than in the native organism from which the cellulase is derived. Protein folding and secretion can also be a barrier to heterologous cellulase expression.

[0011] Therefore, in order to address the limitations of heterologous cellulase expression in CBP systems, the present invention provides the expression of several BGLs in host cells, such as the yeast *S. cerevisiae*. The expression level and secreted activity level of the BGLs was characterized. In addition, the BGLs were purified and their specific activity on hardwood derived pretreated solids (C6 solids) and hardwood derived hemicellulose liquor (C5 liquor) was determined. The corresponding BGL genes, or variants and combinations thereof, in such host cells were well expressed and resulted in improved specific activity of the expressed BGLs. Also, the combination of purified BGLs with one or more other cellulases, or host cells expressing the BGLs and one or more other cellulases, also resulted in improved specific activity of the expressed BGLs. Thus, such genes and expression systems are useful for efficient and cost-effective CBP systems.

BRIEF SUMMARY OF THE INVENTION

[0012] The present invention provides for the heterologous expression of *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* beta-glucosidases (BGLs), or fragments thereof, in host cells. The host cell can comprise one or more polynucleotides encoding a BGL that is (i) at least about 90% identical to any one of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27, 30 or 31-40, (ii) at least about 95% identical to any one of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27, 30 or 31-40, or (iii) identical to any one of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27, 30 or 31-40. The host cell can comprise one or more polynucleotides encoding a BGL having (i) an amino acid

sequence at least about 90% identical to any one of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28, (ii) an amino acid sequence at least about 90% identical to any one of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28 without the signal peptide sequence, (iii) an amino acid sequence at least about 95% identical to any one of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28, or (iv) an amino acid sequence at least about 95% identical to any one of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28 without the signal peptide sequence.

[0013] In some embodiments of the invention, the fragment of the BGL can be a BGL signal peptide. The signal peptide can comprise an amino acid sequence that is (i) at least about 90% identical to any one of SEQ ID NOs:2, 5, 11, 14, 17, 20, 23, 26 or 29, (ii) at least about 95% identical to any one of SEQ ID NOs:2, 5, 11, 14, 17, 20, 23, 26 or 29, or (iii) identical to any one of SEQ ID NOs:2, 5, 11, 14, 17, 20, 23, 26 or 29.

[0014] In some embodiments of the invention, the host cell further comprises one or more additional polynucleotides encoding a heterologous cellulase. The heterologous cellulase can be a xylanase, xylosidase, acetylxylanesterase (AXE), endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase, or cellobiohydrolase (CBH). The endoglucanase can be *A. fumigatus* endoglucanase I, *N. fischeri* endoglucanase III, *T. reesei* endoglucanase I, or *C. formosanus* endoglucanase I. The CBH can be CBH1 or CBH2. The CBH can also be *T. emersonii* cellobiohydrolase I, *C. lucknowense* cellobiohydrolase IIb, or *T. reesei* cellobiohydrolase II. The host cell can further comprise a polynucleotide encoding *S. fibuligera* BGL. The host cell can also further comprise one or more polynucleotides encoding *T. emersonii* CBH1, *T. reesei* CBD, *C. lucknowense* CBH2, *A. fumigatus* EG1, *N. fischeri* EG3, *S. fibuligera* BGL, or *A. niger* xylanase. The host cell can further comprise one or more polynucleotides encoding *A. niger* xylanase, *P.t.r.* xylosidase, *N. fischeri* AXE, *A. fumigatus* EG1, *T. reesei* AGL1, *T. reesei* beta-mannanase, *A. fumigatus* alpha-glucuronidase (FC110), *A. fumigatus* acetyl esterase (FC136), *N. fischeri* beta-mannosidase (FC124), or *S. fibuligera* BGL.

[0015] In some embodiments of the invention, the host cell can saccharify and/or ferment crystalline cellulose. In other embodiments, the host cell can hydrolyze hardwood solids or C5 liquor derived from hardwoods.

[0016] In some embodiments of the invention, the yeast is selected from *Saccharomyces cerevisiae*, *Saccharomyces pastorianus*, *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*,

Pichia stipitis, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus*, *Schizosaccharomyces pombe*, *Schwanniomyces occidentalis*, or derivatives thereof. In some embodiments, the yeast is *Saccharomyces cerevisiae*.

[0017] Other embodiments of the invention are directed to a BGL peptide isolated from a host cell of the invention, or a purified BGL peptide isolated from a host cell of the invention. Other embodiments of the invention include a co-culture comprising (i) a host cell of the invention and (ii) a second host cell comprising one or more polynucleotides encoding a xylanase, xylosidase, AXE, endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase or CBH. In other embodiments, the invention is directed to a composition comprising (i) a peptide or purified peptide of the invention and (ii) a host cell comprising one or more polynucleotides encoding a xylanase, xylosidase, AXE, endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase or CBH.

[0018] The present invention also provides a method for hydrolyzing a cellulosic substrate, comprising contacting the cellulosic substrate with a host cell, co-culture, composition, peptide or purified peptide of the invention. The cellulosic substrate can comprise a lignocellulosic biomass. The lignocellulosic biomass can be grass, switch grass, cord grass, rye grass, reed canary grass, miscanthus, sugar-processing residues, sugarcane bagasse, agricultural wastes, rice straw, rice hulls, barley straw, corn cobs, cereal straw, wheat straw, canola straw, oat straw, oat hulls, corn fiber, stover, soybean stover, corn stover, forestry wastes, recycled wood pulp fiber, paper sludge, sawdust, hardwood, softwood, or combinations thereof. The cellulosic substrate can be hydrolyzed to xylose, glucose, mannose, galactose, arabinose, or combinations thereof. In some embodiments, the cellulose substrate is hydrolyzed to xylose, glucose, mannose, galactose or arabinose at a rate at least about 10% greater than the rate of a host cell comprising a polynucleotide encoding a BGL from *S. fibuligera*. In some embodiments of the method, the BGL is present in an amount of about 0.2 mg or less per gram of xylose.

[0019] The present invention also provides a method of fermenting cellulose, comprising culturing a host cell, co-culture, composition, peptide or purified peptide of the invention in medium that contains crystalline cellulose under suitable conditions for a period sufficient to allow saccharification and fermentation of the cellulose. In some embodiments, the host cell produces ethanol.

[0020] The present invention also provides yeast strains M4860, M4861, M4862, M4863, M4864, and M4865, and expression vectors pMU3557, pMU3558, pMU3559, pMU3560, pMU3561, pMU3562, pMU3563, pMU3564, pMU3565, and pMU3566.

[0021] The present invention also provides a fermentation product produced by a host cell, co-culture or yeast strain of the invention. The fermentation product can be ethanol.

[0021a] In yet another aspect, the present invention provides a recombinant yeast host cell comprising a polynucleotide encoding a heterologous beta-glucosidase (BGL) having (i) an amino acid sequence at least 90% or at least 95% identical to SEQ ID NO: 7; or (ii) an amino acid sequence at least 90% or at least 95% identical to a fragment of SEQ ID NO: 7 without a signal peptide, wherein said polynucleotide is codon-optimized for expression in a yeast host cell.

[0021b] In yet another aspect, the present invention provides a recombinant yeast host cell comprising a polynucleotide encoding a heterologous beta-glucosidase (BGL) having (i) an amino acid sequence at least 90% or at least 95% identical to SEQ ID NO: 7; or (ii) an amino acid sequence at least 90% or at least 95% identical to a fragment of SEQ ID NO: 7 without a signal peptide, wherein said polynucleotide is codon-optimized for expression in the yeast host cell, wherein the codon adaption index (CAI) of the codon-optimized polynucleotide corresponds to from 0.8 to 1.0.

BRIEF DESCRIPTION OF THE DRAWINGS/FIGURES

[0022] FIG. 1 depicts a plasmid map of pMU3557.

[0023] FIG. 2 depicts a plasmid map of pMU3558.

[0024] FIG. 3 depicts a plasmid map of pMU3559.

[0025] FIG. 4 depicts a plasmid map of pMU3560.

[0026] FIG. 5 depicts a plasmid map of pMU3561.

[0027] FIG. 6 depicts a plasmid map of pMU3562.

[0028] FIG. 7 depicts a plasmid map of pMU3563.

[0029] FIG. 8 depicts a plasmid map of pMU3564.

[0030] FIG. 9 depicts a plasmid map of pMU3565.

[0031] FIG. 10 depicts a plasmid map of pMU3566.

[0032] FIG. 11 depicts a beta-glucosidase activity assay with cellobiose of the transformants described in Example 1.

[0033] FIGs. 12A-12B depict SDS-PAGE and western blot analysis of the supernatants from beta-glucosidase (BGL) producing strains. The left-hand panels are SDS-PAGE gel results. The right-hand panels are western blot results.

[0034] FIG. 13 depicts a comparison of several BGL enzymes for activity against cellobiose at several protein loadings. The enzymes are identified by the two letter abbreviation of the source organism in the figure legend.

[0035] FIG. 14 depicts a comparison of several BGL enzymes for activity against cellobiose at a 5 ug/mL protein loading. The enzymes are identified by the two letter abbreviation of the source organism in the figure legend.

[0036] FIG. 15 depicts a comparison of several BGL enzymes for their impact on pretreated hardwood hydrolysis in a low concentration (2% total solids). "Big 6" refers to yeast made and purified cellulases, *T. emersonii* CBH1 with the *T. reesei* CBD, *C. lucknowense* CBH2, *A. fumigatus* EG1, *N. fischeri* EG3, *S. fibuligera* BGL, and *A. niger* xylanase. 2 mg/g of total

solids of this mixture along with 4 mg/g of a commercial enzyme preparation termed "flashzyme" was loaded in the assay, and additional purified BGL was added in small amounts (0.1 mg enzyme protein per gram of total solids) in addition to a commercial enzyme preparation which was loaded at a typical loading of 4 mg enzyme protein per gram of total solids. Released sugars were measured by HPLC.

[0037] FIG. 16 depicts xylose (combination of xylose, galactose and mannose) release from pretreated hardwood derived C5 liquor during enzymatic assay using purified BGL enzymes. BGL was added in small amounts (0.2 mg enzyme protein per gram xylose) in addition to other yeast-made purified enzymes (all added at 0.2 mg/g xylose except xld = 0.6 mg/g xylose). "Original set" represents the following set of genes: *A. niger* xylanase, *P.t.r.* xylosidase, *N. fischeri* AXE, *A. fumigatus* EG1, *T. reesei* AGL1, *T. reesei* beta-mannanase, *A. fumigatus* alpha-glucuronidase (FC110), *A. fumigatus* acetyl esterase (FC136), *N. fischeri* beta-mannosidase (FC124), and *S. fibuligera* BGL. "Original set + Ao BGL" represents the original set, except that the *S. fibuligera* BGL was not included, and the *A. oryzae* BGL was used in its place. Released sugars were measured by HPLC.

[0038] FIG. 17 depicts glucose release from pretreated hardwood derived C5 liquor during enzymatic assay using purified BGL enzymes. BGL was added in small amounts (0.2 mg enzyme protein per gram xylose) in addition to other yeast-made purified enzymes (all added at 0.2 mg/g xylose except xld = 0.6 mg/g xylose). BGL was added in small amounts (0.1 mg enzyme protein per gram of total solids) in addition to a commercial enzyme preparation which was loaded at a typical loading of 4 mg enzyme protein per gram of total solids. Released sugars were measured by HPLC.

[0039] FIG. 18 depicts sugar release from pretreated hardwood derived C5 liquor during enzymatic assay using purified BGL enzymes. BGL was added in small amounts (0.2 mg enzyme protein per gram xylose) in addition to other yeast-made purified enzymes (all added at 0.2 mg/g xylose except xld = 0.6 g/g xylose). BGL was added in small amounts (0.1 mg enzyme protein per gram of total solids) in addition to a commercial enzyme preparation which was loaded at a typical loading of 4 mg enzyme protein per gram of total solids. Released sugars were measured using HPLC using the BioRad Aminex 87P column to separate xylose, galactose, and mannose.

[0040] FIG. 19 depicts sugar release from pretreated hardwood derived C5 liquor during enzymatic assay using purified BGL enzymes. BGL was added in small amounts (0.2 mg

enzyme protein per gram xylose) in addition to other yeast-made purified enzymes (all added at 0.2 mg/g xylose except xld = 0.6 mg/g xylose). Also, BGL was added in small amounts (0.1 mg enzyme protein per gram of total solids) in addition to a commercial enzyme preparation which was loaded at a typical loading of 4 mg enzyme protein per gram of total solids. Sets with more than one BGL as indicated in the figure legend were created by adding an additional 0.2 mg/g xylose protein loading for the particular BGLs noted to the reaction. Released sugars were measured using HPLC using the BioRad Aminex 87H column to separate xylose, galactose, and mannose.

DETAILED DESCRIPTION OF THE INVENTION

[0041] The present invention relates to, *inter alia*, the heterologous expression of BGL genes from *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, and *Phytophthora infestans* in host cells, including yeast, e.g., *Saccharomyces cerevisiae*. The present invention provides important tools to enable growth of yeast on cellulosic substrates for production of products such as ethanol.

Definitions

[0042] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. In case of conflict, the present application including the definitions will control. Also, unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular.

[0043] A used herein, the terms "comprises," "comprising," "includes," "including," "has," "having," "contains," or "containing," or any other variation thereof, will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers. For example, a composition, a mixture, a process, a method, an article, or an apparatus that comprises a list of elements not expressly listed or inherent to only those elements but can include other elements not expressly listed or inherent to such composition, mixture, process, method, article, or apparatus. Further, unless expressly stated to the contrary, "or" refers to an inclusive or and not to exclusive or. For example, a condition A or

B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

[0044] Also, the indefinite articles "a" and "an" preceding an element or component of the invention are intended to be nonrestrictive regarding the number of instances, *i.e.*, occurrences of the element or component. Therefore, "a" or "an" should be read to include one or at least one, and the singular word form of the element or component also includes the plural unless the number is obviously meant to be singular.

[0045] The term "invention" or "present invention" as used herein is a non-limiting term and is not intended to refer to any single embodiment of the particular invention but encompasses all possible embodiments as described in the application.

[0046] As used herein, the term "about" modifying a quantity or amount related to the invention refers to variation in the numerical quantity that can occur, for example, through typical measuring and liquid handling procedures used for making concentrates or solutions in the real world; through inadvertent error in these procedures; through differences in the manufacture, source, or purity of the ingredients employed to make the compositions or to carry out the methods; and the like. The term "about" also encompasses amounts that differ due to different equilibrium conditions for a composition resulting from a particular initial mixture. Whether or not modified by the term "about," the claims include equivalents to the quantities. In one embodiment, the term "about" means within 10% of the reported numerical value, alternatively within 5% of the reported numerical value.

[0047] A "vector," *e.g.*, a "plasmid" or "YAC" (yeast artificial chromosome) refers to an extrachromosomal element often carrying one or more genes that are not part of the central metabolism of the cell. They can be in the form of a circular double-stranded DNA molecule. Such elements can be autonomously replicating sequences, genome integrating sequences, or phage sequences. Such elements can be linear, circular, or supercoiled and can be single- or double-stranded. They can also be DNA or RNA, derived from any source. They can include a number of nucleotide sequences which have been joined or recombined into a unique construction which is capable of introducing a promoter fragment and DNA sequence for a selected gene product along with appropriate 3' untranslated sequence into a cell. The plasmids or vectors of the present invention can be stable and self-replicating. The plasmids or vectors of the present invention can also be suicide vectors, or vectors that cannot replicate in the host cell.

Such vectors are useful for forcing insertion of the nucleotide sequence into the host chromosome.

[0048] An "expression vector" is a vector that is capable of directing the expression of at least one polypeptide encoded by a polynucleotide sequence of the vector.

[0049] The term "heterologous" as used herein refers to an element of a vector, plasmid or host cell that is derived from a source other than the endogenous source. Thus, for example, a heterologous sequence could be a sequence that is derived from a different gene or plasmid from the same host, from a different strain of host cell, or from an organism of a different taxonomic group (*e.g.*, different kingdom, phylum, class, order, family genus, or species, or any subgroup within one of these classifications). The term "heterologous" is also used synonymously herein with the term "exogenous."

[0050] The term "domain" as used herein refers to a part of a molecule or structure that shares common physical or chemical features, for example hydrophobic, polar, globular, helical domains or properties, *e.g.*, a DNA binding domain or an ATP binding domain. Domains can be identified by their homology to conserved structural or functional motifs. Examples of domains of BGL have been described and include, for example, a glycosyl hydrolase family 3 n-terminal domain, a glycosyl hydrolase family 3 c-terminal domain, and a fibronectin type III like domain.

[0051] A "nucleic acid," "polynucleotide," or "nucleic acid molecule" is a polymeric compound comprised of covalently linked subunits called nucleotides. Nucleic acid includes polyribonucleic acid (RNA) and polydeoxyribonucleic acid (DNA), both of which can be single-stranded or double-stranded. DNA includes cDNA, genomic DNA, synthetic DNA, and semi-synthetic DNA.

[0052] An "isolated nucleic acid molecule" or "isolated nucleic acid fragment" refers to the phosphate ester polymeric form of ribonucleosides (adenosine, guanosine, uridine or cytidine; "RNA molecules") or deoxyribonucleosides (deoxyadenosine, deoxyguanosine, deoxythymidine, or deoxycytidine; "DNA molecules"), or any phosphoester analogs thereof, such as phosphorothioates and thioesters, in either single stranded form, or a double-stranded helix. Double stranded DNA-DNA, DNA-RNA and RNA-RNA helices are possible. The term nucleic acid molecule, and in particular DNA or RNA molecule, refers only to the primary and secondary structure of the molecule, and does not limit it to any particular tertiary forms. Thus, this term includes double-stranded DNA found, *inter alia*, in linear or circular DNA molecules (*e.g.*, restriction fragments), plasmids, and chromosomes. In discussing the structure of particular

double-stranded DNA molecules, sequences are generally described herein according to the normal convention of giving only the sequence in the 5' to 3' direction along the non-transcribed strand of DNA (*i.e.*, the strand having a sequence homologous to the mRNA).

[0053] A "gene" refers to an assembly of nucleotides that encode a polypeptide, and includes cDNA and genomic DNA nucleic acids. "Gene" also refers to a nucleic acid fragment that expresses a specific protein, including intervening sequences (introns) between individual coding segments (exons), as well as regulatory sequences preceding (5' non-coding sequences) and following (3' non-coding sequences) the coding sequence. "Native gene" refers to a gene as found in nature with its own regulatory sequences.

[0054] A nucleic acid molecule is "hybridizable" to another nucleic acid molecule, such as a cDNA, genomic DNA, or RNA, when a single stranded form of the nucleic acid molecule can anneal to the other nucleic acid molecule under the appropriate conditions of temperature and solution ionic strength. Hybridization and washing conditions are well known and exemplified, *e.g.*, in Sambrook *et al.*, MOLECULAR CLONING: A LABORATORY MANUAL, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989, particularly Chapter 11 and Table 11.1 therein (hereinafter "Maniatis"). The conditions of temperature and ionic strength determine the "stringency" of the hybridization. Stringency conditions can be adjusted to screen for moderately similar fragments, such as homologous sequences from distantly related organisms, to highly similar fragments, such as genes that duplicate functional enzymes from closely related organisms. Post-hybridization washes determine stringency conditions. One set of conditions uses a series of washes starting with 6X SSC, 0.5% SDS at room temperature for 15 min, then repeated with 2X SSC, 0.5% SDS at 45°C for 30 min, and then repeated twice with 0.2X SSC, 0.5% SDS at 50°C for 30 min. For more stringent conditions, washes are performed at higher temperatures in which the washes are identical to those above except for the temperature of the final two 30 min washes in 0.2X SSC, 0.5% SDS are increased to 60°C. Another set of highly stringent conditions uses two final washes in 0.1X SSC, 0.1% SDS at 65°C. An additional set of highly stringent conditions are defined by hybridization at 0.1X SSC, 0.1% SDS, 65°C and washed with 2X SSC, 0.1% SDS followed by 0.1X SSC, 0.1% SDS.

[0055] Hybridization requires that the two nucleic acids contain complementary sequences, although depending on the stringency of the hybridization, mismatches between bases are possible. The appropriate stringency for hybridizing nucleic acids depends on the length of

the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of similarity or homology between two nucleotide sequences, the greater the value of T_m for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA. For hybrids of greater than 100 nucleotides in length, equations for calculating T_m have been derived (*see, e.g.*, Maniatis at 9.50-9.51). For hybridizations with shorter nucleic acids, *i.e.*, oligonucleotides, the position of mismatches becomes more important, and the length of the oligonucleotide determines its specificity (*see, e.g.*, Maniatis at 11.7-11.8). In one embodiment the length for a hybridizable nucleic acid is at least about 10 nucleotides. A minimum length for a hybridizable nucleic acid can also be at least about 15 nucleotides, at least about 20 nucleotides, or at least 30 nucleotides. Furthermore, the skilled artisan will recognize that the temperature and wash solution salt concentration can be adjusted as necessary according to factors such as length of the probe.

[0056] The term "percent identity", as known in the art, is a relationship between two or more polypeptide sequences or two or more polynucleotide sequences, as determined by comparing the sequences. In the art, "identity" also means the degree of sequence relatedness between polypeptide or polynucleotide sequences, as determined by the match between strings of such sequences.

[0057] 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 can 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 can be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence can be inserted into the reference sequence.

[0058] As a practical matter, whether any particular nucleic acid molecule or polypeptide is at least about 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.*, 6:237-245, 1990. 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.

[0059] 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.

[0060] 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.

[0061] 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.

[0062] Suitable nucleic acid sequences or fragments thereof (isolated polynucleotides of the present invention) encode polypeptides that are at least about 70% to 75% identical to the amino acid sequences reported herein, at least about 80%, 85%, or 90% identical to the amino acid sequences reported herein, or at least about 95%, 96%, 97%, 98%, 99%, or 100% identical to the amino acid sequences reported herein. Suitable nucleic acid fragments are at least about 70%, 75%, or 80% identical to the nucleic acid sequences reported herein, at least about 80%, 85%, or 90% identical to the nucleic acid sequences reported herein, or at least about 95%, 96%, 97%, 98%, 99%, or 100% identical to the nucleic acid sequences reported herein. Suitable nucleic acid fragments not only have the above identities/similarities but typically encode a polypeptide having at least 50 amino acids, at least 100 amino acids, at least 150 amino acids, at least 200 amino acids, at least 250 amino acids, at least 300 amino acids, or at least 350 amino acids.

