



(86) **Date de dépôt PCT/PCT Filing Date:** 2009/05/11
(87) **Date publication PCT/PCT Publication Date:** 2009/11/19
(45) **Date de délivrance/Issue Date:** 2018/09/25
(85) **Entrée phase nationale/National Entry:** 2010/11/10
(86) **N° demande PCT/PCT Application No.:** US 2009/002902
(87) **N° publication PCT/PCT Publication No.:** 2009/139839
(30) **Priorité/Priority:** 2008/05/11 (US61/052,214)

(51) **Cl.Int./Int.Cl. C12N 9/42** (2006.01),
C12P 7/10 (2006.01)
(72) **Inventeurs/Inventors:**
MCBRIDE, JOHN, US;
LYND, LEE, US;
DELEAULT, KRISTEN, US
(73) **Propriétaire/Owner:**
LALLEMAND HUNGARY LIQUIDITY MANAGEMENT
LLC, HU
(74) **Agent:** GOWLING WLG (CANADA) LLP

(54) **Titre : CONSTRUCTION DE SOUCHES DE LEVURES PROTOTROPHIQUES/CELLULOLYTIQUES EXPRIMANT DES
CELLULASES ANCREES ET SECRETEES**
(54) **Title: CONSTRUCTION OF PROTOTROPHIC/CELLULOYTIC YEAST STRAINS EXPRESSING TETHERED AND SECRETED
CELLULASES**

(57) **Abrégé/Abstract:**

The present invention is directed to the construction of prototrophic, cellulo lytic strains of *Saccharomyces cere-visiae* with tethered and secreted cellulases and selection-based improvement of growth on cellulose by these strains. In some em-bodiments, host cells of the invention are able to produce ethanol using crystalline cellulose as a sole carbon source.

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

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
19 November 2009 (19.11.2009)

PCT

(10) International Publication Number
WO 2009/139839 A8(51) International Patent Classification:
C12N 9/42 (2006.01) *C12P 7/10* (2006.01)(21) International Application Number:
PCT/US2009/002902(22) International Filing Date:
11 May 2009 (11.05.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/052,214 11 May 2008 (11.05.2008) US(71) Applicant (for all designated States except US): **MAS-COMA CORPORATION** [US/US]; 67 Etna Road, Suite 300, Lebanon, NH 03766 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **MCBRIDE, John** [US/US]; 174 Dartmouth College Highway, Lebanon, NH 03768 (US). **LYND, Lee** [US/US]; 181 Bean Road, Neri-den, NH 03770 (US). **DELEAULT, Kristen** [US/US]; 409 Fernwood Farms Road, Canaan, NH 03741 (US).(74) Agents: **JACKMAN, Peter, A.** et al.; STERNE, KESSLER, GOLDSTEIN & FOX P.L.L.C., 1100 New York Avenue, N.w., Washington, DC 20005-3934 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

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

(48) Date of publication of this corrected version:

19 August 2010

(15) Information about Correction:

see Notice of 19 August 2010

(54) Title: CONSTRUCTION OF PROTROPHIC/CELLULOYTIC YEAST STRAINS EXPRESSING TETHERED AND SECRETED CELLULASES

(57) Abstract: The present invention is directed to the construction of prototrophic, cellulolytic strains of *Saccharomyces cerevisiae* with tethered and secreted cellulases and selection-based improvement of growth on cellulose by these strains. In some embodiments, host cells of the invention are able to produce ethanol using crystalline cellulose as a sole carbon source.

WO 2009/139839 A8

CONSTRUCTION OF PROTROPHIC/CELLULOYTIC YEAST STRAINS EXPRESSING TETHERED AND SECRETED CELLULASES

BACKGROUND OF THE INVENTION

[0001] The potential of plant biomass as a cheap and renewable substrate for the production of fuel and chemicals has gained considerable interest in recent years. The biological saccharification of cellulose, the main component of plant biomass, is of particular interest in the field of fuel ethanol production. At least four biologically mediated process steps are involved in the current cellulose-to-ethanol technology: (i) cellulose enzyme production; (ii) enzymatic saccharification of cellulose; (iii) fermentation of hexose sugars (end-products of cellulose hydrolysis); and (iv) fermentation of pentose sugars (end-products of hemicellulose hydrolysis) to ethanol. Lynd, L. R. *et al.*, "Microbial cellulose utilization: fundamentals and biotechnology," *Microbiol. Mol. Biol. Rev.* 66:506-577 (2002) Combining the four process steps above into a one-step conversion of cellulose to fuel ethanol (termed consolidated bioprocessing (CBP)) would result in a considerable reduction in processing costs *See id.*