[0063] The term "probe" refers to a single-stranded nucleic acid molecule that can base pair with a complementary single stranded target nucleic acid to form a double-stranded molecule.

[0064] The term "complementary" is used to describe the relationship between nucleotide bases that are capable to hybridizing to one another. For example, with respect to DNA, adenosine is complementary to thymine and cytosine is complementary to guanine. Accordingly, the instant invention also includes isolated nucleic acid fragments that are complementary to the complete sequences as reported in the accompanying Sequence Listing as well as those substantially similar nucleic acid sequences.

[0065] As used herein, the term "oligonucleotide" refers to a nucleic acid, generally of about 18 nucleotides, that is hybridizable to a genomic DNA molecule, a cDNA molecule, or an mRNA molecule. Oligonucleotides can be labeled, *e.g.*, with ³²P-nucleotides or nucleotides to which a label, such as biotin, has been covalently conjugated. An oligonucleotide can be used as a probe to detect the presence of a nucleic acid according to the invention. Similarly,

oligonucleotides (one or both of which can be labeled) can be used as PCR primers, either for cloning full length or a fragment of a nucleic acid of the invention, or to detect the presence of nucleic acids according to the invention. Generally, oligonucleotides are prepared synthetically, for example, on a nucleic acid synthesizer. Accordingly, oligonucleotides can be prepared with non-naturally occurring phosphoester analog bonds, such as thioester bonds, etc.

[0066] A DNA or RNA "coding region" is a DNA or RNA molecule which is transcribed and/or translated into a polypeptide in a cell *in vitro* or *in vivo* when placed under the control of appropriate regulatory sequences. "Suitable regulatory regions" refer to nucleic acid regions located upstream (5' non-coding sequences), within, or downstream (3' non-coding sequences) of a coding region, and which influence the transcription, RNA processing or stability, or translation of the associated coding region. Regulatory regions can include promoters, translation leader sequences, RNA processing site, effector binding site and stem-loop structure. The boundaries of the coding region are determined by a start codon at the 5' (amino) terminus and a translation stop codon at the 3' (carboxyl) terminus. A coding region can include, but is not limited to, prokaryotic regions, cDNA from mRNA, genomic DNA molecules, synthetic DNA molecules, or RNA molecules. If the coding region is intended for expression in a eukaryotic cell, a polyadenylation signal and transcription termination sequence will usually be located 3' to the coding region.

[0067] "Open reading frame" is abbreviated ORF and means a length of nucleic acid, either DNA, cDNA or RNA, that comprises a translation start signal or initiation codon, such as an ATG or AUG, and a termination codon and can be potentially translated into a polypeptide sequence.

[0068] "Promoter" refers to a DNA fragment capable of controlling the expression of a coding sequence or functional RNA. In general, a coding region is located 3' to a promoter. Promoters can be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, or even comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters can direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental or physiological conditions. Promoters which cause a gene to be expressed in most cell types at most times are commonly referred to as "constitutive promoters". It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of different lengths can have

identical promoter activity. A promoter is generally bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter will be found a transcription initiation site (conveniently defined for example, by mapping with nuclease S1), as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

[0069] A coding region is "under the control" of transcriptional and translational control elements in a cell when RNA polymerase transcribes the coding region into mRNA, which is then trans-RNA spliced (if the coding region contains introns) and translated into the protein encoded by the coding region.

[0070] "Transcriptional and translational control regions" are DNA regulatory regions, such as promoters, enhancers, terminators, and the like, that provide for the expression of a coding region in a host cell. In eukaryotic cells, polyadenylation signals are control regions.

[0071] The term "operably associated" refers to the association of nucleic acid sequences on a single nucleic acid fragment so that the function of one is affected by the other. For example, a promoter is operably associated with a coding region when it is capable of affecting the expression of that coding region (*i.e.*, that the coding region is under the transcriptional control of the promoter). Coding regions can be operably associated to regulatory regions in sense or antisense orientation.

[0072] The term "expression," as used herein, refers to the transcription and stable accumulation of sense (mRNA) or antisense RNA derived from the nucleic acid fragment of the invention. Expression can also refer to translation of mRNA into a polypeptide.

Polynucleotides of the Invention

[0073] The present invention provides for the use of BGL polynucleotide sequences from *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans*. Nucleic acid sequences for BGL from *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* are available in GenBank and examples of such sequences are shown in Example 1.

[0074] The present invention also provides for the use of an isolated polynucleotide comprising a nucleic acid at least about 70%, at least about 75%, 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, or any range of values thereof, to any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27, or 30, or a fragment, variant, derivative, or codon-optimized version thereof. The present invention also provides for the use of an isolated polynucleotide comprising a nucleic acid having from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identity to any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27, or 30, or a fragment, variant, derivative, or codon-optimized version thereof.

[0075] In certain aspects, the present invention relates to a polynucleotide comprising a nucleic acid encoding a functional and/or structural domain of a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL. The present invention also encompasses an isolated polynucleotide comprising a nucleic acid that is at least about 70%, at least about 75%, 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, or any range of values thereof, to a nucleic acid encoding a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL domain. The present invention also encompasses an isolated polynucleotide comprising a nucleic acid having from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identity to a nucleic acid encoding a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL domain. Examples of BGL domains have been described and include, for example, a glycosyl hydrolase family 3 n-terminal domain, a glycosyl hydrolase family 3 c-terminal domain, and a fibronectin type III like domain.

[0076] The present invention also encompasses variants of BGL genes. Variants can 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, *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polynucleotide variants can be produced for a variety of reasons, e.g., to optimize codon expression for a particular host (e.g., change codons in the BGL mRNA to those preferred by a host such as the yeast *Saccharomyces cerevisiae*). Codon-optimized polynucleotides of the present invention are discussed further herein.

[0077] The present invention also encompasses an isolated polynucleotide comprising a nucleic acid that is at least about 70%, at least about 75%, 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, or any range of values thereof, to a nucleic acid encoding a fusion protein, wherein the nucleic acid comprises (1) a first polynucleotide, where the first polynucleotide encodes for a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or domain, fragment, variant, or derivative thereof; and (2) a second polynucleotide.

[0078] The present invention also encompasses an isolated polynucleotide comprising a nucleic acid that is from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identity to a nucleic acid encoding a fusion protein, wherein the nucleic acid comprises (1) a first polynucleotide, where the first polynucleotide encodes a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or domain, fragment, variant, or derivative thereof; and (2) a second polynucleotide.

[0079] In further embodiments of the fusion polynucleotide, 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 is either 5' or 3' to 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 expression in *S. cerevisiae*.

[0080] 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:3, 6, 9, 12, 15, 18, 21, 24, 27 or 30, using information from the sequences disclosed herein or the clones deposited with the ATCC. For example, allelic variants and/or species homologs can 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 homologue.

[0081] Polynucleotides comprising sequences that are at least about 70%, at least about 75%, 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, or any range of values thereof, to the entire sequence of any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27 or 30 or any fragment or domain therein can be used according to the methods described herein. In addition, polynucleotides comprising sequences that are from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identity to the entire sequence of any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27 or 30 or any fragment or domain therein can be used according to the methods described herein. Some embodiments of the invention encompass a nucleic acid molecule comprising at least about 10, at least about 20, at least about 30, at least about 35, at least about 40, at least about 50, at least about 60, at least about 70, at least about 80, at least about 90, at least about 100, at least about 200, at least about 300, at least about 400, at least about 500, at least about 600, at least about 700, or at least about 800 consecutive nucleotides, or more, or any range of values thereof, of any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24, 27 or 30, or domains, fragments, variants, or derivatives thereof.

[0082] In further aspects of the invention, nucleic acid molecules disclosed herein, encode a polypeptide having BGL functional activity. The phrase "a polypeptide having BGL functional activity" is intended to refer to a polypeptide exhibiting activity similar, but not necessarily identical, to a functional activity of the BGL polypeptides of the present invention, as measured, for example, in a particular biological assay. For example, a BGL functional activity can routinely be measured by determining the ability of a BGL polypeptide to hydrolyze

oligomers of glucose which are linked via beta 1-4 type bonds, including dimers (cellobiose), where they usually have higher activity, as well as longer chain oligomers where they usually have less activity.

[0083] Of course, 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 of a described identity to a nucleic acid sequence, or fragments thereof, will encode polypeptides "having BGL functional activity." In fact, since degenerate variants of any of these nucleotide sequences all 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 BGL functional activity.

[0084] Fragments of the full length gene of the present invention can be used as a hybridization probe for a cDNA library to isolate the full length cDNA and to isolate other cDNAs which have a high sequence similarity to the BGL genes of the present invention, or a gene encoding for a protein with similar biological activity. The probe length can vary from 5 bases to tens of thousands of bases, and will depend upon the specific test to be done. Typically a probe length of about 15 bases to about 30 bases is suitable. Only part of the probe molecule need be complementary to the nucleic acid sequence to be detected. In addition, the complementarity between the probe and the target sequence need not be perfect. Hybridization does occur between imperfectly complementary molecules with the result that a certain fraction of the bases in the hybridized region are not paired with the proper complementary base.

[0085] In certain embodiments, a hybridization probe can have at least 30 bases and can contain, for example, 50 or more bases. The probe can also be used to identify a cDNA clone corresponding to a full length transcript and a genomic clone or clones that contain the complete gene including regulatory and promoter regions, exons, and introns. An example of a screen comprises isolating the coding region of the gene by using the known DNA sequence to synthesize an oligonucleotide probe. Labeled oligonucleotides having a sequence complementary to that of the gene of the present invention are used to screen a library of bacterial or fungal cDNA, genomic DNA or mRNA to determine to which members of the library the probe hybridizes.

[0086] The present invention further relates to polynucleotides which hybridize to the hereinabove-described sequences if there is at least about 70%, at least about 90%, or at least

about 95% identity between the sequences. The present invention particularly relates to polynucleotides which hybridize under stringent conditions to the hereinabove-described polynucleotides. As herein used, the term "stringent conditions" means hybridization will occur only if there is at least about 95% or at least about 97% identity between the sequences. In certain aspects of the invention, the polynucleotides which hybridize to the hereinabove described polynucleotides encode polypeptides which either retain substantially the same biological function or activity as the mature polypeptide encoded by the DNAs of any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24 or 30.

[0087] Alternatively, polynucleotides which hybridize to the hereinabove-described sequences can have at least 20 bases, at least 30 bases, or at least 50 bases which hybridize to a polynucleotide of the present invention and which has an identity thereto, as hereinabove described, and which may or may not retain activity. For example, such polynucleotides can be employed as probes for the polynucleotide of any of SEQ ID NOs:3, 6, 9, 12, 15, 18, 21, 24 or 30, for example, for recovery of the polynucleotide or as a diagnostic probe or as a PCR primer.

[0088] Hybridization methods are well defined and have been described above. Nucleic acid hybridization is adaptable to a variety of assay formats. One of the most suitable is the sandwich assay format. The sandwich assay is particularly adaptable to hybridization under non-denaturing conditions. A primary component of a sandwich-type assay is a solid support. The solid support has adsorbed to it or covalently coupled to it immobilized nucleic acid probe that is unlabeled and complementary to one portion of the sequence.

[0089] For example, genes encoding similar proteins or polypeptides to those of the instant invention could be isolated directly by using all or a portion of the instant nucleic acid fragments as DNA hybridization probes to screen libraries from any desired bacteria using methodology well known to those skilled in the art. Specific oligonucleotide probes based upon the instant nucleic acid sequences can be designed and synthesized by methods known in the art (*see, e.g., Maniatis, 1989*). Moreover, the entire sequences can be used directly to synthesize DNA probes by methods known to the skilled artisan such as random primers DNA labeling, nick translation, or end-labeling techniques, or RNA probes using available in vitro transcription systems.

[0090] In certain aspects of the invention, polynucleotides which hybridize to the hereinabove-described sequences having at least 20 bases, at least 30 bases, or at least 50 bases which hybridize to a polynucleotide of the present invention can be employed as PCR primers.

Typically, in PCR-type amplification techniques, the primers have different sequences and are not complementary to each other. Depending on the desired test conditions, the sequences of the primers should be designed to provide for both efficient and faithful replication of the target nucleic acid. Methods of PCR primer design are common and well known in the art. Generally two short segments of the instant sequences can be used in polymerase chain reaction (PCR) protocols to amplify longer nucleic acid fragments encoding homologous genes from DNA or RNA. The polymerase chain reaction can also be performed on a library of cloned nucleic acid fragments wherein the sequence of one primer is derived from the instant nucleic acid fragments, and the sequence of the other primer takes advantage of the presence of the polyadenylic acid tracts to the 3' end of the mRNA precursor encoding microbial genes. Alternatively, the second primer sequence can be based upon sequences derived from the cloning vector. For example, the skilled artisan can follow the RACE protocol (Frohman *et al.*, *PNAS USA* 85:8998 (1988)) to generate cDNAs by using PCR to amplify copies of the region between a single point in the transcript and the 3' or 5' end. Primers oriented in the 3' and 5' directions can be designed from the instant sequences. Using commercially available 3' RACE or 5' RACE systems (BRL), specific 3' or 5' cDNA fragments can be isolated (Ohara *et al.*, *PNAS USA*, 86:5673, 1989; Loh *et al.*, *Science*, 243:217, 1989).

[0091] In addition, specific primers can be designed and used to amplify a part of or the full-length of the instant sequences. The resulting amplification products can be labeled directly during amplification reactions or labeled after amplification reactions, and used as probes to isolate full length DNA fragments under conditions of appropriate stringency.

[0092] Therefore, the nucleic acid sequences and fragments thereof of the present invention can be used to isolate genes encoding homologous proteins from the same or other fungal species or bacterial species. Isolation of homologous genes using sequence-dependent protocols is well known in the art. Examples of sequence-dependent protocols include, but are not limited to, methods of nucleic acid hybridization, and methods of DNA and RNA amplification as exemplified by various uses of nucleic acid amplification technologies (*e.g.*, polymerase chain reaction, Mullis *et al.*, U.S. Pat. No. 4,683,202; ligase chain reaction (LCR) (Tabor *et al.*, *Proc. Acad. Sci. USA*, 82:1074, 1985); or strand displacement amplification (SDA), (Walker *et al.*, *Proc. Natl. Acad. Sci. USA*, 89:392, 1992).

[0093] The polynucleotides of the present invention also comprise nucleic acids encoding a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium*

decumbens, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or domain, fragment, variant, or derivative thereof, fused to a polynucleotide encoding a marker sequence which allows for selection and/or detection of the presence of the polynucleotide in an organism.

Expression of the marker can be independent from expression of the BGL polypeptide. The marker sequence can be a yeast selectable marker such as one or more of *URA3*, *HIS3*, *LEU2*, *TRP1*, *LYS2*, *ADE2* or *SMR1*. See, e.g., Casey *et al.*, *J. Inst. Brew.*, 94:93-97, 1988.

[0094] In other embodiments of the present invention, the BGL is derived from *Saccharomycopsis fibuligera*. In other embodiments, the BGL is a beta-glucosidase I or a beta-glucosidase II isoform, paralogue or orthologue. In other embodiments, the BGL expressed by the cells of the present invention is recombinant beta-glucanase I from a *Saccharomycopsis fibuligera* source.

Codon Optimization

[0095] As used herein the term "codon-optimized" means a nucleic acid (e.g., a nucleic acid coding region) that has been adapted for expression in the cells of a given organism by replacing 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.

[0096] 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. The Codon Adaptation Index is described in more detail in Sharp *et al.* (*Nucleic Acids Research*, 15:1281-1295, 1987).

[0097] The CAI of codon-optimized sequences of the present invention correspond to from about 0.6 to about 1.0, from about 0.7 to about 1.0, from about 0.8 to about 1.0, from about 0.9 to about 1.0, from about 0.95 to about 1.0, or about 1.0. A codon-optimized sequence can 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 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 can 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.

[0098] 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 1. 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 1: 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 "

[0099] 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.

[0100] 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 and codon-optimizing programs are readily available, for example, at <http://phenotype.biosci.umbc.edu/codon/sgd/index.php> (visited September 4, 2009) or at <http://www.kazusa.or.jp/codon/> (visited September 4, 2009), and these tables can be adapted in a number of ways. See Nakamura *et al.*, *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 2. This table uses mRNA nomenclature, and so instead of thymine (T) which is found in DNA, the tables use 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 2: 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

Amino Acid	Codon	Number	Frequency per hundred
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
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

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

[0101] 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.

[0102] 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 2 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.

[0103] 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 2 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.

[0104] 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 will can vary significantly using this method; however, the sequence always encodes the same polypeptide.

[0105] When using the methods above, the term "about" is used precisely to account for fractional percentages of codon frequencies for a given amino acid. For such methods, "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.

[0106] 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 March 14, 2013). 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.

[0107] 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 are 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.

[0108] In certain embodiments, an entire polypeptide sequence, or fragment, variant, or derivative thereof is codon-optimized by any of the methods described herein. Various desired fragments, variants or derivatives are designed, and each is then codon-optimized individually. In addition, partially codon-optimized coding regions of the present invention can be designed and constructed. For example, the invention includes a nucleic acid fragment of a codon-optimized coding region encoding a polypeptide in which at least about 1%, 2%, 3%, 4%, 5%,

10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100% of the codon positions have been codon-optimized for a given species. That is, they contain a codon that is preferentially used in the genes of a desired species, *e.g.*, a yeast species such as *Saccharomyces cerevisiae*, in place of a codon that is normally used in the native nucleic acid sequence.

[0109] 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.

[0110] Codon-optimized sequences (*e.g.*, coding regions) can be versions encoding a BGL from *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans*, or a domain, fragment, variant, or derivative thereof.

[0111] Codon optimization is carried out for a particular species by methods described herein. For example, in certain embodiments, codon-optimized sequences (*e.g.*, coding regions) encoding polypeptides of a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or a domain, fragment, variant, or derivative thereof are optimized according to yeast codon usage, *e.g.*, *Saccharomyces cerevisiae*. In particular, the present invention relates to codon-optimized coding regions encoding polypeptides of a *Humicola grisea*, *Aspergillus aculeatus*, or *Aspergillus oryzae* BGL, or a domain, variant, or derivative thereof which have been optimized according to yeast codon usage, for example, *Saccharomyces cerevisiae* codon usage. Also

provided are polynucleotides, vectors, and other expression constructs comprising codon-optimized coding regions encoding BGL polypeptides of *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* or a domain, fragment, variant, or derivative thereof, and various methods of using such polynucleotides, vectors and other expression constructs.

[0112] In certain embodiments described herein, a codon-optimized sequence encoding the polypeptide sequence of any of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28, or a domain, fragment, variant, or derivative thereof, is optimized according to codon usage in yeast (*Saccharomyces cerevisiae*). Alternatively, a codon-optimized coding region encoding the polypeptide sequence of any of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25 or 28, can be optimized according to codon usage in any plant, animal, or microbial species.

BGL Polypeptides

[0113] The present invention further relates to the expression of *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, and *Phytophthora infestans* BGL polypeptides. The sequences of these peptides are available in GenBank and examples are set forth in Example 1.

[0114] The present invention further encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is at least about 70%, at least about 75%, 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, or any range of values thereof, for example, to the polypeptide sequences shown in any of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25, or 28, and/or domains, fragments, variants, or derivative thereof, of any of these polypeptides (e.g., those fragments described herein, or domains of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29).

[0115] The present invention further encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identity, for example, to the polypeptide sequences shown in any of SEQ ID NOs:1, 4, 7, 10, 13, 16, 19, 22, 25, or 28, and/or domains, fragments, variants, or derivative thereof, of any of these polypeptides (e.g., those fragments described herein, or

domains of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29). Examples of BGL domains have been described and include, for example, a glycosyl hydrolase family 3 n-terminal domain, a glycosyl hydrolase family 3 c-terminal domain, and a fibronectin type III like domain.

[0116] By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a query amino acid sequence of the present invention, it is intended that the amino acid sequence of the subject polypeptide is identical to the query sequence except that the subject polypeptide sequence can include up to five amino acid alterations per each 100 amino acids of the query amino acid sequence. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a query amino acid sequence, up to 5% of the amino acid residues in the subject sequence can be inserted, deleted or substituted with another amino acid. These alterations of the reference sequence can occur at the amino- or carboxy- terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence.

[0117] As a practical matter, whether any particular polypeptide is at least 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% identical to, for instance, any of the amino acid sequences of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29 can be determined conventionally using known computer programs. As discussed above, 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.* 6:237-245(1990)). In a sequence alignment the query and subject sequences are either both nucleotide sequences or both amino acid sequences. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB amino acid alignment are: Matrix=PAM 0, k-tuple=2, Mismatch Penalty=1, Joining Penalty=20, Randomization Group Length=0, Cutoff Score=1, Window Size=sequence length, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the subject amino acid sequence, whichever is shorter. Also as discussed above, manual corrections can be made to the results in certain instances.

[0118] 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.

[0119] The present invention also encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is at least about 70%, at least about 75%, 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, or any range of values thereof, to the polypeptide of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29, or to portions of such polypeptide, wherein the portion can contain at least 30 amino acids, at least 50 amino acids, at least 100 amino acids, at least 150 amino acids, at least 200 amino acids, at least 250 amino acids, at least 300 amino acids, or at least 350 amino acids.

[0120] The present invention also encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence is from about 70% to 100%, from about 75% to 100%, from about 80% to 100%, from about 85% to 100%, from about 90% to 100%, from about 95% to 100% identical to the polypeptide of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29, or to portions of such polypeptide, wherein the portion can contain at least 30 amino acids, at least 50 amino acids, at least 100 amino acids, at least 150 amino acids, at least 200 amino acids, at least 250 amino acids, at least 300 amino acids, or at least 350 amino acids.

[0121] The present invention further relates to a domain, fragment, variant, derivative, or analog of the polypeptide of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29.

[0122] 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.

[0123] Fragments of BGL polypeptides of the present invention can encompass domains, proteolytic fragments, and deletion fragments of *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptides. The fragments can optionally retain a specific biological activity of the BGL protein. Exemplary fragments include those described in Example 1. Polypeptide fragments further include any portion of the polypeptide which comprises a catalytic activity of the BGL protein.

[0124] The variant, derivative or analog of the polypeptide of any of SEQ ID NOs:1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29 can be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-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 or (v) one in which a fragment of the polypeptide is soluble, *i.e.*, not membrane bound, yet still binds ligands to the membrane bound receptor. Such variants, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

[0125] 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 does 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.

[0126] By an "allelic variant" is intended 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 can 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 *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL protein.

[0127] The allelic variants, the conservative substitution variants, and members of the BGL protein family, will have an amino acid sequence having at least 75%, at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% amino acid sequence

identity, or any range of values thereof, with a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL amino acid sequence set forth in any one of SEQ ID NOs: 1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29. 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.

[0128] Thus, the proteins and peptides of the present invention include molecules comprising the amino acid sequence of any one of SEQ ID NOs: 1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23, 25, 26, 28 or 29 or fragments thereof having a consecutive sequence of at least about 3, 4, 5, 6, 10, 15, 20, 25, 30, 35, 50, 100, 150, 200, 250, 300, 350, or more amino acid residues, or any range of values thereof, of the *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide sequence; 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 organisms, the alleles or other naturally occurring variants of the family of proteins; and derivatives wherein the protein has been covalently modified by substitution, chemical, enzymatic, or other appropriate means with a moiety other than a naturally occurring amino acid (for example, a detectable moiety such as an enzyme or radioisotope).

[0129] Using known methods of protein engineering and recombinant DNA technology, variants can be generated to improve or alter the characteristics of the BGL polypeptides. For instance, one or more amino acids can be deleted from the N-terminus or C-terminus of the secreted protein without substantial loss of biological function.

[0130] Thus, the invention further includes *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL 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 to have little effect on activity.

[0131] The skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly affect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid), as further described below.

[0132] For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie *et al.*, *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.

[0133] 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 indicate 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.

[0134] 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. See, e.g., Cunningham *et al.*, *Science*, 244:1081-1085, 1989. The resulting mutant molecules can then be tested for biological activity.

[0135] As the authors state, 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.

[0136] The terms "derivative" and "analog" refer to a polypeptide differing from the *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide, but retaining essential properties thereof. Generally, derivatives and analogs are overall closely similar, and, in many regions, identical to the *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide. The term "derivative" and "analog" when referring to *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, and *Phytophthora infestans* BGL polypeptides of the present invention include polypeptides which retain at least some of the activity of the corresponding native polypeptide.

[0137] Derivatives of *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, and *Phytophthora infestans* BGL polypeptides of the present invention are polypeptides which have been altered so as to exhibit additional features not found on the native polypeptide. Derivatives can be covalently modified by substitution, chemical, enzymatic, or other appropriate means with a moiety other than a naturally occurring amino acid (for example, a detectable moiety such as an enzyme or radioisotope). Examples of derivatives include fusion proteins.

[0138] An analog is another form of a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide of the present invention. An "analog" also retains substantially the same biological function or activity as the polypeptide of interest, i.e., functions as a beta-glucosidase. An analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

[0139] The polypeptide of the present invention can be a recombinant polypeptide, a natural polypeptide, or a synthetic polypeptide.

BGL Fusion Polypeptides

[0140] The present invention also encompasses fusion proteins comprising two or more polypeptides. For example, the fusion proteins can be a fusion of a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL and a second peptide. The BGL 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 BGL and/or a second peptide that is C-terminal to the heterologous cellulase. Thus, in certain embodiments, the polypeptide of the present invention comprises a first polypeptide and a second polypeptide, wherein the first polypeptide comprises a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide.

[0141] According to the present invention, the fusion protein can comprise a first and second polypeptide wherein the first polypeptide comprises a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide and the second polypeptide comprises a signal sequence. The signal sequence can be from any organism. For example, in some embodiments, the second polypeptide is a *Saccharomyces cerevisiae* (*S. cerevisiae*) polypeptide. In one particular embodiment, the *S. cerevisiae* polypeptide is the *S. cerevisiae* alpha mating factor signal sequence. In some embodiments, the signal sequence comprises the amino acid sequence of any one of SEQ ID NOs:2, 5, 8, 11, 17, 20, 23, 26 or 29, or any fragment or variant thereof described herein.

[0142] According to another embodiment, the fusion protein can comprise a first and second polypeptide, wherein the first polypeptide comprises a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide 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, a MYC-tag, or a fluorescent protein.

[0143] In certain other embodiments, the first polypeptide and the second polypeptide are fused via a linker sequence. The linker sequence can, in some embodiments, comprise the sequence: GGSPPS (SEQ ID NO:41). The linker sequence can, in other embodiments, be encoded by a codon-optimized polynucleotide of the invention described further herein.

[0144] In further embodiments of the fusion protein, the first and second polypeptide are in the same orientation, or the second polypeptide is in the reverse orientation of the first polypeptide. In additional embodiments, the first polypeptide is either N-terminal or C-terminal to the second polypeptide. In certain other embodiments, the first polypeptide and/or the second polypeptide are encoded by codon-optimized polynucleotides, for example, polynucleotides codon-optimized for *S. cerevisiae*.

Vectors and Host Cells

[0145] The present invention also 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.

[0146] Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which can be, for example, a cloning vector or an expression vector. The vector can 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.

[0147] The polynucleotides of the present invention can be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide can 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. Such vectors also include "suicide vectors" which are not self-replicating but can be replicated after insertion into the host chromosome. Other vectors can also be used.

[0148] 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.

[0149] The DNA sequence in the expression vector is operatively associated with an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. Representative examples of such promoters are as follows:

Table 3: Promoters

Gene	Organism	Systematic name	Reason for use/benefits
<i>PGK1</i>	<i>S. cerevisiae</i>	YCR012W	Strong constitutive promoter
<i>ENO1</i>	<i>S. cerevisiae</i>	YGR254W	Strong constitutive promoter
<i>TDH3</i>	<i>S. cerevisiae</i>	YGR192C	Strong constitutive promoter
<i>TDH2</i>	<i>S. cerevisiae</i>	YJR009C	Strong constitutive promoter
<i>TDH1</i>	<i>S. cerevisiae</i>	YJL052W	Strong constitutive promoter
<i>ENO2</i>	<i>S. cerevisiae</i>	YHR174W	Strong constitutive promoter
<i>GPM1</i>	<i>S. cerevisiae</i>	YKL152C	Strong constitutive promoter
<i>TPI1</i>	<i>S. cerevisiae</i>	YDR050C	Strong constitutive promoter

[0150] In addition, *Escherichia coli* (*E. coli*) promoters, such as lac or trp, are known to control expression of genes in prokaryotic or lower eukaryotic cells. The expression vector can also contain a ribosome binding site for translation initiation and a transcription terminator. The vector can also include appropriate sequences for amplifying expression, or can include additional regulatory regions. The vector can also include an enterokinase site for linking to a C-terminal tag to allow for cleavage of the target protein following protein purification.

[0151] In addition, the expression vectors can 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 or neomycin (G418) resistance or zeocin resistance for eukaryotic cell culture, or chloramphenicol, thiamphenicol, streptomycin, tetracycline, kanamycin, hygromycin, phleomycin or ampicillin resistance in *E. coli*.

[0152] The vector containing the appropriate DNA sequence as herein, as well as an appropriate promoter or control sequence, can be employed to transform an appropriate host to permit the host to express the protein.

[0153] Thus, in certain aspects, the present invention relates to host cells containing the above-described constructs. The host cell can be a higher eukaryotic cell, such as a mammalian

cell, or a lower eukaryotic cell, such as a yeast cell, e.g., *Saccharomyces cerevisiae*, or the host cell can be a prokaryotic cell, such as a bacterial cell.

[0154] Representative examples of appropriate hosts include, for example, bacterial cells, such as *E. coli*, *Streptomyces*, *Salmonella typhimurium*; thermophilic or mesophilic bacteria; fungal cells, such as yeast; and plant cells, etc. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

[0155] Appropriate fungal hosts include yeast. In certain aspects of the invention the yeast is *Saccharomyces cerevisiae*, *Saccharomyces pastorianus* (also known as *Saccharomyces carlsbergensis*), *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus* or *Schwanniomyces occidentalis*. In some embodiments, the host cell can be an oleaginous yeast cell. In some particular embodiments, the oleaginous yeast cell is a *Blakeslea*, *Candida*, *Cryptococcus*, *Cunninghamella*, *Lipomyces*, *Mortierella*, *Mucor*, *Phycomces*, *Pythium*, *Rhodospiridium*, *Rhodotorula*, *Trichosporon* or *Yarrowia* cell.

[0156] According to the methods described herein, the yeast strains can be modified, e.g. to improve growth, selection, and/or stability. Thus, for example, the *Saccharomyces cerevisiae*, *Saccharomyces pastorianus* (also known as *Saccharomyces carlsbergensis*), *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus* or *Schwanniomyces occidentalis* can include deletions, insertions, and/or rearrangements and still be considered *Saccharomyces cerevisiae*, *Saccharomyces pastorianus* (also known as *Saccharomyces carlsbergensis*), *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus* or *Schwanniomyces occidentalis*. Derivatives of the aforementioned yeast cells, i.e., yeast that have been adapted sufficiently to diverge the genome to the extent that it is a different species can also be used according to the present methods. Thus, the host cells described herein include derivatives of *Saccharomyces cerevisiae*, *Saccharomyces pastorianus*

(also known as *Saccharomyces carlsbergensis*), *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus* and *Schwanniomyces occidentalis*.

[0157] More particularly, the present invention also includes recombinant constructs comprising one or more of the sequences as broadly described above. The constructs comprise a vector, such as a plasmid or viral vector, into which a sequence of the invention has been inserted, in a forward or reverse orientation. In one aspect of this embodiment, the construct further comprises regulatory sequences, including, for example, a promoter, operably associated to the sequence. Large numbers of suitable vectors and promoters are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example.

[0158] Yeast: Yeast vectors include those of five general classes, based on their mode of replication in yeast, YIp (yeast integrating plasmids), YRp (yeast replicating plasmids), YCp (yeast replicating plasmids with centromere (CEN) elements incorporated), YEpl (yeast episomal plasmids), and YLp (yeast linear plasmids). With the exception of the YLp plasmids, all of these plasmids can be maintained in *E. coli* as well as in *Saccharomyces cerevisiae* and thus are also referred to as yeast shuttle vectors. In certain aspects, these plasmids contain two types of selectable genes: plasmid-encoded drug-resistance genes and cloned yeast genes, where the drug resistant gene is typically used for selection in bacterial cells and the cloned yeast gene is used for selection in yeast. Drug-resistance genes include ampicillin, kanamycin, tetracycline, neomycin and sulfometuron methyl. Cloned yeast genes include *HIS3*, *LEU2*, *LYS2*, *TRP1*, *URA3*, *TRP1* and *SMR1*. pYAC vectors can also be utilized to clone large fragments of exogenous DNA on to artificial linear chromosomes.

[0159] In certain aspects of the invention, YCp plasmids, which have high frequencies of transformation and increased stability due to the incorporated centromere elements, are utilized. In certain other aspects of the invention, YEpl plasmids, which provide for high levels of gene expression in yeast, are utilized. In additional aspects of the invention, YRp plasmids are utilized.

[0160] In certain embodiments, the vector comprises (1) a first polynucleotide, where the first polynucleotide encodes for a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*,

Debaryomyces hansenii, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or domain, fragment, variant, or derivative thereof; and (2) a second polynucleotide, where the second polynucleotide encodes for a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL, or domain, fragment, variant, or derivative thereof.

[0161] 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 is either N-terminal or C-terminal to 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*.

[0162] In particular embodiments, the vector of the present invention is a plasmid selected from pMU3557, pMU3558, pMU3559, pMU3560, pMU3561, pMU3562, pMU3563, pMU3564, pMU3565, or pMU3566 (SEQ ID NOs:31-40). Descriptions of these plasmids are found in Example 1 and FIGs. 1-10. However, any other plasmid or vector can be used as long as they are replicable and viable in the host.

[0163] Promoter regions can be selected from any desired gene. Particular named yeast promoters include the *ENO1* promoter, the *PGK1* promoter, the *TEF1* promoter, and the *HXT7* promoter. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, PL and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

[0164] Introduction of the construct into a host yeast cell, *e.g.*, *Saccharomyces cerevisiae*, can be effected by lithium acetate transformation, spheroplast transformation, or transformation by electroporation, as described, for example, in Current Protocols in Molecular Biology, 13.7.1-13.7.10.

[0165] Introduction of the construct in other host cells can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation. *See e.g.*, Davis *et al.*, Basic Methods in Molecular Biology, 1986.

[0166] The constructs in host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence. Alternatively, the polypeptides of the invention can be synthetically produced by conventional peptide synthesizers.

[0167] Following creation of a suitable host cell and growth of the host cell to an appropriate cell density, the selected promoter is induced by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period.

[0168] Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

[0169] Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, such methods are well known to those skilled in the art.

[0170] Yeast cells, e.g., *Saccharomyces cerevisiae*, employed in expression of proteins can be manipulated as follows. The BGL polypeptides can be secreted by cells and therefore can be easily recovered from supernatant using methods known to those of skill in the art. Proteins can also be recovered and purified from recombinant cell cultures by methods including spheroplast preparation and lysis, cell disruption using glass beads, and cell disruption using liquid nitrogen, for example.

[0171] Various mammalian cell culture systems can also be employed to express recombinant protein. Expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences.

[0172] Additional methods include ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

[0173] The BGL polypeptides can be prepared in any suitable manner. Such polypeptides include isolated naturally occurring polypeptides; recombinantly produced polypeptides, synthetically produced polypeptides, or polypeptides produced by a combination of these methods. Means for preparing such polypeptides are well understood in the art.

[0174] BGL polypeptides are provided in an isolated form, and, in certain aspects, are substantially purified. A recombinantly produced version of a BGL polypeptide can be substantially purified using techniques described herein or otherwise known in the art, such as, for example, by the one-step method described in Smith *et al.*, *Gene*, 67:31-40, 1988. BGL polypeptides also can be purified from natural, synthetic or recombinant sources using techniques described herein or otherwise known in the art.

[0175] The BGL polypeptides of the present invention can be in the mature form, or can be a part of a larger protein, such as a fusion protein. It can be advantageous to include an additional amino acid sequence which contains secretory or leader sequences, pro-sequences, sequences which aid in purification, such as multiple histidine residues, or an additional sequence for stability during recombinant production.

[0176] Secretion of desired proteins into the growth media has the advantages of simplified and less costly purification procedures. It is well known in the art that secretion signal sequences are often useful in facilitating the active transport of expressible proteins across cell membranes. The creation of a transformed host capable of secretion can be accomplished by the incorporation of a DNA sequence that codes for a secretion signal which is functional in the host production host. Methods for choosing appropriate signal sequences are well known in the art (*see, e.g.*, European Pub. No. 546049; Int'l. Pub. No. WO 93/24631). The secretion signal DNA or facilitator can be located between the expression-controlling DNA and the instant gene or gene fragment, and in the same reading frame with the latter.

Heterologous expression of BGL polypeptides in host cells

[0177] In order to address the limitations of the previous systems, the present invention provides *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptides, and domains, variants, and derivatives thereof that can be effectively and efficiently utilized in a consolidated bioprocessing system.

[0178] In particular, the invention relates to the production of a heterologous beta-glucosidase (BGL) in a host organism. In certain embodiments, this host organism is yeast, such as *Saccharomyces cerevisiae*.

[0179] In certain embodiments of the present invention, a host cell comprising a vector which encodes and expresses a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*,

Aspergillus oryzae, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL that is utilized for consolidated bioprocessing is co-cultured with additional host cells expressing one or more additional heterologous cellulases. Additional heterologous cellulases can be derived from for example, a fungal or bacterial source.

[0180] In some embodiments, the cellulase is a xylanase, xylosidase, acetylxyLANesterase (AXE), endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase, cellobiohydrolase (CBH), or combinations thereof. In other embodiments, the endoglucanase is *Aspergillus fumigatus* (*A. fumigatus*) endoglucanase I, *Neosartorya fischeri* (*N. fischeri*) endoglucanase III, *Trichoderma reesei* (*T. reesei*) endoglucanase I, *Coptotermes formosanus* (*C. formosanus*) endoglucanase I, or combinations thereof. In some embodiments, the CBH is CBH1 or CBH2, or combinations thereof. In some embodiments, the CBH is *Talaromyces emersonii* (*T. emersonii*) cellobiohydrolase I, *Chrysosporium lucknowense* (*C. lucknowense*) cellobiohydrolase IIB, *T. reesei* cellobiohydrolase II, or combinations thereof. In other embodiments of the invention, the CBH is a CBH1 or CBH2 isoform, paralogue or orthologue.

[0181] In certain embodiments of the invention, the endoglucanase can be an endoglucanase I or an endoglucanase II isoform, paralogue or orthologue. In another embodiment, the endoglucanase expressed by the host cells of the present invention can be recombinant endo-1,4- β -glucanase. In certain embodiments of the present invention, the endoglucanase is an endoglucanase I from *T. reesei*, *A. fumigatus* EG1, *N. fischeri* EG3, *C. formosanus* endoglucanase I, or combinations thereof.

[0182] In some embodiments, a host cell of the invention can further comprise a polynucleotide encoding *Saccharomycopsis fibuligera* (*S. fibuligera*) BGL.

[0183] In some embodiments, a host cell of the invention can further comprise one or more polynucleotides encoding *T. emersonii* CBH1, *T. reesei* CBD, *C. lucknowense* CBH2, *A. fumigatus* EG1, *N. fischeri* EG3, *S. fibuligera* BGL, or *Aspergillus niger* xylanase. In other embodiments, a host cell of the invention can further comprise one or more polynucleotides encoding *A. niger* xylanase, *P.t.r.* xylosidase, *N. fischeri* AXE, *A. fumigatus* EG1, *T. reesei* AGL1, *T. reesei* beta-mannanase, *A. fumigatus* alpha-glucuronidase (FC110), *A. fumigatus* acetyl esterase (FC136), *N. fischeri* beta-mannosidase (FC124), or *S. fibuligera* BGL.

[0184] DNA and polypeptide sequences encoding these cellulases, and other exemplary cellulases, are available in GenBank and described, for example, in Int'l Pub. No. WO 2011/051806, Int'l Pub. No. WO 2011/153516, Int'l Pub. No. WO 2010/005553, Int'l Pub. No. WO 2009/139839, Int'l Pub. No. WO 2009/138877, Int'l Pub. No. WO 2010/060056, Int'l Appl. No. PCT/US2012/057952, filed September 28, 2012 and U.S. Appl. No. 61/694,690, filed August 28, 2012.

[0185] The transformed host cells or cell cultures described herein are measured for recombinant protein content. For the use of secreted cellulases, protein content can be determined by analyzing the host (*e.g.*, yeast) cell supernatants. Proteins, including tethered heterologous biomass degrading enzymes, can also be recovered and purified from recombinant cell cultures by methods including spheroplast preparation and lysis, cell disruption using glass beads, and cell disruption using liquid nitrogen for example. Additional protein purification methods include trichloroacetic acid, ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography, gel filtration, and lectin chromatography. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps.

[0186] Protein analysis methods include methods such as the traditional Lowry method, the bicinchoninic acid protein assay reagent (Pierce) or the protein assay method according to BioRad's manufacturer's protocol. Using such methods, the protein content of saccharolytic enzymes can be estimated. Additionally, to accurately measure protein concentration a BGL can be expressed with a tag, for example a His-tag or HA-tag and purified by standard methods using, for example, antibodies against the tag, a standard nickel resin purification technique or similar approach.

[0187] The transformed host cells or cell cultures described herein can be further analyzed for hydrolysis of cellulase (*e.g.*, by a sugar detection assay), for a particular type of cellulase activity (*e.g.*, by measuring the individual enzyme activity) or for total cellulase activity. Endoglucanase activity can be determined, for example, by measuring an increase of reducing ends in an endoglucanase specific carboxymethylcellulose (CMC) substrate. Cellobiohydrolase activity can be measured, for example, by using insoluble cellulosic substrates such as the amorphous substrate phosphoric acid swollen cellulose (PASC) or microcrystalline

cellulose (Avicel) and determining the extent of the substrate's hydrolysis. BGL activity, such as the "specific activity" described herein, can be measured by a variety of assays, for example, using cellobiose. Unit measurements of BGL activity and hydrolysis include, for example, umol glucose/mol or mg BGL/time (for example, seconds). Alternatively, one unit of BGL activity can be defined as the amount of enzyme required to liberate 1 umol of para-nitrophenol (pNP) from a pNP beta-glucoside or cellobiose per minute under assay conditions.

[0188] A total cellulase activity, which can include, for example, the activity of endoglucanase, CBHI, CBHII and BGL, can hydrolyze crystalline cellulose synergistically. Total cellulase activity can thus be measured using insoluble substrates including pure cellulosic substrates such as Whatman No. 1 filter paper, cotton linter, microcrystalline cellulose, bacterial cellulose, algal cellulose, and cellulose-containing substrates such as dyed cellulose, alpha-cellulose or pretreated lignocellulose.

[0189] It will be appreciated that suitable lignocellulosic material can be any feedstock that contains soluble and/or insoluble cellulose, where the insoluble cellulose can be in a crystalline or non-crystalline form. In various embodiments, the lignocellulosic biomass comprises, for example, wood, corn, corn cobs, corn stover, corn fiber, sawdust, bark, leaves, agricultural and forestry residues, grasses such as switchgrass, cord grass, rye grass or reed canary grass, miscanthus, ruminant digestion products, municipal wastes, paper mill effluent, newspaper, cardboard, miscanthus, sugar-processing residues, sugarcane bagasse, agricultural wastes, rice straw, rice hulls, barley straw, cereal straw, wheat straw, canola straw, oat straw, oat hulls, stover, soybean stover, forestry wastes, recycled wood pulp fiber, paper sludge, sawdust, hardwood, softwood or combinations thereof.

[0190] In certain embodiments of the present invention, a host cell comprising a vector which encodes and expresses a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL that is utilized for consolidated bioprocessing is co-cultured with additional host cells expressing one or more additional heterologous cellulases. In other embodiments of the invention, a host cell transformed with a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL is transformed with and/or expresses one or more other heterologous xylanase, xylosidase, AXE, endoglucanase, alpha-

galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase, or CBH, as described further herein.