[0002] The yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) has superior ethanol formation properties, but is noncellulolytic. The expression of cellulases in *S. cerevisiae* would be a prerequisite for cellulose conversion via CBP. *S. cerevisiae* has received a great deal of interest regarding heterologous protein expression as well as the production of ethanol and other commodity products. *See id.*; Romanos, M. A. *et al.*, "Foreign gene expression in yeast: a review," *Yeast* 8:423-488 (1992) Expression of a functional cellulase system in *S. cerevisiae* would require the co-expression of at least three groups of enzymes, namely endoglucanases (EC 3.2.1.4); exoglucanases (EC 3.2.1.91) and β -glucosidases (EC 3.2.1.21). These enzymes act synergistically to efficiently degrade cellulose Mansfield, S.D. and R. Meder, "Cellulose hydrolysis – the role of the mono-component cellulases in crystalline cellulose degradation," *Cellulose* 10, 159–169 (2003)

[0003] Various cellulase genes have been expressed in *S. cerevisiae* with the aim of direct ethanol production from cellulose. Often, however, heterologous cellulase enzymes are produced by recombinant organisms in such low concentrations that the amount of saccharified substrate available is unable to sustain growth of the organisms. In an attempt to alleviate enzyme concentration deficiencies, yeast strains displaying cell

surface proteins have been developed. Fujita, Y *et al.*, "Direct and Efficient Production of Ethanol from Cellulosic Material with a Yeast Strain Displaying Cellulolytic Enzymes," *Applied and Environmental Microbiology* 68: 5136-5141 (2002) describes an *S. cerevisiae* strain expressing tethered β -glucosidase I (BglI) and endoglucanase II (EgII). However, this strain, while able to grown on a linear, soluble polysaccharide, is unable to grown on insoluble cellulose.

- [0004] Improvements on such strains have been described and characterized, where the expression of four different tethered cellulase enzymes result in a strain having the capability of growth on insoluble cellulose. See, U.S. Application, entitled "Recombinant Yeast Strains Expressing Tethered Cellulase Enzyme," to McBride *et al.*, filed November 20, 2007, and assigned to Dartmouth University. In this previously-described strain, tethered versions of endoglucanase I (EgI), cellobiohydrolase I (Cbh1), and cellobiohydrolase II (Cbh2) from *Trichoderma reesei* (*T. reesei*) and the β -glucosidase I (BglI) from *Saccharomycopsis fibuligera* (*S. fibuligera*) were used to transform *S. cerevisiae*. This tethered EgI/Cbh1/Cbh2/BglI transformed yeast strain was capable of growth on the insoluble cellulose substrate phosphoric acid swollen cellulose (PASC) and the crystalline insoluble cellulose substrate bacterial microcrystalline cellulose (BMCC).
- [0005] Given that both the EgI and Cbh1 from the *T. reesei* have cellulose binding domains (CBDs) at their carboxy terminus, and that this is also where they are attached to the anchoring domain, these tethered constructs may not, however, necessarily provide sufficient activity on insoluble substrates. Additionally, *T. reesei* Cbh1 is typically not well secreted. While a codon optimized version may be somewhat improved, evidence suggests the improvement is not large if at all. Finally, tethered cellulase enzymes may not gain the access to the substrate that secreted versions do for steric reasons. Thus, there is a need in the art to improve such tethered cellulase enzyme systems.
- [0006] An additional approach to increase cellulose conversion via CBP in *S. cerevisiae* is to improve cellulose utilization by selection-based methods. Selection-based improvement of strains, including yeast strains, for improving cellulose utilization promises to be a powerful tool for engineering recombinant organisms for consolidated bioprocessing. However, to date, no demonstration of this technique has been accomplished.