[0191] Specific activity of cellulases can also be detected by methods known to one of ordinary skill in the art. To accurately measure protein concentration a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL can be expressed with a tag, for example a His-tag or hemagglutinin (HA)-tag and purified by standard methods using, for example, antibodies against the tag, a standard nickel resin purification technique or similar approach.

[0192] In other embodiments, the host cell produces the BGL in a culture. In some embodiments, BGL is produced in an amount of at least about 0.6 mg, at least about 0.7 mg, at least about 0.8 mg, at least about 0.9 mg, at least about 1 mg, at least about 1.5 mg, at least about 2 mg, at least about 2.5 mg, at least about 3 mg, at least about 3.5 mg, at least about 4 mg, at least about 4.5 mg, at least about 5 mg, at least about 6 mg, at least about 7 mg, at least about 8 mg, at least about 9 mg or at least about 10 mg, of any ranges thereof. In other embodiments, BGL is produced in an amount of from about 0.6 mg to about 10 mg, from about 1 mg to about 10 mg, or from about 1 mg to about 5 mg.

[0193] In other embodiments, the host cell produces the BGL in a concentration of at least about 0.2 mg/ml in culture. In some embodiments, the concentration is at least about 0.2 mg/ml, at least about 0.5 mg/ml, at least about 1 mg/ml, at least about 1.5 mg/ml, at least about 2 mg/ml, at least about 2.5 mg/ml, at least about 3 mg/ml, at least about 3.5 mg/ml, at least about 4 mg/ml, at least about 4.5 mg/ml, at least about 5 mg/ml, at least about 5.5 mg/ml, or at least about 6 mg/ml, or any range of values thereof. In some embodiments, the concentration is from about 0.2 mg/ml to about 6 mg/ml, from about 0.2 mg/ml to about 5 mg/ml, from about 0.2 mg/ml to about 0.2 mg/ml to about 3 mg/ml.

[0194] In other embodiments, the present invention also provides a method for hydrolyzing a cellulosic substrate. In embodiments, the method comprises contacting the cellulosic substrate with a host cell, co-culture, composition, peptide or purified peptide of the invention. In some embodiments, the cellulosic substrate comprises a lignocellulosic biomass. In other embodiments, the lignocellulosic biomass is grass, switch grass, cord grass, rye grass, reed canary grass, miscanthus, sugar-processing residues, sugarcane bagasse, agricultural wastes, rice straw, rice hulls, barley straw, corn cobs, cereal straw, wheat straw, canola straw, oat straw,

oat hulls, corn fiber, stover, soybean stover, corn stover, forestry wastes, recycled wood pulp fiber, paper sludge, sawdust, hardwood, softwood, or combinations thereof. In other embodiments, the cellulosic substrate can be hydrolyzed to xylose, glucose, mannose, galactose, arabinose, or combinations thereof. In some embodiments, the cellulose substrate is hydrolyzed to cellulosic substrate is hydrolyzed to xylose, glucose, mannose, galactose or arabinose at a rate at least about 10% greater than the rate of a host cell comprising a polynucleotide encoding a BGL from *S. fibuligera*. In other embodiments, the rate is at least about 10% greater, at least about 20% greater, at least about 30% greater, at least about 40% greater, at least about 50% greater, at least about 60% greater, at least about 70% greater, at least about 80% greater, at least about 90% greater, or at least about 100% greater, or any range of values thereof. In other embodiments, the rate is from about 10% greater to about 100% greater, from about 10% greater to about 70% greater, from about 10% greater to about 60% greater, from about 10% greater to about 50% greater, from about 20% greater to about 70% greater, from about 30% greater to about 70% greater, or from about 30% greater to about 60% greater.

[0195] In some embodiments of the methods of the invention, the BGL is present in an amount of about 0.2 mg or less per gram of xylose.

[0196] The present invention also provides a method of fermenting cellulose, comprising culturing a host cell, co-culture, composition, peptide or purified peptide of the invention in medium. In some embodiments, the medium contains crystalline cellulose. In some embodiments, the culturing is under suitable conditions for a period sufficient to allow saccharification and fermentation of the cellulose. In other embodiments, the host cell produces ethanol.

[0197] 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. 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.

Co-Cultures

[0198] The present invention is also directed to co-cultures comprising at least two yeast host cells wherein the at least one yeast host cell comprises a polynucleotide encoding a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium*

decumbens, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide and at least one other yeast host cell comprises a polynucleotide encoding a heterologous cellulase. As used herein, "co-culture" refers to growing two different strains or species of host cells together in the same vessel. In some embodiments of the invention, at least one host cell of the co-culture comprises a heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase, at least one host cell of the co-culture comprises a heterologous polynucleotide comprising a nucleic acid which encodes a β -glucosidase and at least one host cell comprises a heterologous polynucleotide comprising a nucleic acid which encodes a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL polypeptide. In a further embodiment, the co-culture further comprises a host cell comprising a heterologous polynucleotide comprising a nucleic acid which encodes a second BGL.

[0199] The co-culture can comprise two or more strains of yeast host cells and the heterologous cellulases can be expressed in any combination in the two or more strains of host cells. For example, according to the present invention, the co-culture can comprise two strains: one strain of host cells that expresses one or more cellulases described herein and a second strain of host cells that expresses a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL. Alternatively, the co-culture can comprise three, four, five, six, seven, eight, or more strains of host cells that each express one or more cellulases described herein and/or a *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL.

[0200] 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 can be present in equal or unequal numbers.

[0201] The co-cultures of the present invention can include tethered cellulases, secreted cellulases or both tethered and secreted cellulases. In addition, other cellulases, such as externally added cellulases can be present in the co-culture.

[0202] According to the methods described herein, a host cell or group of host cells can comprise a vector or vectors which encode and express a combination of heterologous cellulases including one or more cellulases selected from *Humicola grisea*, *Candida wickerhamii*, *Aspergillus aculeatus*, *Aspergillus oryzae*, *Penicillium decumbens*, *Chaetomium globosum*, *Neocallimastix frontalis*, *Debaryomyces hansenii*, *Kluyveromyces marxianus*, or *Phytophthora infestans* BGL. For example, a single host cell may express endoglucanase, BGL, CBH1 and CBH2. Alternatively, a group of cells could express a combination of cellulases, for example such that a first host cell expresses endoglucanase, a second host cell expresses BGL, a third host cell expresses CBH1, and a fourth host cell expresses a CBH2. Similarly, a first host cell can express both endoglucanase and BGL and a second host cell can express both CBH1 and CBH2.

EXAMPLES

MATERIALS AND METHODS

Media and strain cultivation

[0203] Unless otherwise specified, yeast strains were routinely grown in YPD (10 g/L yeast extract, 20 g/L peptone, 20 g/L glucose), YPC (10 g/L yeast extract, 20 g/L peptone, 20 g/L cellobiose), or YNB + glucose (6.7 g/L Yeast Nitrogen Base without amino acids, and supplemented with appropriate amino acids for strain, 20 g/L glucose) media and, if needed, antibiotics for selection. 15 g/L agar was added for solid media.

Molecular methods

[0204] Unless otherwise specified, standard protocols were followed for DNA manipulations (Sambrook *et al.* 1989). Polymerase chain reaction (PCR) was performed using Phusion polymerase (New England Biolabs) for cloning, and Taq polymerase (New England Biolabs) for screening transformants, and in some cases Advantage Polymerase (Clontech) for PCR of genes for correcting auxotrophies. Manufacturers guidelines were followed as supplied. Restriction enzymes were purchased from New England Biolabs and digests were set up according to the supplied guidelines. Ligations were performed using the Quick ligation kit

(New England Biolabs) as specified by the manufacturer. Gel purification was performed using either Qiagen or Zymo research kits, PCR product and digest purifications were performed using Zymo research kits, and Qiagen midi and miniprep kits were used for purification of plasmid DNA.

Yeast transformation

[0205] A protocol for electrotransformation of yeast was developed based on Cho *et al.* (*Enzyme And Microbial Technology*, 25:23-30, 1999) and Ausubel *et al.* (Current Protocols in Molecular Biology. USA: John Wiley and Sons, Inc., 1994). Linear fragments of DNA are created by restriction enzyme digestion utilizing unique restriction sites within the plasmid. The fragments are purified by precipitation with 3M sodium acetate and ice cold ethanol, subsequent washing with 70% ethanol, and resuspension in USB dH₂O (DNAse and RNAse free, sterile water) after drying in a 70°C vacuum oven.

[0206] Unless otherwise specified, yeast cells, *e.g.*, *Saccharomyces cerevisiae*, for transformation were prepared by growing to saturation in 5 mL YPD cultures. 4 mL of the culture was sampled, washed 2X with cold distilled water, and resuspended in 640 µL cold distilled water. 80 µL of 100mM Tris-HCl, 10mM EDTA, pH 7.5 (10X TE buffer - filter sterilized) and 80 µL of 1M lithium acetate, pH 7.5 (10X LiAc - filter sterilized) was added and the cell suspension incubated at 30°C for 45 minutes with gentle shaking. 20 µL of 1M DTT was added and incubation continued for 15 minutes. The cells were then centrifuged, washed once with cold distilled water, and once with electroporation buffer (1M sorbitol, 20mM HEPES), and finally resuspended in 267 µL electroporation buffer.

[0207] For electroporation, 10 µg of linearized DNA (measured by estimation on gel) was combined with 50 µL of the cell suspension in a sterile 1.5 mL microcentrifuge tube. The mixture was then transferred to a 0.2 cm electroporation cuvette, and a pulse of 1.4 kV (200Ω, 25 µF) applied to the sample using, *e.g.*, the BioRad Gene Pulser device. 1mL of YPD with 1M sorbitol adjusted to pH 7.0 (YPDS) was placed in the cuvette and the cells allowed to recover for ~3 hrs. 100-200 µL of cell suspension was spread out on YPDS agar plates with appropriate selection, which were incubated at 30°C for 3-4 days until colonies appeared.

SDS-PAGE and gel staining

[0208] Unless otherwise specified, SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) was carried out as described by Laemmli (*Nature*, 227:680-685, 1970) on a

10% gel at 100 V. A 20 µl sample of culture supernatant was mixed with SDS-PAGE loading buffer and incubated at 95°C for 5 minutes before loading onto the gel. After protein separation, the gels were silver stained. Silver staining was performed by incubating the gels with shaking at room temperature in 1) 30% ethanol and 0.5% acetic acid (3x30 min); 2) 20% ethanol (10 min); 3) water (10 min); 4) sodium thiosulfate (0.2 g/L) (1 min); 5) water (2x20 seconds); 6) silver nitrate (2 g/L) (30 min); 7) water (5-10 seconds); 8) 37% formaldehyde (0.7 ml/L) and potassium carbonate (anhydr.) (30 g/L) and sodium thiosulfate (10 mg/L) (2x3 min or to desired intensity); 9) Tris base (50 g/L) and 2.5% acetic acid (1 min); and 10) water.

Determination of protein concentration

[0209] To estimate specific activity of the BGLs the Bradford method (BioRad protein assay) was used as it is prescribed for use in microtiter plates, using the Gamma globulin standard. Before determination of protein concentration, supernatant samples were first subjected to the buffer exchange procedure as directed for the 2 mL Zeba desalt spin columns (Thermo Scientific).

Western Blot Protocol for Supernatants of Strains:

1. Test top performing strains for activity, along with randomly selected alpha-glucuronidase strains (no activity assay available) and run on a 4-20% Tris glycine SDS-PAGE gel (Invitrogen, EC6025BOX), transfer to PVDF membrane (Amersham Hybond P, GE Healthcare, RPN303F) and block overnight in TBS (10 mM Tris, 150 mM NaCl, pH 7.5) + 2% BSA (bovine serum albumin)
2. Dilute primary Qiagen muα Penta-His 1:5000 in TBST (TBS with 0.1% Tween 20). Pour off blocker and add primary antibody. Incubate at room temperature for 1h.
3. Pour off primary antibody and wash 3x5min in THST (10 mM Tris, 500 mM NaCl, pH 7.5 with 0.1% Tween 20).
4. Dilute Thermo gtamu-HRP (cat. No. 31439) 1:7500 in TBST and add to blots. Incubate at room temperature for 1h, pour off and wash again with THST
5. Add ECL (Thermo, 32166) substrate and visualize using a Syngene G:BOX with a CCD camera.

BGL activity assay on cellobiose

- Standard curve and samples in duplicate:

- 100 µg/mL Bgl-His diluted in 50 mM Na citrate, pH 5.5; then 1:2
 - Samples diluted 1:10 in 50 mM Na citrate, pH 5.5
- To a PCR plate, add
 - 50 µL sample or standard
 - 50 µL 50 mM Na citrate, pH 5.5
 - 50 µL 50 mM Na citrate, 20 mM cellobiose, pH 5.5
 - For the blank, use 50 µL sample + 100 µL 50 mM Na citrate, pH 5.5 (no cellobiose)
- Incubate x 45 – 60 min at 35°C
- Heat 100°C x 5 min in the thermocycler
- To flat bottom clear plate, add
 - 10 µL sample
 - 100 µL HK reagent
 - Add 0.15 M Tris base to the vial to improve buffer capacity
 - Sigma kit – GAHK20 glucose HK kit
 - Unused reagent can be stored at -20°C
- Incubate x 2 hours to overnight at RT
- Read at 340 nm
- Subtract the results from the blank (residual glucose from the media) from the sample results

Purification of His tagged BGL

- Grow cells in YPD
- Centrifuge cells, filter thru 0.2 µm membrane then concentrate in a 10kDa MWCO filter
- pH adjust the sample to ~7 with 1M Tris, pH 9
- Purify on the FPLC with the following conditions:
 - Column: GE HisTrap 5 mL column
 - Mobile phase A: 25 mM Tris, pH 6.8
 - Mobile phase B: 25 mM Tris, 150 mM imidazole, pH 6.8
 - Flow rate: 5 mL/min
 - Step elution to 100% B
 - Collect 1 mL fractions

- Buffer exchange into 50 mM NaAc, pH 5
- Determine concentration by absorbance at 280 nm using the theoretical molar extinction coefficient of the protein based on its amino acid sequence (Edelhoch, (1967), Biochemistry, 6, 1948-1954).

Example 1: Screening of Yeast Produced Beta-Glucosidases for Efficient Cellobiose and Oligomer Hydrolysis

[0210] In order to find beta-glucosidase (BGL) enzymes that are well expressed in *Saccharomyces cerevisiae*, and highly active on hardwood derived substrates, several BGLs were designed and synthesized by DNA 2.0. The enzymes and sequences tested are below in Table 4.

Table 4: Beta-glucosidase enzymes tested for expression in yeast

(FC)#	Cazy family	E.C. #	Activity	Source Organism	Accession #	Strain #	Plasmid #
141	GH3	3.2.1.21	Beta-glucosidase	<i>Saccharomycopsis fibuligera</i>	P22506	M1429	pMU1172*
146	GH1	3.2.1.21	Beta-glucosidase	<i>Humicola grisea</i>	BAA74958	M4860	pMU3557
147	GH1	3.2.1.21	Beta-glucosidase	<i>Candida Wickerhamii</i>	AAC49036		pMU3558
148	GH3	3.2.1.21	Beta-glucosidase	<i>Aspergillus Aculeatus</i>	P48825	M4861	pMU3559
149	GH3	3.2.1.21	Beta-glucosidase	<i>Aspergillus oryzae</i>	XP 001816831	M4862	pMU3560
150	GH3	3.2.1.21	Beta-glucosidase	<i>Penicillium decumbens</i>	ADB82653	M4863	pMU3561
151	GH3	3.2.1.21	Beta-glucosidase	<i>Chaetomium globosum</i>	XP 001229937	M4864	pMU3562
152	GH3	3.2.1.21	Beta-glucosidase	<i>Neocallimastix frontalis</i>	AEX92706	M4865	pMU3563
153	GH3	3.2.1.21	Beta-glucosidase	<i>Debaryomyces hansenii</i>	XP 457283		pMU3564
154	GH3	3.2.1.21	Beta-glucosidase	<i>Kluyveromyces marxianus</i>	P07337		pMU3565
155	GH30	3.2.1.21	Beta-glucosidase/ Beta-xylosidase	<i>Phytophthora infestans</i>	AAK19754		pMU3566

* As described, for example, in Int'l Pub. No. WO 2011/153516.

[0211] A six-repeat histidine (6X HIS) tag was added to the C-terminus of these synthetic genes and they were cloned into an expression vector for testing in yeast.

[0212] The full amino acid sequence, with signal peptide, for *Humicola grisea* beta-glucosidase (Accession No. BAA74958) is in SEQ ID NO:1.

[0213] The native signal peptide for *Humicola grisea* beta-glucosidase (Accession No. BAA74958) is in SEQ ID NO:2.

[0214] The corresponding *Humicola grisea* beta-glucosidase DNA sequence is in SEQ ID NO:3.

[0215] The full amino acid sequence, with signal peptide, for *Candida wickerhamii* beta-glucosidase (Accession No. AAC49036) is in SEQ ID NO:4.

[0216] The native signal peptide for *Candida wickerhamii* beta-glucosidase (Accession No. AAC49036) is in SEQ ID NO:5.

[0217] The corresponding *Candida wickerhamii* beta-glucosidase DNA sequence is in SEQ ID NO:6.

[0218] The full amino acid sequence, with signal peptide, for *Aspergillus aculeatus* beta-glucosidase (Accession No. P48825) is in SEQ ID NO:7.

[0219] The native signal peptide for *Aspergillus aculeatus* beta-glucosidase (Accession No. P48825) is in SEQ ID NO:8.

[0220] The corresponding *Aspergillus aculeatus* beta-glucosidase DNA sequence is in SEQ ID NO:9.

[0221] The full amino acid sequence, with signal peptide, for *Aspergillus oryzae* beta-glucosidase (Accession No. XP_001816831) is in SEQ ID NO:10.

[0222] The native signal peptide for *Aspergillus oryzae* beta-glucosidase (Accession No. XP_001816831) is in SEQ ID NO:11.

[0223] The corresponding *Aspergillus oryzae* beta-glucosidase DNA sequence is in SEQ ID NO:12.

[0224] The full amino acid sequence, with signal peptide, for *Penicillium decumbens* beta-glucosidase (Accession No. ADB82653) is in SEQ ID NO:13.

- [0225] The native signal peptide for *Penicillium decumbens* beta-glucosidase (Accession No. ADB82653) is in SEQ ID NO:14.
- [0226] The corresponding *Penicillium decumbens* beta-glucosidase DNA sequence is in SEQ ID NO:15.
- [0227] The full amino acid sequence, with signal peptide, for *Chaetomium globosum* beta-glucosidase (Accession No. XP_001229937) is in SEQ ID NO:16.
- [0228] The native signal peptide for *Chaetomium globosum* beta-glucosidase (Accession No. XP_001229937) is in SEQ ID NO:17.
- [0229] The corresponding *Chaetomium globosum* beta-glucosidase DNA sequence is in SEQ ID NO:18.
- [0230] The full amino acid sequence, with signal peptide, for *Neocallimastix frontalis* beta-glucosidase (Accession No. AEX92706) is in SEQ ID NO:19.
- [0231] The native signal peptide for *Neocallimastix frontalis* beta-glucosidase (Accession No. AEX92706) is in SEQ ID NO:20.
- [0232] The corresponding *Neocallimastix frontalis* beta-glucosidase DNA sequence is in SEQ ID NO:21.
- [0233] The full amino acid sequence, with signal peptide, for *Debaryomyces hansenii* beta-glucosidase (Accession No. XP_457283) is in SEQ ID NO:22.
- [0234] The added signal peptide for *Debaryomyces hansenii* beta-glucosidase is in SEQ ID NO:23.
- [0235] The corresponding *Debaryomyces hansenii* beta-glucosidase DNA sequence is in SEQ ID NO:24.
- [0236] The full amino acid sequence, with signal peptide, for *Kluyveromyces marxianus* beta-glucosidase (Accession No. P07337) is in SEQ ID NO:25.
- [0237] The added signal peptide for *Kluyveromyces marxianus* beta-glucosidase is in SEQ ID NO:26.
- [0238] The corresponding *Kluyveromyces marxianus* beta-glucosidase DNA sequence is in SEQ ID NO:27.
- [0239] The full amino acid sequence, with signal peptide, for *Phytophthora infestans* beta-glucosidase (Accession No. AAK19754) is in SEQ ID NO:28.
- [0240] The native signal peptide for *Phytophthora infestans* beta-glucosidase (Accession No. AAK19754) is in SEQ ID NO:29.

[0241] The corresponding *Phytophthora infestans* beta-glucosidase DNA sequence is in SEQ ID NO:30.

[0242] The sequence of pMU3557 is in SEQ ID NO:31 (*see also* FIG. 1).

[0243] The sequence of pMU3558 is in SEQ ID NO:32 (*see also* FIG. 2).

[0244] The sequence of pMU3559 is in SEQ ID NO:33 (*see also* FIG. 3).

[0245] The sequence of pMU3560 is in SEQ ID NO:34 (*see also* FIG. 4).

[0246] The sequence of pMU3561 is in SEQ ID NO:35 (*see also* FIG. 5).

[0247] The sequence of pMU3562 is in SEQ ID NO:36 (*see also* FIG. 6).

[0248] The sequence of pMU3563 is in SEQ ID NO:37 (*see also* FIG. 7).

[0249] The sequence of pMU3564 is in SEQ ID NO:38 (*see also* FIG. 8).

[0250] The sequence of pMU3565 is in SEQ ID NO:39 (*see also* FIG. 9).

[0251] The sequence of pMU3566 is in SEQ ID NO:40 (*see also* FIG. 10).

[0252] The plasmids described in Table 4 above were transformed into the yeast strain M1744 (described in, for example, Int'l Pub. No. WO 2011/153516), and selected on synthetic complete media without uracil (SD-ura) in order to isolate transformants. These transformants were then screened for activity using a beta-glucosidase activity assay with cellobiose as the substrate to assess if functional protein was being produced (FIG. 11). FIG. 11 shows the results of screening 12 colonies for each plasmid transformed. In each case, the colony showing the best activity is shown. These results show BGL enzyme activity was present in transformants. BGL from *Aspergillus aculeatus*, *Aspergillus oryzae*, and *Humicola grisea* showed the highest functional activity.

[0253] SDS-PAGE was also used to assess if BGL protein was being produced in the transformants (FIGs. 12A-12C, left panel). These results show recombinant BGL protein was present in the transformants. BGL from *Aspergillus aculeatus* and *Aspergillus oryzae* showed the highest levels of production. In addition, western blots were conducted to further assess the presence of recombinant BGL protein. FIGs. 12A-12C (right panel) shows the results of these blots. These results show recombinant BGL protein was present in the transformants. BGL protein produced by strains harboring the pMU3557, pMU3559 and pMU3560 plasmids showed the highest levels of production.

[0254] BGLs that showed activity and/or protein production were subsequently purified and used in hydrolysis assays with both pretreated hardwood solids and concentrated C5 liquor. Several strains were grown in shake flask culture in order to purify the beta-glucosidase enzyme

via the associated 6X HIS tag. These strains included: M4860, M4861, M4862, M4863, M4864 and M4865. The associated BGL protein concentration recovered after purification is listed below in Table 5.

Table 5: Amount of protein purified from cultures of BGL producing strains of *S. cerevisiae*

Strain	Source Organism	Concentration (mg/mL)	Volume (mL)	Total Protein (mg)
M4860	<i>Humicola grisea</i>	0.2	3	0.6
M4861	<i>Aspergillus Aculeatus</i>	0.6	2	1.2
M4862	<i>Aspergillus oryzae</i>	3.2	1.5	4.8
M4863	<i>Penicillium decumbens</i>	0.11	1.5	0.165
M4864	<i>Chaetomium globosum</i>	0.26	4	1.04
M4865	<i>Neocallimastix frontalis</i>	0.09	1	0.09

[0255] The data in Table 5 indicate BGL protein was present in the strains. BGL from *Aspergillus aculeatus*, *Aspergillus oryzae* and *Chaetomium globosum* showed the highest concentrations.

[0256] After the BGL enzymes were purified, their specific activities were compared by hydrolysis assays on cellobiose at pH 5 and 37°C (FIGs. 13 and 14). The hydrolysis assays contained pretreated hardwood solids (2% solids loading) or diluted C5 liquor, along with sodium citrate buffered to pH 5.2, purified enzyme and sodium azide to prevent contamination. The resultant sugars were analyzed by BioRad Aminex 87H and 87P high performance liquid chromatography (HPLC) to determine the usefulness of each enzyme. The 87H column can measure acetic acid, but also results in xylose, galactose, and mannose co-eluting, while the 87P column can resolve xylose, galactose, and mannose, but cannot measure acetic acid release. For this reason, both columns were employed to analyze the release of sugars.

[0257] Over the range of enzyme loadings tested in FIG. 13, it is clear that the purified BGLs had specific activity against cellobiose. BGL from *Aspergillus oryzae*, *Aspergillus aculeatus*, *Penicillium decumbens*, and *Saccharomycopsis fibuligera* enzymes showed the highest specific activity. In FIG. 14, a lower enzyme loading was tested. BGL from *Aspergillus aculeatus* and *Aspergillus oryzae* showed the highest specific activity at the lower enzyme loading.

[0258] The purified BGL enzymes were also tested for their activity on both pretreated hardwood solids and C5 liquor derived from pretreated hardwoods. FIG. 15 demonstrates that

the addition of small quantities of the BGLs increase hydrolysis rates. BGL from *Aspergillus oryzae* and *Aspergillus aculeatus* showed the highest hydrolysis rates relative to a control BGL from *Saccharomycopsis fibuligera*. In addition, BGL from *Aspergillus aculeatus* and *Humicola grisea* lead to the highest total yields at the end of hydrolysis. For FIG. 15, the purified BGLs were added along with *Saccharomycopsis fibuligera* BGL and were compared to a reaction where additional *Saccharomycopsis fibuligera* BGL was added. FIG. 15 indicates that the test BGL enzymes had hydrolysis rates greater than a reaction where additional *Saccharomycopsis fibuligera* BGL was added.

[0259] The BGL enzymes also improved hydrolysis of C5 oligomers from hardwoods. FIGs. 16 and 17 show the time course release of xylose and glucose, respectively, from C5 oligomers in an assay where the BGLs were added with other enzymes targeting hydrolysis of the oligomers, and where BGLs were used in place of *Saccharomycopsis* BGL. In contrast to the assays on hardwood solids, the enzyme mixtures in these assays utilized either the *Saccharomycopsis* BGL or the test BGLs at equal loadings. FIG. 16 shows that xylose release in the assay stayed constant, whereas glucose release increased by >35% for reactions where the new BGLs were included. In particular, inclusion of the *Aspergillus aculeatus* enzyme resulted in the highest yield of glucose.

[0260] FIG. 18 shows data collected from the same assay using the Biorad Aminex 87P column. This data also shows increases in glucose and mannose relative to control by the addition of the test BGL enzymes. The hydrolysis of glucose relative to acid hydrolysis increased from ~35% to ~50% (~40% increase), and the hydrolysis of mannose increased from ~42% to ~50% (a 16% increase) by adding the *Aspergillus aculeatus* BGL enzyme. Finally, several mixtures of BGLs were added to hydrolyze the C5 oligomers, and there was an increase in glucose release in all mixtures containing *Aspergillus aculeatus* and *Aspergillus oryzae* BGLs (FIG. 19).

[0261] These examples illustrate possible embodiments of the present invention. While the invention has been particularly shown and described with reference to some embodiments thereof, it will be understood by those skilled in the art that they have been presented by way of example only, and not limitation, and various changes in form and details can be made therein without departing from the spirit and scope of the invention. Thus, the breadth and scope of the present invention should not be limited by any of the above-described exemplary embodiments, but should be defined only in accordance with the following claims and their equivalents.

We Claim:

1. A recombinant yeast host cell comprising a polynucleotide encoding a heterologous beta-glucosidase (BGL) having
 - (i) an amino acid sequence at least 90% or at least 95% identical to SEQ ID NO: 7; or
 - (ii) an amino acid sequence at least 90% or at least 95% identical to a fragment of SEQ ID NO: 7 without a signal peptide, wherein said polynucleotide is codon-optimized for expression in the yeast host cell, wherein the codon adaption index (CAI) of the codon-optimized polynucleotide corresponds to from 0.8 to 1.0.
2. The host cell of claim 1, wherein the polynucleotide is at least 90% identical to SEQ ID NO: 9.
3. The host cell of claim 1, wherein the polynucleotide is at least 95% identical to SEQ ID NO: 9.
4. The host cell of claim 1, wherein the polynucleotide is identical to SEQ ID NO: 9.
5. The host cell of claim 1, wherein the signal peptide comprises an amino acid sequence identical to SEQ ID NO: 8.
6. The host cell of any one of claims 1 to 5, further comprising one or more additional polynucleotides encoding a heterologous cellulase.
7. The host cell of claim 6, wherein the heterologous cellulase is a xylanase, xylosidase, acetylxy lanesterase (AXE), endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase, cellobiohydrolase (CBH), or combinations thereof.
8. The host cell of claim 7, wherein the CBH is CBH1.
9. The host cell of claim 7, wherein the CBH is CBH2.

10. The host cell of any one of claims 1 to 4, wherein the BGL is *Aspergillus aculeatus* (*A. aculeatus*) BGL.
11. The host cell of claim 7, wherein the endoglucanase is *A. fumigatus* endoglucanase I, *N. fischeri* endoglucanase III, *T. reesei* endoglucanase I, or *C. formosanus* endoglucanase I.
12. The host cell of claim 7, wherein the CBH is *T. emersonii* cellobiohydrolase I, *C. lucknowense* cellobiohydrolase IIb or *T. reesei* cellobiohydrolase II.
13. The host cell of claim 7, further comprising a polynucleotide encoding *S. fibuligera* BGL.
14. The host cell of claim 7, further comprising one or more polynucleotides encoding *T. emersonii* CBH1, *T. reesei* CBD, *C. lucknowense* CBH2, *A. fumigatus* EG1, *N. fischeri* EG3, *S. fibuligera* BGL or *A. niger* xylanase.
15. The host cell of claim 7, further comprising one or more polynucleotides encoding *A. niger* xylanase, *P.t.r.* xylosidase, *N. fischeri* AXE, *A. fumigatus* EG1, *T. reesei* AGL1, *T. reesei* beta-mannanase, *A. fumigatus* alpha-glucuronidase (FC110), *A. fumigatus* acetyl esterase (FC136), *N. fischeri* beta-mannosidase (FC124), or *S. fibuligera* BGL.
16. The host cell of any one of claims 1 to 15, wherein the host cell can saccharify crystalline cellulose.
17. The host cell of claim 16, wherein the host cell can ferment the crystalline cellulose.
18. The host cell of any one of claims 1 to 15, wherein the host cell can hydrolyze hardwood solids or C5 liquor derived from hardwoods.
19. The host cell of any one of claims 1 to 18, wherein the host cell produces the BGL in an amount of at least 0.6 mg in culture.
20. The host cell of any one of claims 1 to 18, wherein the host cell produces the BGL in a concentration of at least 0.2 mg/ml in culture.

21. The host cell of any one of claims 1 to 20, wherein the yeast is *Saccharomyces cerevisiae*, *Saccharomyces pastorianus*, *Saccharomyces bayanus*, *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus*, *Schizosaccharomyces pombe*, *Schwanniomyces occidentalis*, or derivatives thereof.

22. The host cell of claim 21, wherein the yeast is *Saccharomyces cerevisiae*.

23. A co-culture comprising (i) a host cell of any one of claims 1 to 22 and (ii) a second host cell comprising one or more polynucleotides encoding a xylanase, xylosidase, AXE, endoglucanase, alpha-galactosidase, glucosidase, mannanase, alpha-glucuronidase, acetyl esterase, beta-mannosidase, glucuronyl esterase or CBH.

24. A method for hydrolyzing a cellulosic substrate, comprising contacting the cellulosic substrate with a host cell of any one of claims 1 to 22, or the co-culture of claim 23.

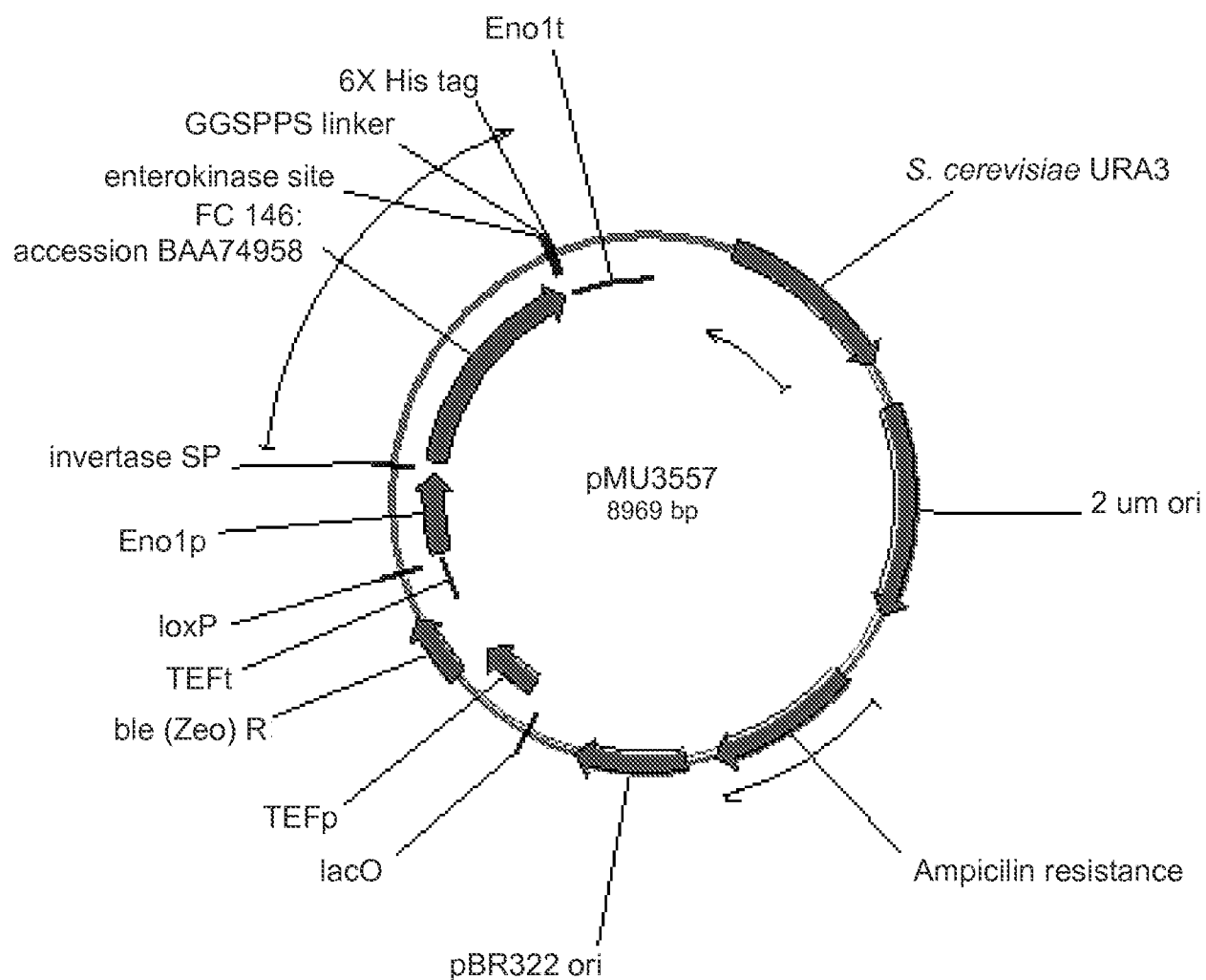
25. The method of claim 24, wherein the cellulosic substrate comprises a lignocellulosic biomass.

26. The method of claim 25, wherein the lignocellulosic biomass is grass, switch grass, cord grass, rye grass, reed canary grass, miscanthus, sugar-processing residues, sugarcane bagasse, agricultural wastes, rice straw, rice hulls, barley straw, corn cobs, cereal straw, wheat straw, canola straw, oat straw, oat hulls, corn fiber, stover, soybean stover, corn stover, forestry wastes, recycled wood pulp fiber, paper sludge, sawdust, hardwood, softwood, or combinations thereof.

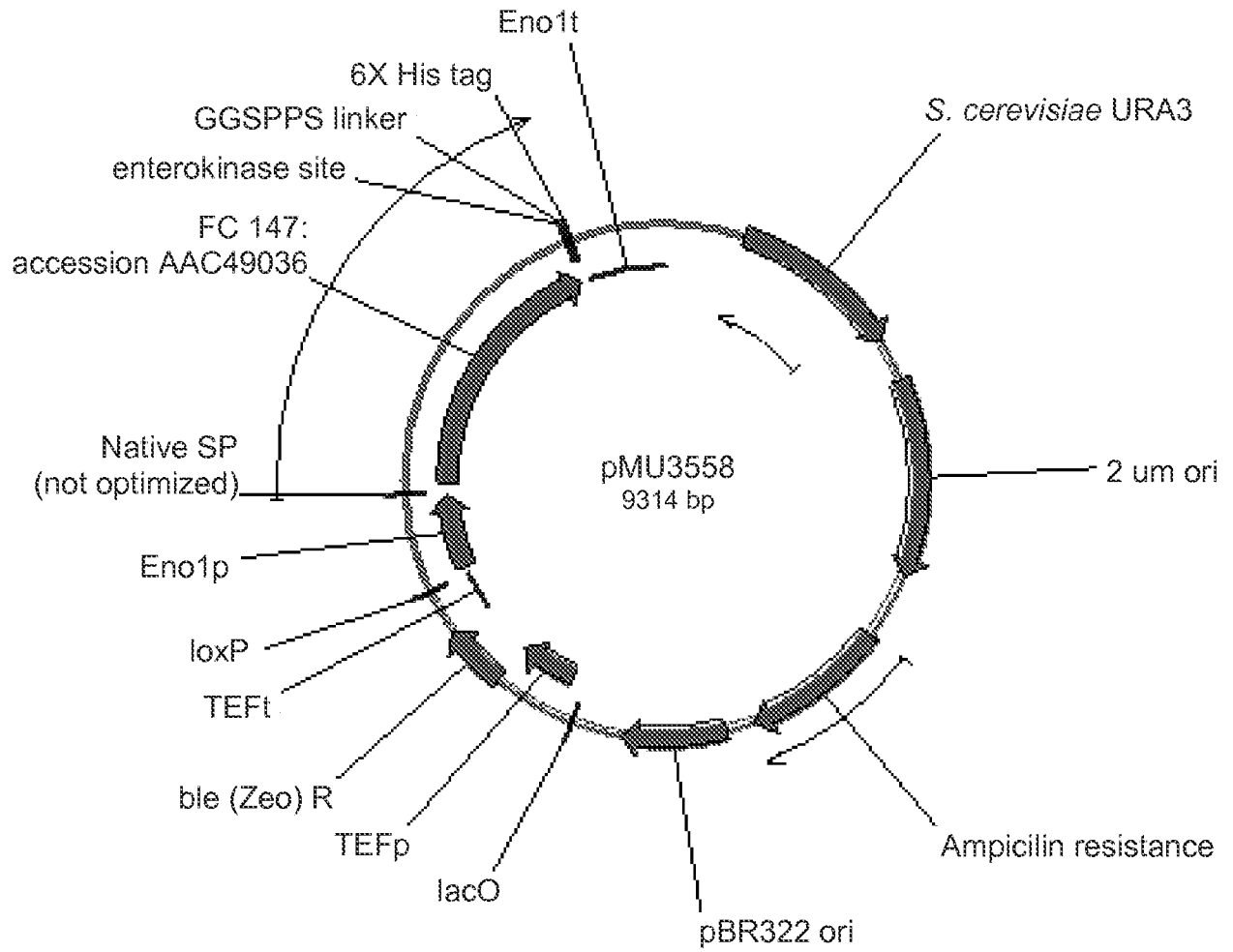
27. The method of claim 25, wherein the cellulosic substrate is hydrolyzed to xylose, glucose, mannose, galactose, arabinose, or combinations thereof.

28. The method of claim 25, wherein the cellulosic substrate is hydrolyzed to xylose, glucose, mannose, galactose or arabinose at a rate at least 10% greater than the rate of a host cell comprising a polynucleotide encoding a BGL from *S. fibuligera*.

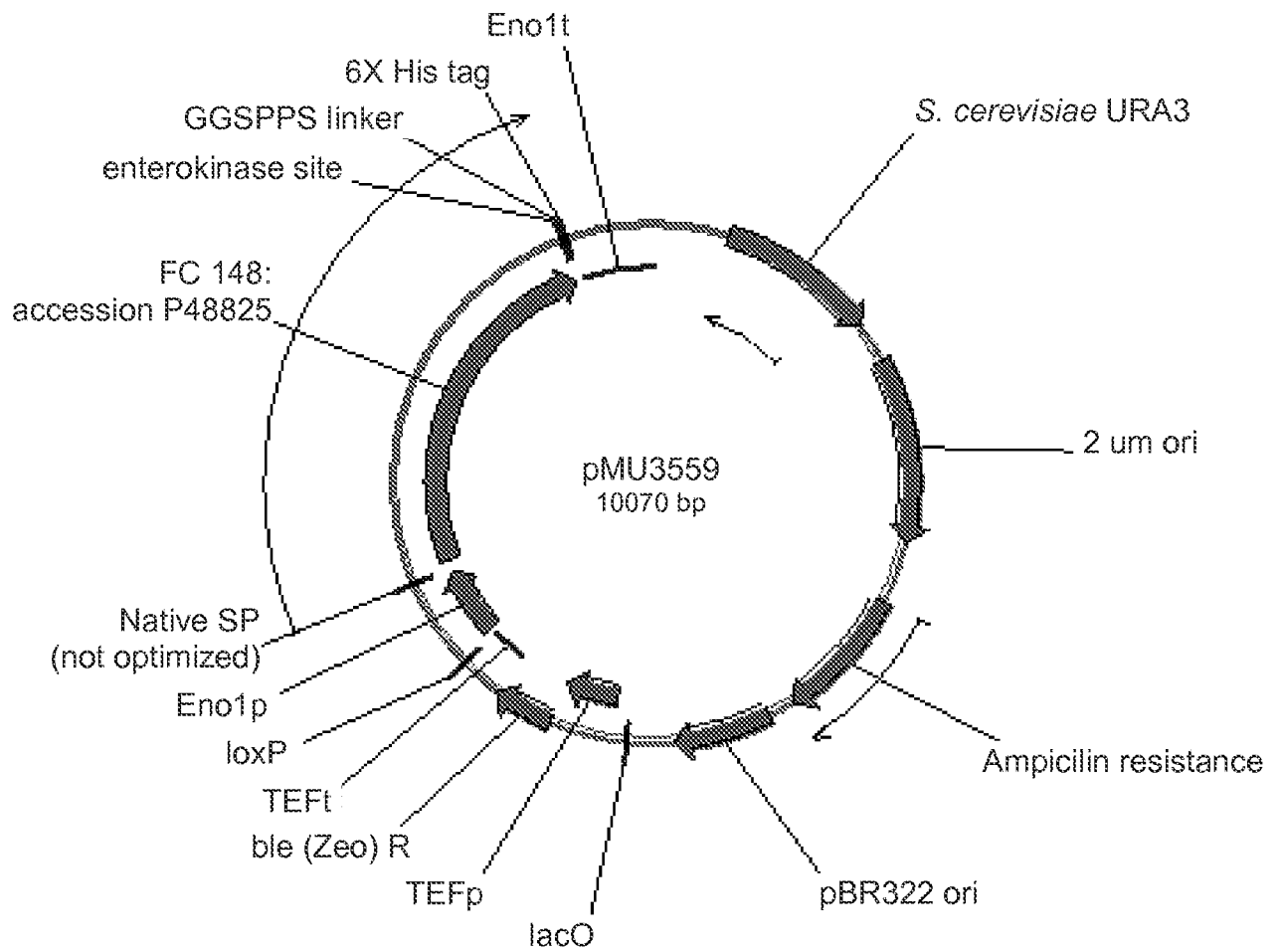
29. The method of claim 26, wherein the BGL is present in an amount of 0.2 mg or less per gram of xylose.
30. A method of fermenting cellulose, comprising culturing the host cell of any one of claims 1 to 22, or the co-culture of claim 23 in medium that contains crystalline cellulose under suitable conditions for a period sufficient to allow saccharification and fermentation of the cellulose.
31. The method of claim 30, wherein ethanol is produced.

**FIG. 1**

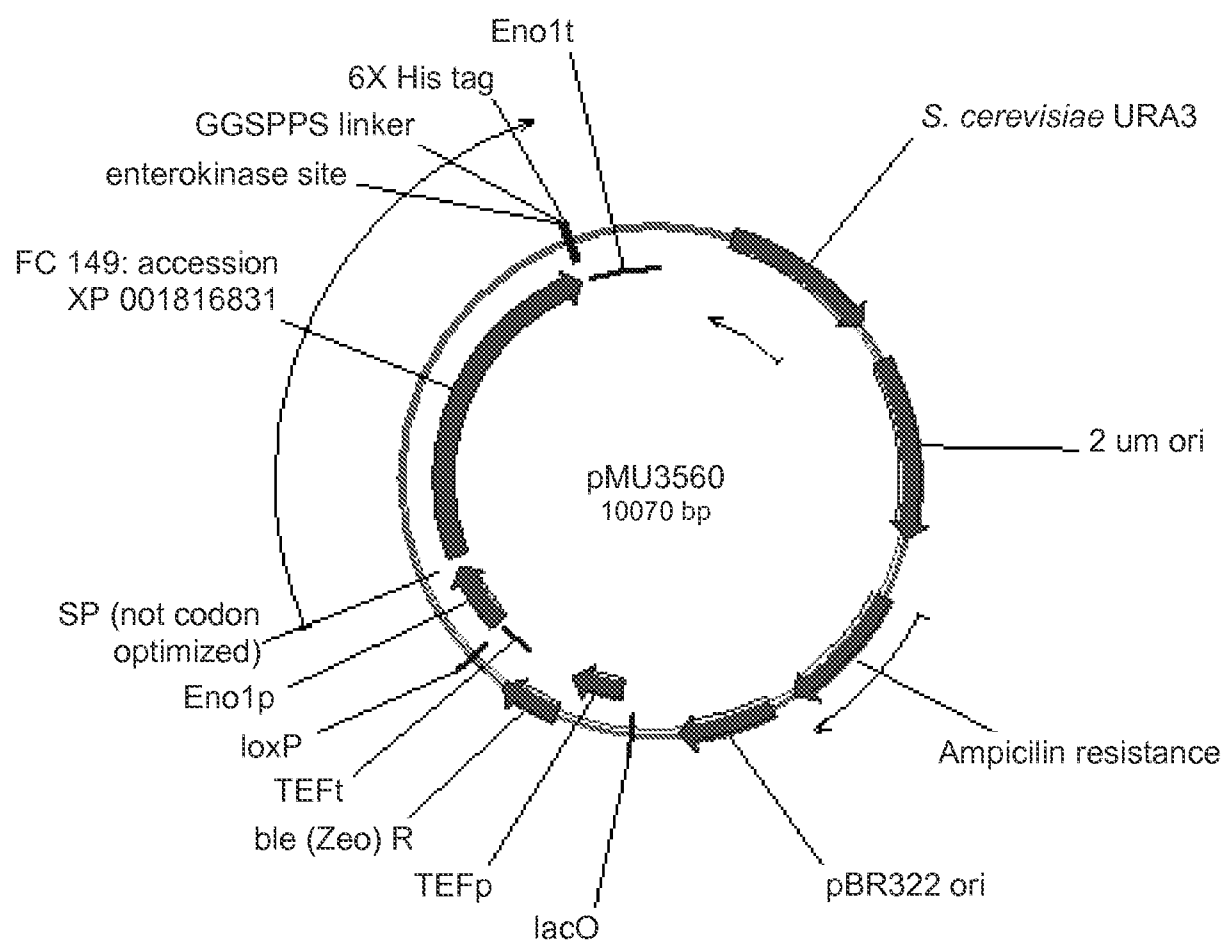
2/21

**FIG. 2**

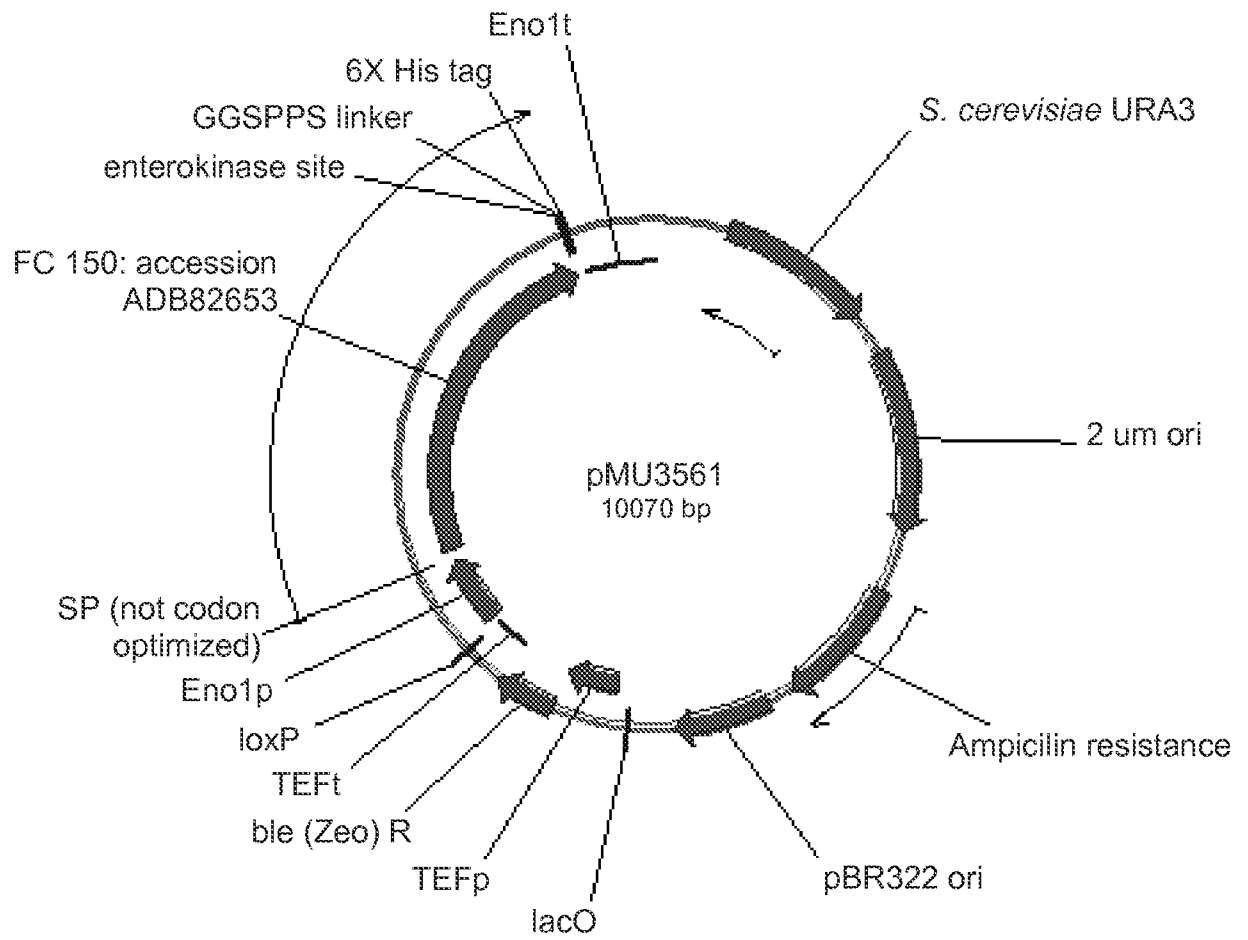
3/21

**FIG. 3**

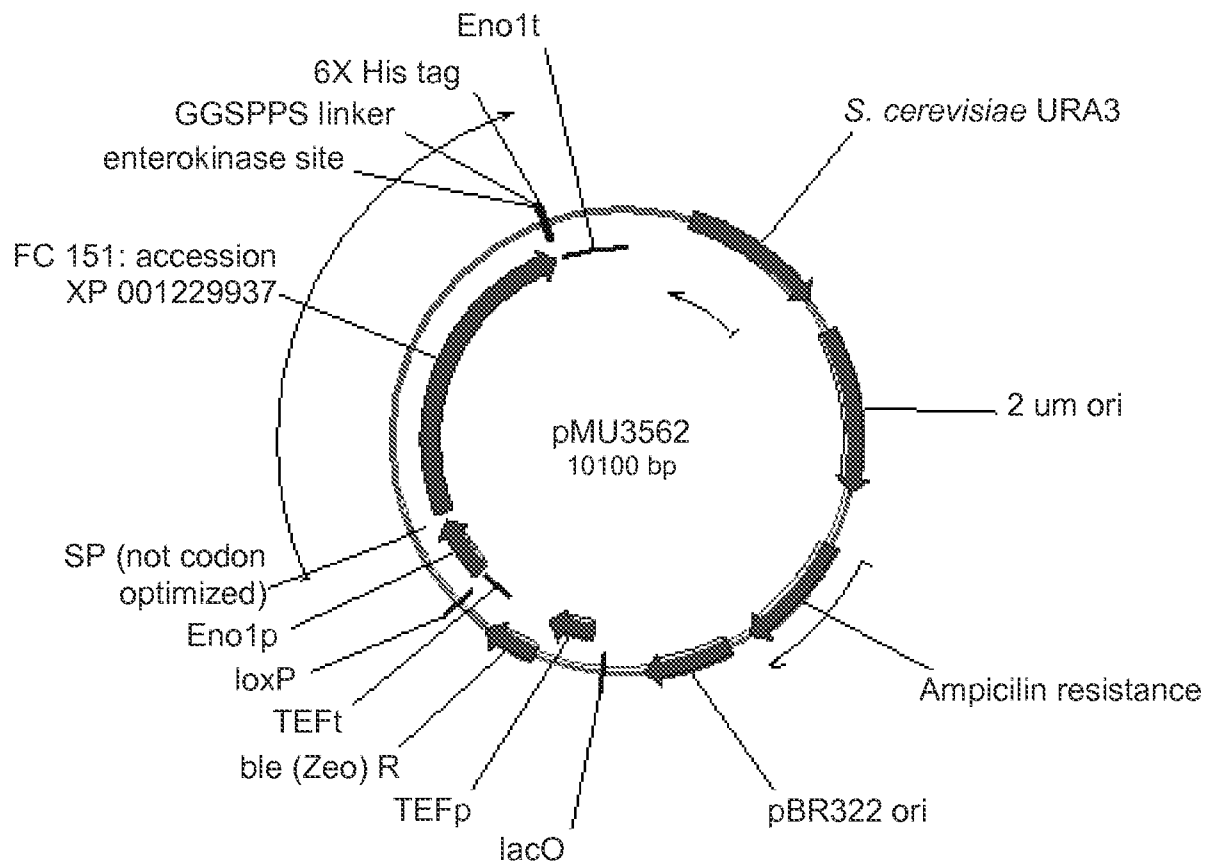
4/21

**FIG. 4**

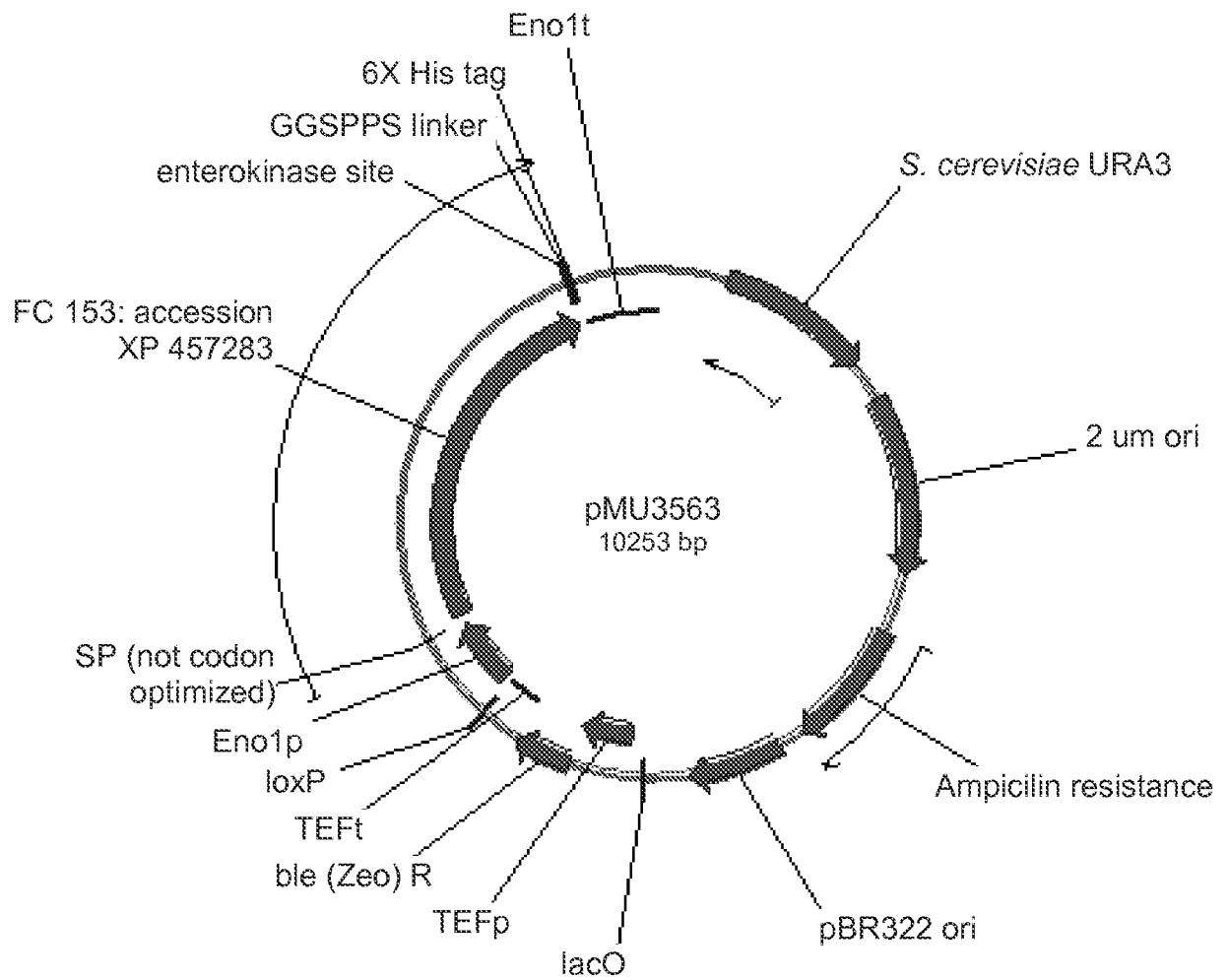
5/21

**FIG. 5**

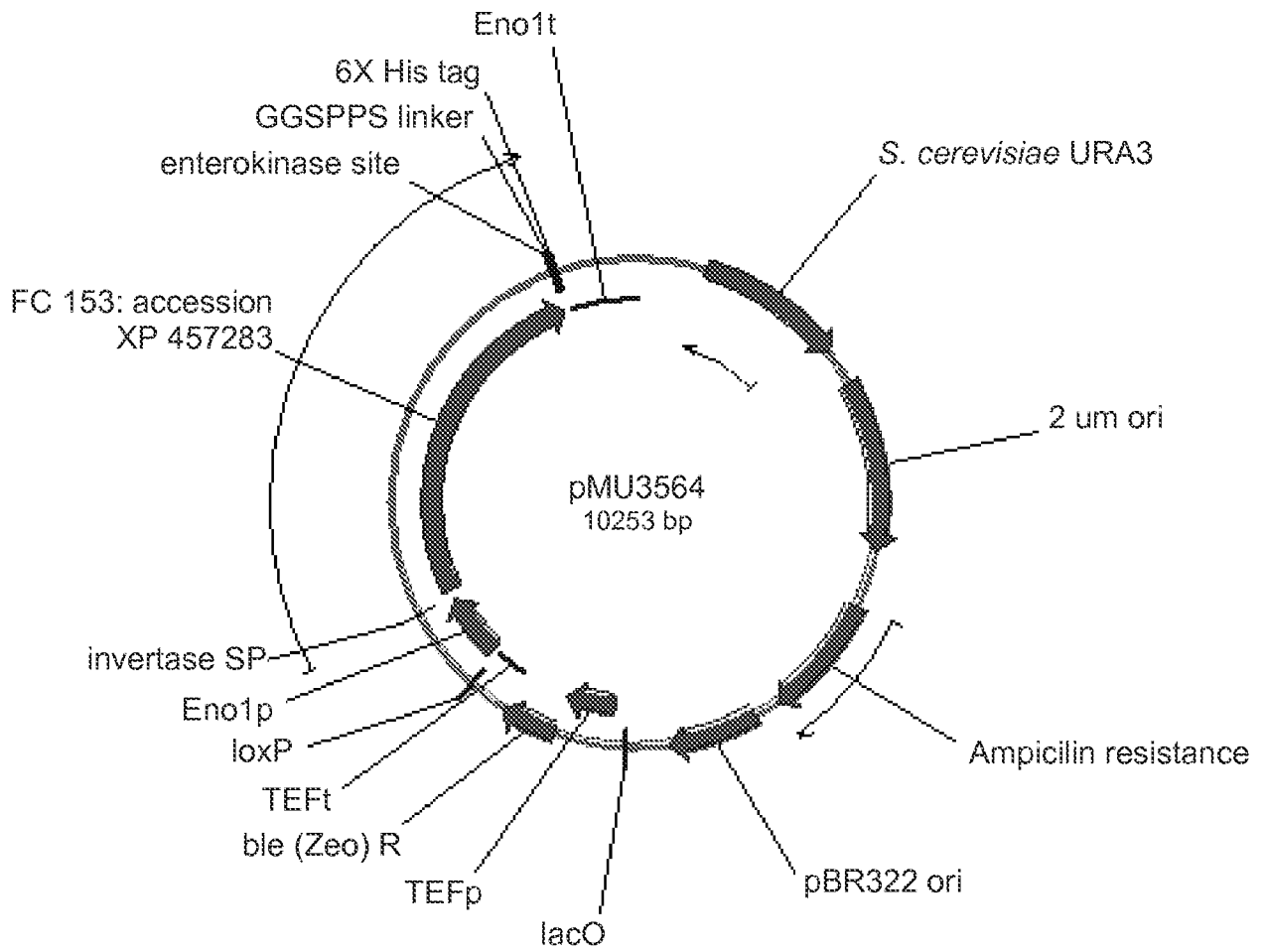
6/21

**FIG. 6**

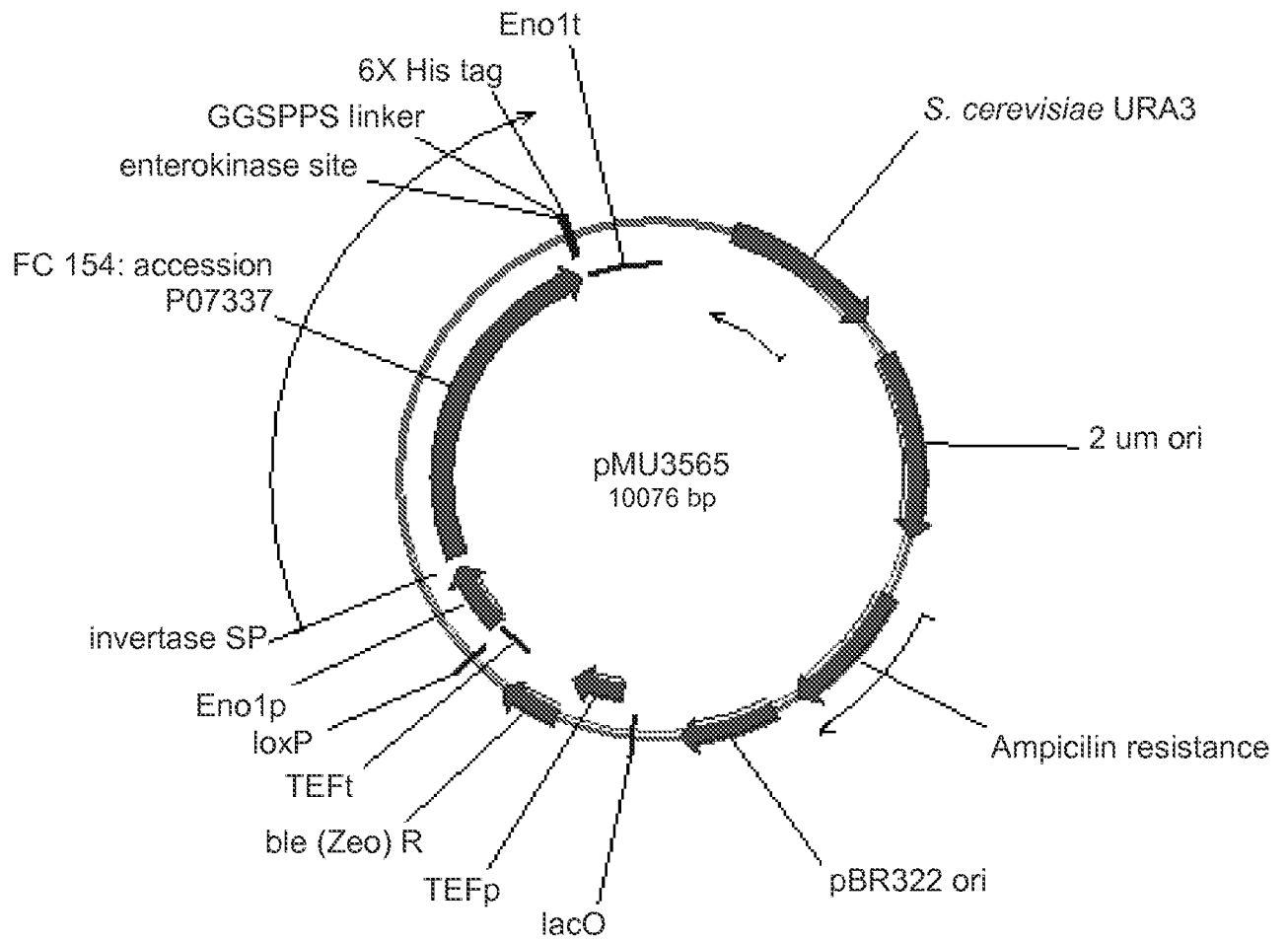
7/21

**FIG. 7**

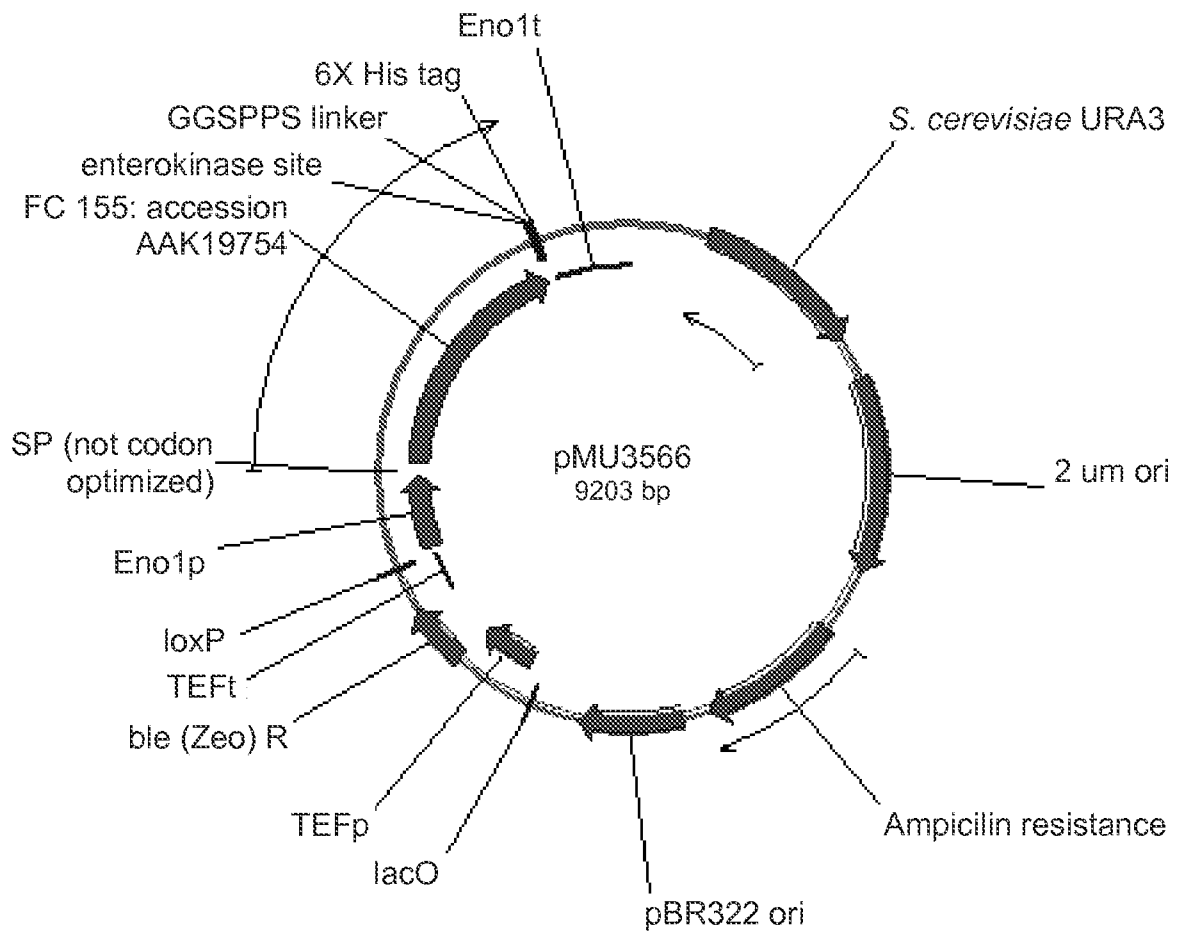
8/21

**FIG. 8**

9/21

**FIG. 9**

10/21

**FIG. 10**

11/21

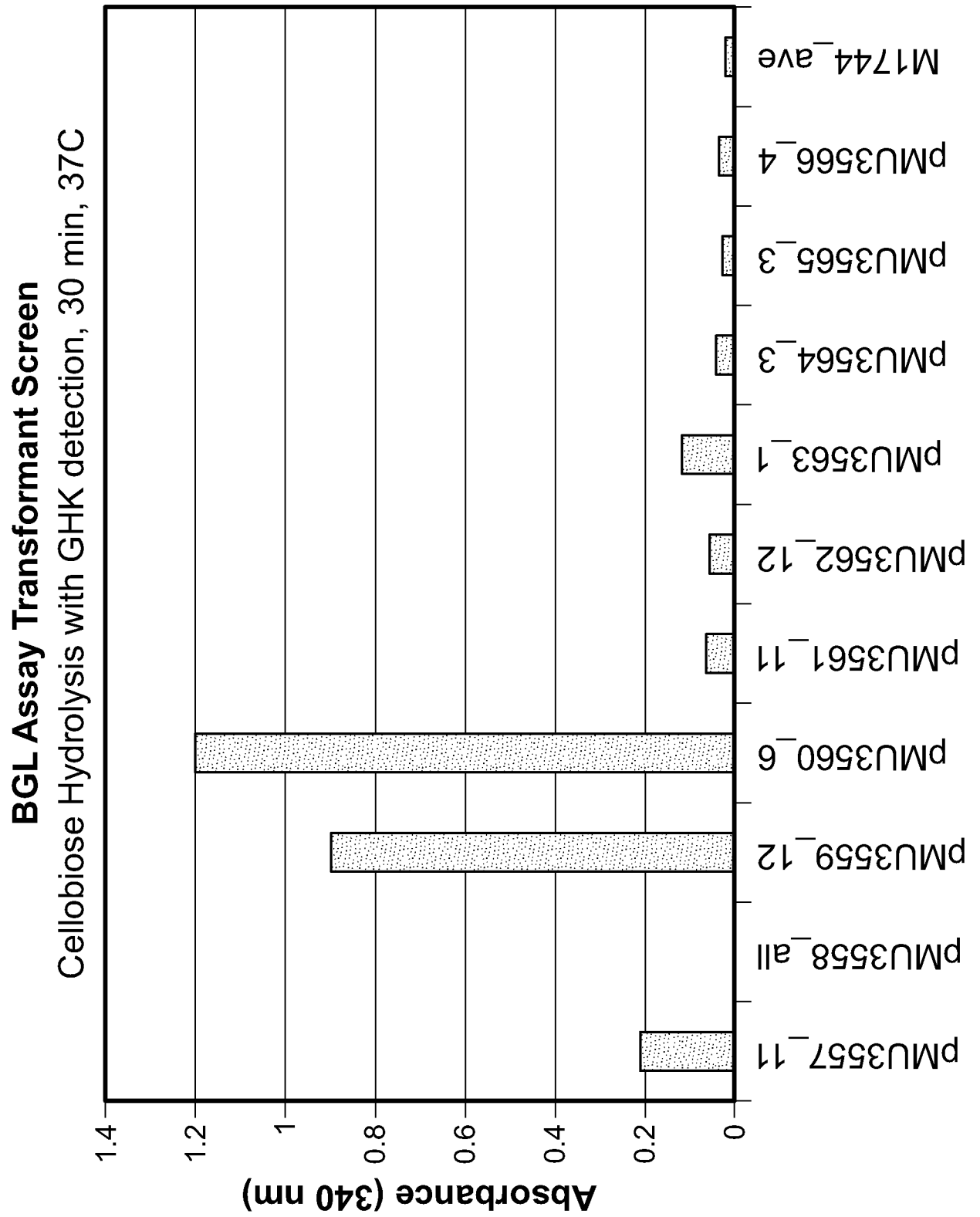


FIG. 11

12/21

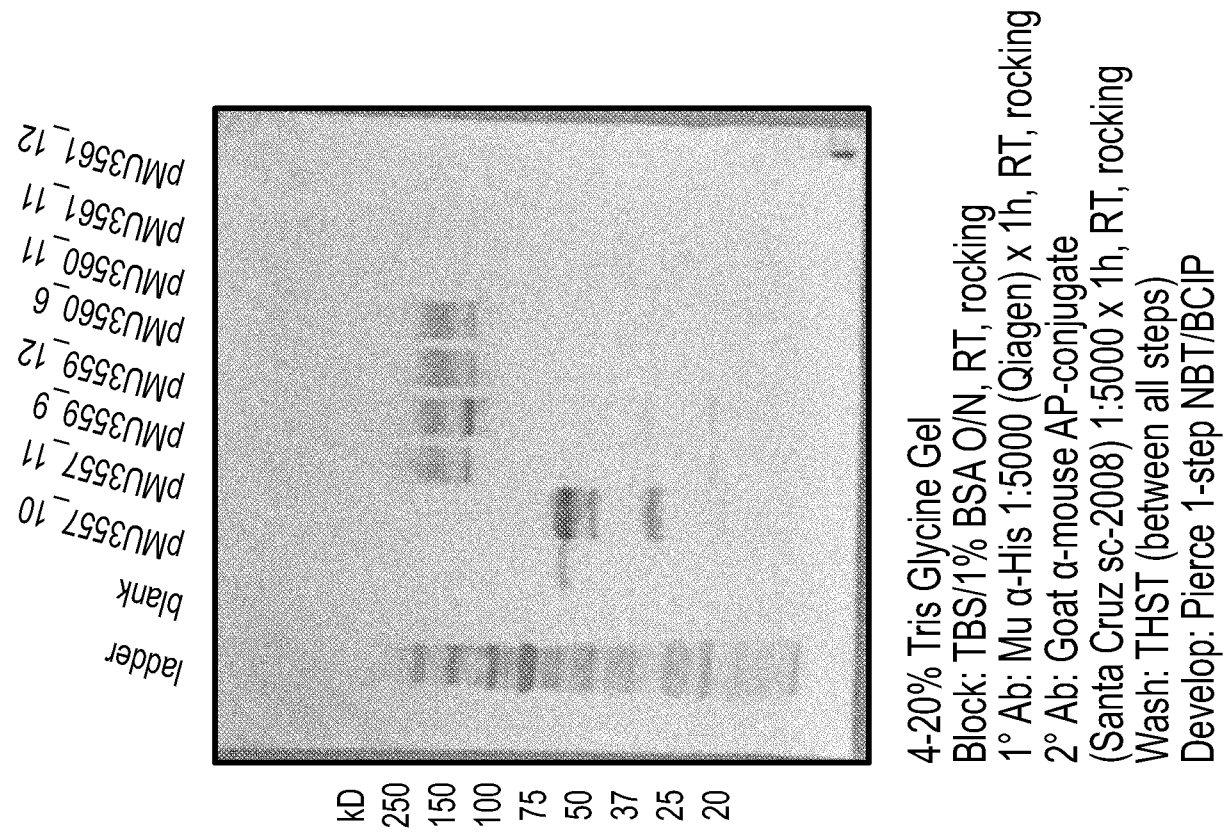
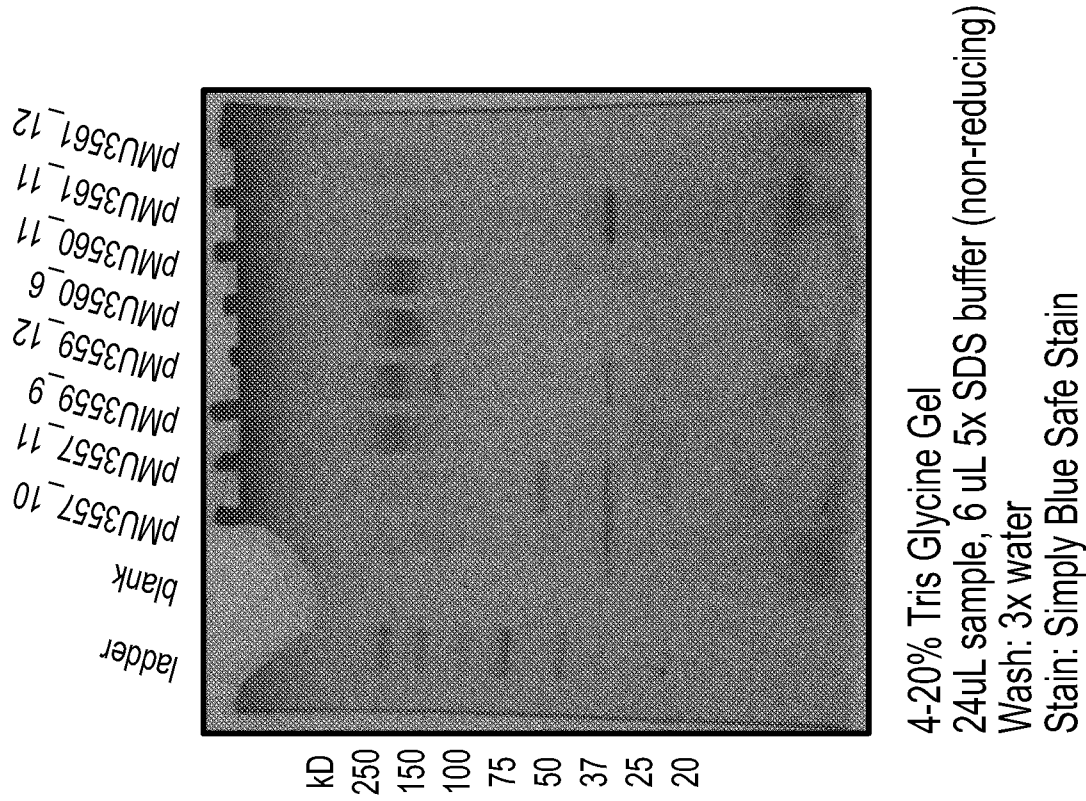


FIG. 12A



13/21

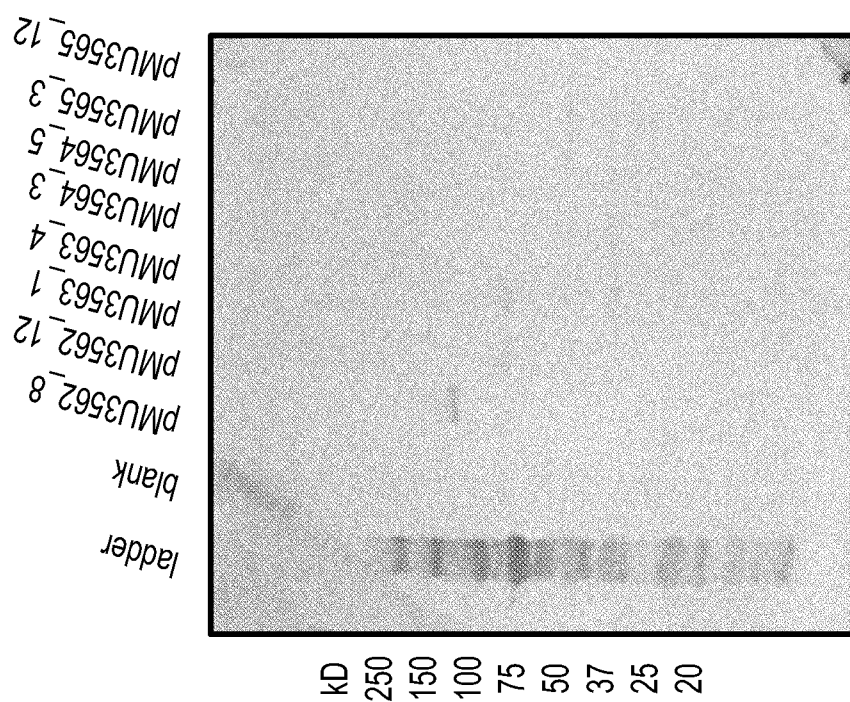
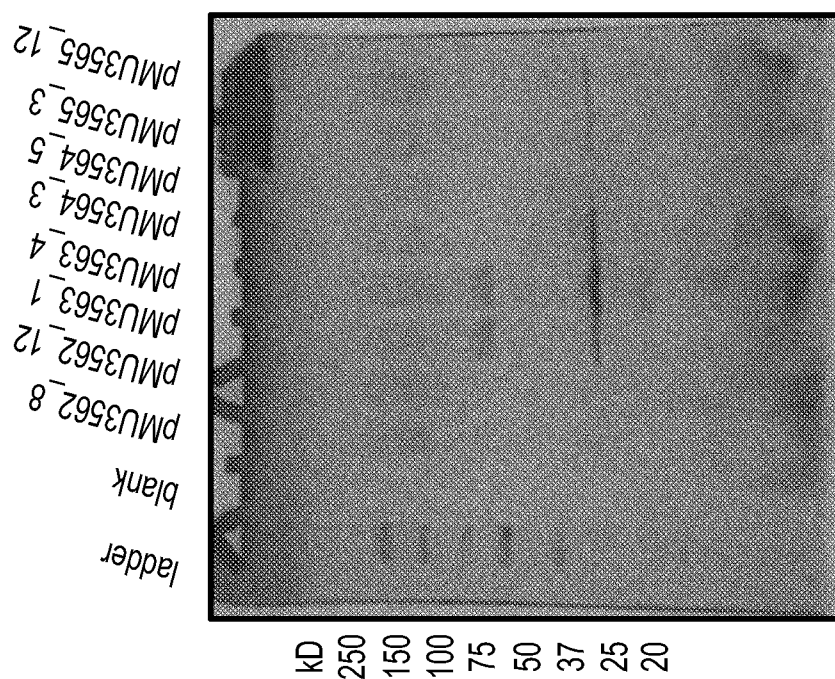


FIG. 12B



14/21

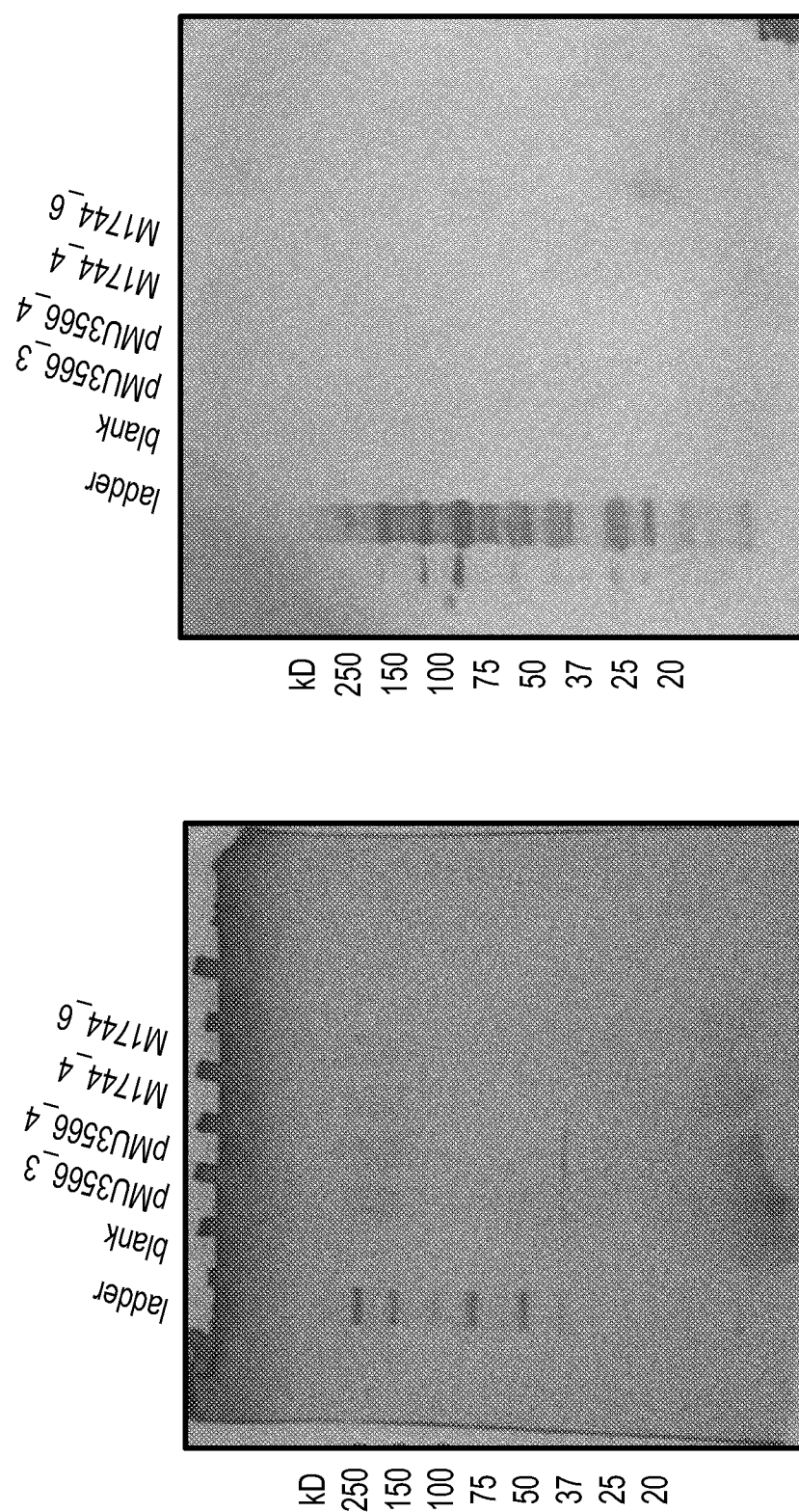


FIG. 12C

15/21

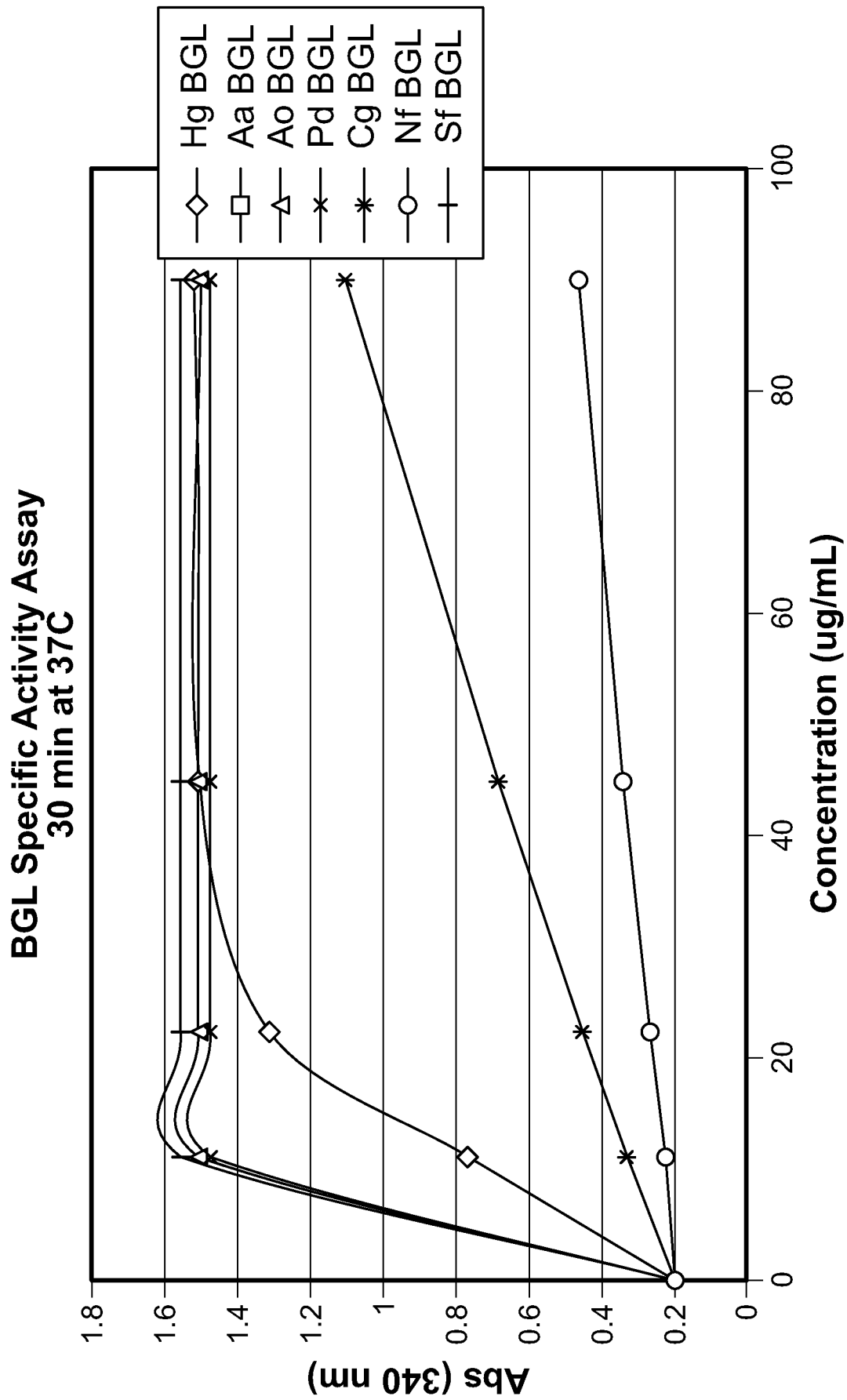


FIG. 13

16/21

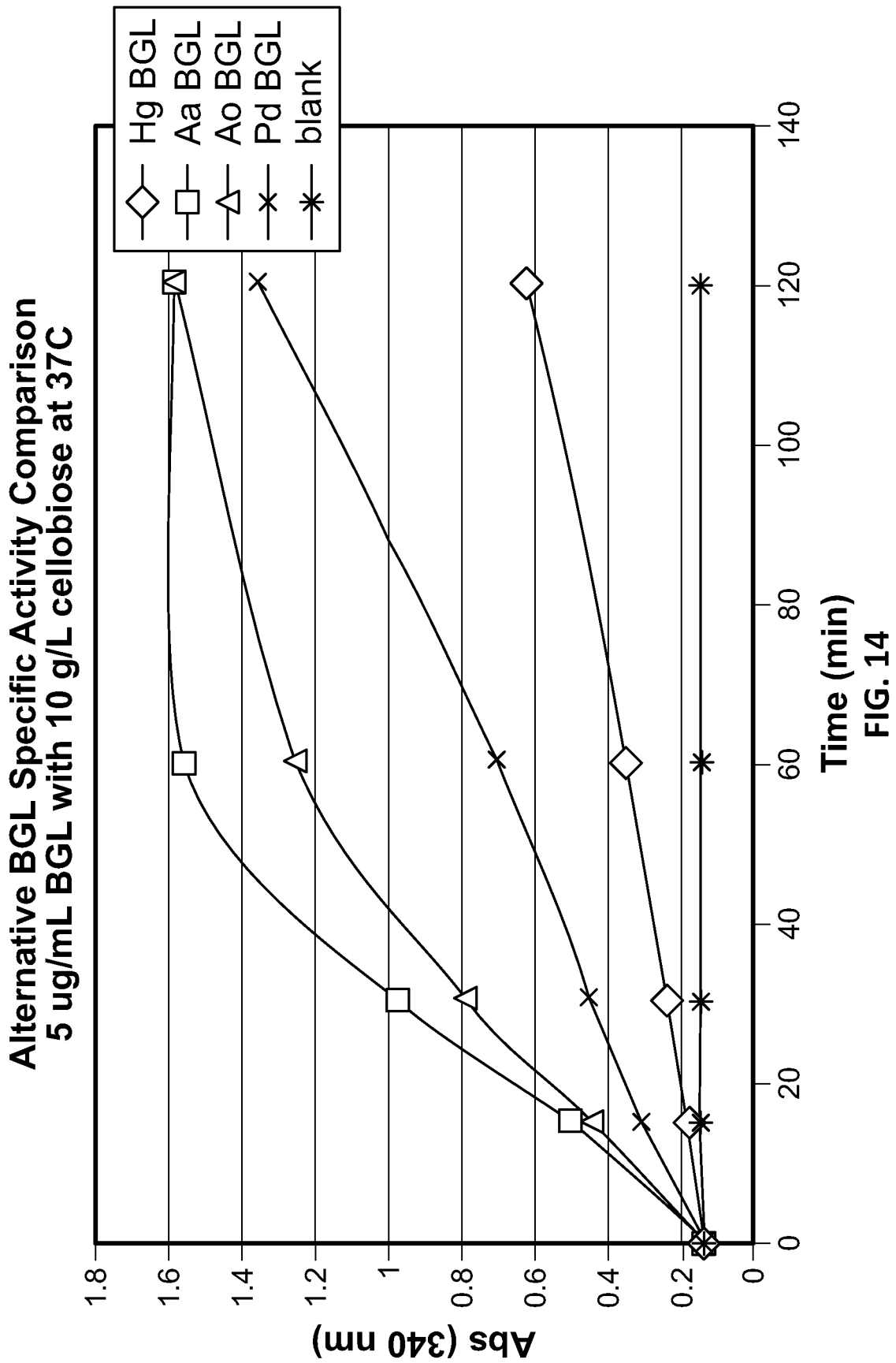


FIG. 14

17/21

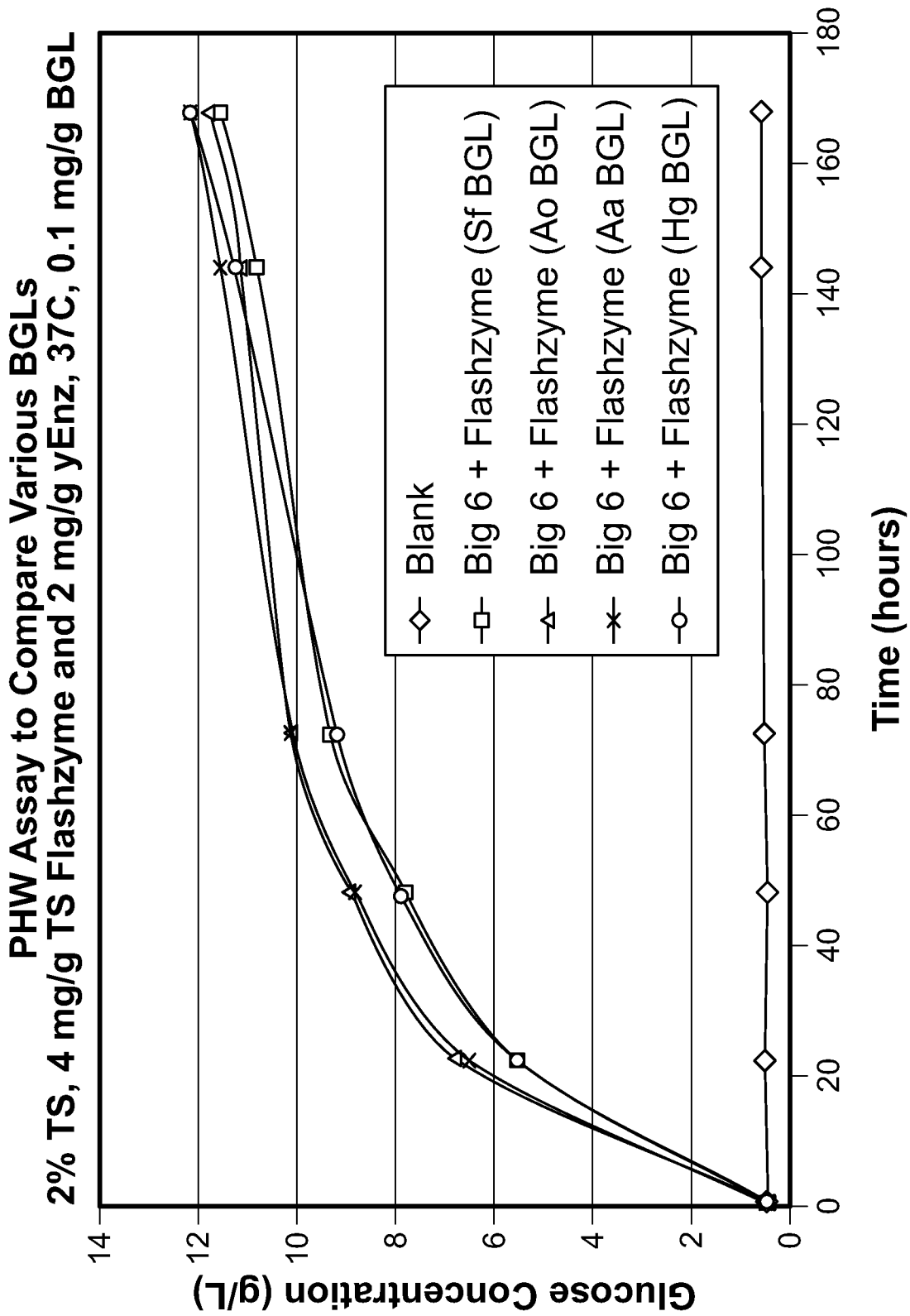


FIG. 15

18/21

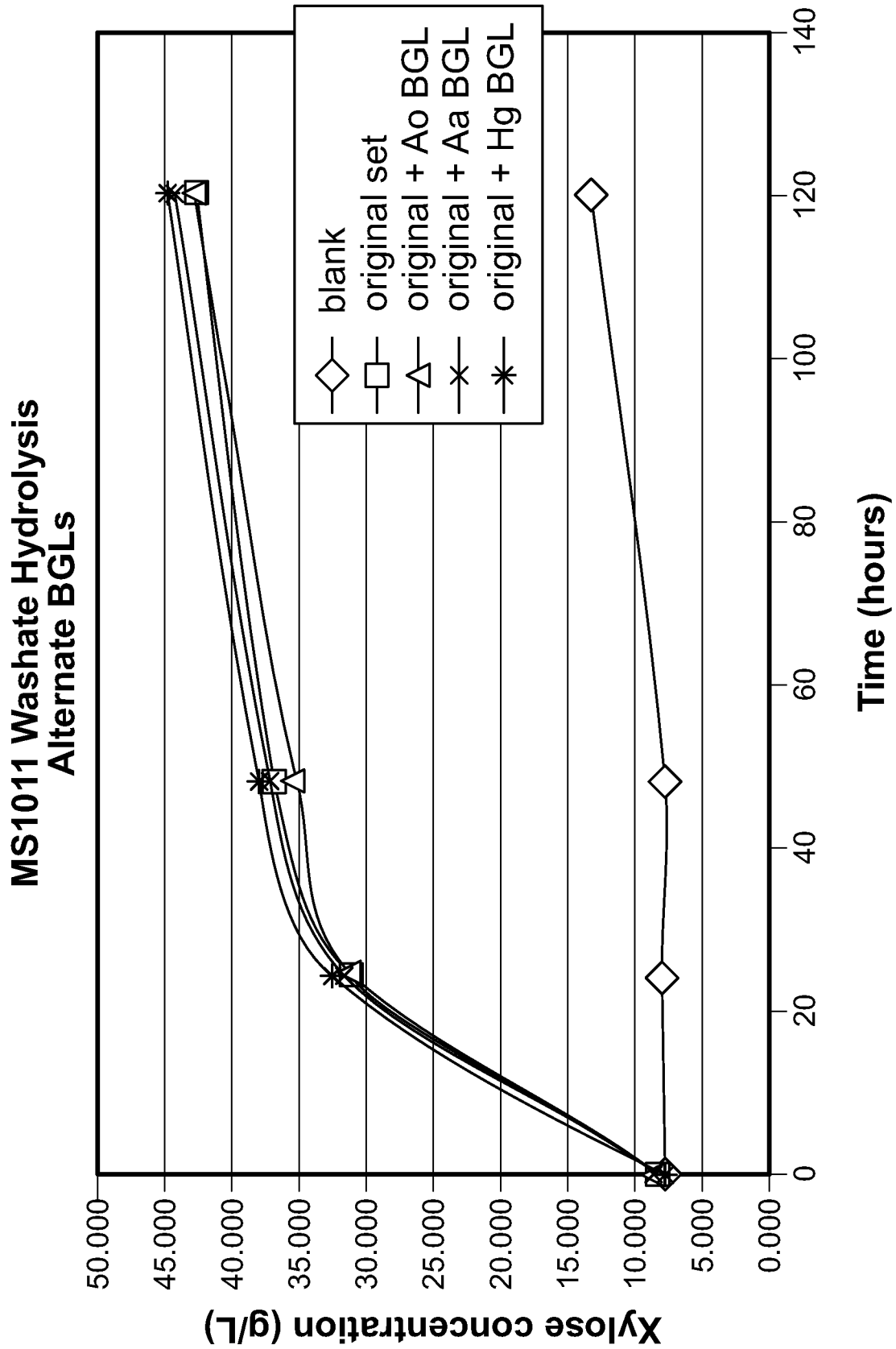


FIG. 16

19/21

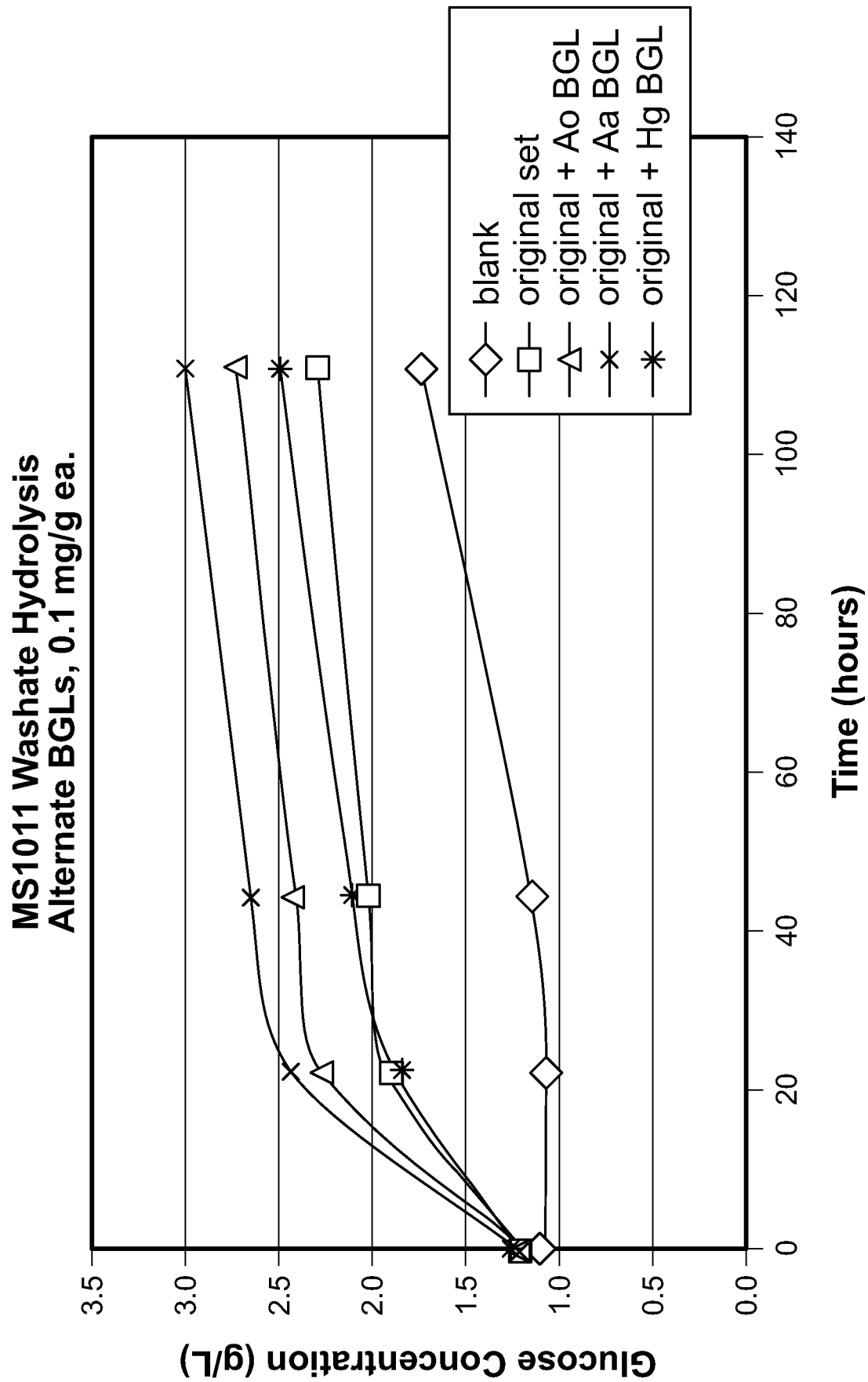


FIG. 17

20/21

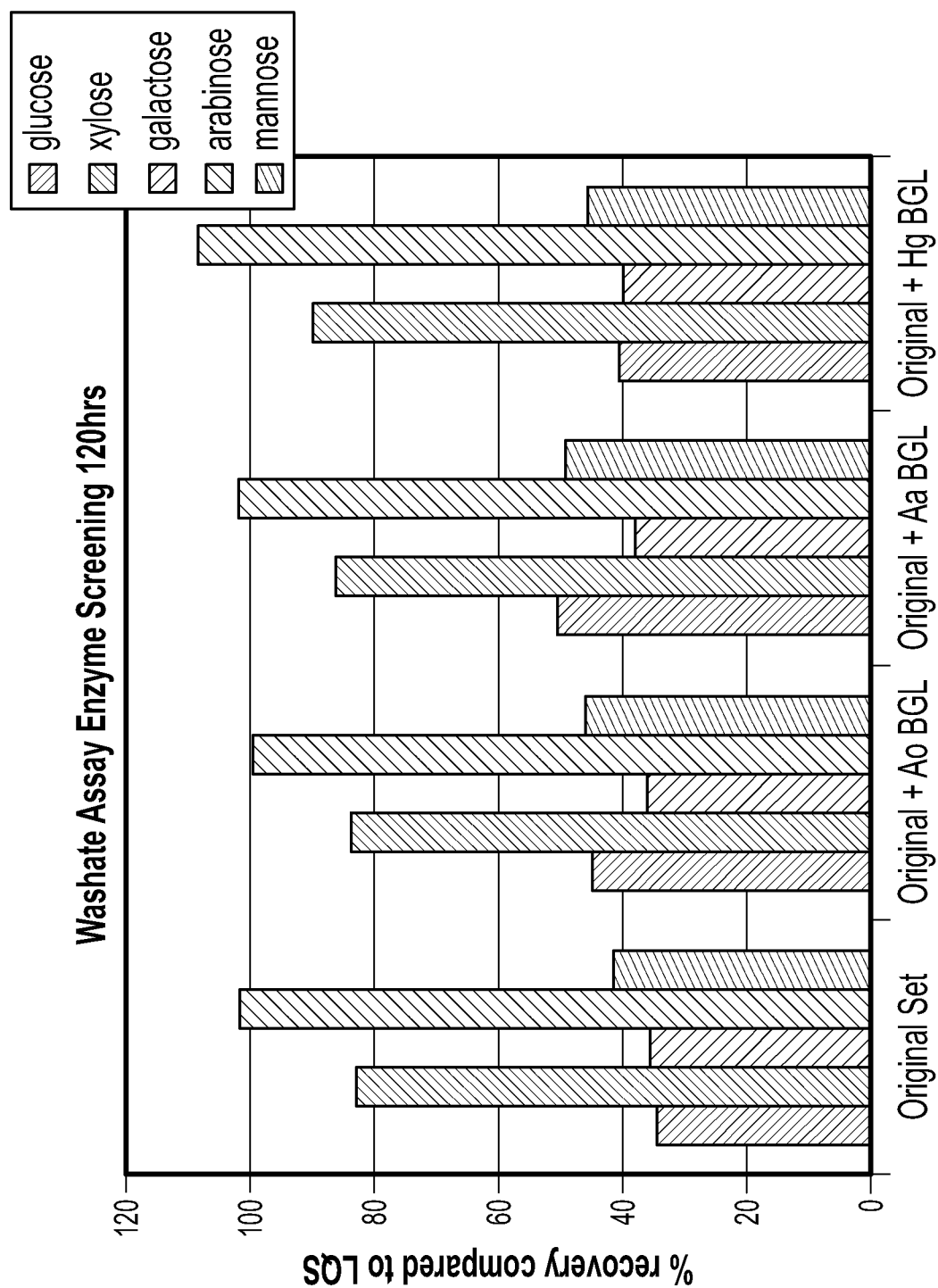


FIG. 18