- [0007] Previous attempts to create strains built for selection experiments were not suitable for further experiments. This is due, in part, to the inability to separate the effect of amino acid utilization from cellulose utilization, and, in part, to the slow rate of growth rate of the previous strains which rendered them unsuitable for continuous culture because those strains were likely to wash out of the continuous culture at elevated dilution rates.
- [0008] The solutions to these issues could come from a number of sources. First, prototrophic versions of these strains could be created, because these versions allow media to be formulated without adding any amino acids. When this is done, it can be calculated that the total carbon available to the cell in synthetic complete media (Yeast Nitrogen Base without amino acids from Difco) is 1.9 mg/L, all of which is present in vitamin components. This virtually eliminates concerns about the utilization of non-cellulose carbon sources during continuous cultures.
- [0009] In addition to strain modification, an easier way to hydrolyze substrate other than AvicelTM PH105 is desired. However, such substrates are generally not available in large quantities, and producing them is prohibitively time consuming. Additionally, AvicelTM PH105 is easy to work with in well mixed systems and does not, for example, clog tubing. One other solution to the issue of slow growth rate and low cell concentration is to add soluble sugar as a co-feed in the system. This co-feed allows the cells to replicate at a relatively high rate and yet still to gain a selective benefit by being cellulolytic, since the soluble sugar concentration in the reactor can be kept very close to zero.
- [0010] The present invention addresses the limitations of the systems described above. First, with regard to the improvement of tethered cellulase systems, the present invention provides for a transformed host cell with greater ability to grow on insoluble cellulose by the addition of a highly expressed, secreted Cbh1 to the Eg1/Cbh1/Cbh2/Bgl1 tethered system.
- [0011] In addition, with regard to the selection-based approach, the present invention provides for a selection method and the creation of a new cellulolytic, prototrophic strain of *S. cerevisiae* utilizing this selection method. The new strain exhibits a number of phenotypic improvements with respect to cellulose utilization, including improved growth on mixes of AvicelTM and cellobiose, improved growth on bacterial microcrystalline cellulose (BMCC)-containing media, and biomass formation on solid AvicelTM containing

media. Improved strains of the present invention attained cell counts on BMCC containing media about ten times faster than previously created strains.

BRIEF DESCRIPTION OF THE INVENTION

[0012] In some embodiments of the present invention a transformed host cell is provided which comprises: (a) at least one heterologous polynucleotide which encodes an endoglucanase which when expressed is tethered to the cell surface; (b) at least one heterologous polynucleotide which encodes a cellobiohydrolase which when expressed is tethered to the cell surface and (c) at least one heterologous polynucleotide which encodes a β -glucosidase which when expressed is tethered to the cell surface, wherein said transformed host cell further comprises a heterologous polynucleotide which encodes at least one additional endoglucanase, cellobiohydrolase, or β -glucosidase which when expressed, is secreted by the cell.

[0013] In other embodiments of the present invention, a host cell with the ability to saccharify cellulose and produce ethanol therefrom is provided. In these embodiments, the host cell comprises a tethered endoglucanase, a tethered cellobiohydrolase, a tethered β -glucosidase and additionally comprising at least one secreted endoglucanase, cellobiohydrolase, or β -glucosidase. In some embodiments, the cellulose is crystalline cellulose.

[0014] In yet another embodiment of the present invention, a method of fermenting cellulose is disclosed. In these embodiments the host cells are transformed with: (a) at least one heterologous polynucleotide which encodes an endoglucanase which when expressed is tethered to the cell surface; (b) at least one heterologous polynucleotide which encodes a cellobiohydrolase which when expressed is tethered to the cell surface and (c) at least one heterologous polynucleotide which encodes a β -glucosidase which when expressed is tethered to the cell surface, wherein said transformed host cell further comprises a heterologous polynucleotide which encodes at least one additional endoglucanase, cellobiohydrolase, or β -glucosidase which when expressed, is secreted by the cell.

[0015] In still other embodiments of the present invention, a recombinant host cell which is capable of producing ethanol when grown using crystalline cellulose as the sole carbon source is provided.

[0016] In yet other embodiments of the present invention, a method of improving the ability of a host cell to use cellulose as a carbon source is provided. These methods comprise: (a) culturing said host cell in media containing cellulose; (b) maintaining the culture conditions in a substantially steady state; (c) allowing variant progeny of the original host cell to acquire a selective advantage in the culture according to the ability of said variant progeny to display increased reproductive capacity in the cellulose-containing media; (d) growing the cells selected in step (c) on media containing cellulose; and, (e) iteratively repeating steps (b) and (c) until a variant of the original cell is produced, wherein said variant has acquired the ability to grow at least 2 fold faster than the original host cell on cellulose as a sole carbon source.

[0017] In other embodiments, a host cell able to use cellulose as a sole carbon source is produced by a process comprising: (a) culturing said host cell in media containing cellulose; (b) maintaining the culture conditions in a substantially steady state whereby variant progeny of the original host cell acquire a selective advantage in the culture according to the ability of said variant progeny to display increased reproductive capacity in the cellulose-containing media; (c) continuously growing the cells of step b on media containing cellulose; and, (d) repeating the selection of steps (b) and (c) until a variant of the original cell is produced, wherein said variant acquires the ability to grow to a cell density at least 2 fold greater than the original, pre-selected host cell on cellulose as a sole carbon source.

Other embodiments of the present invention relate to method of fermenting cellulose using the host cell defined herein, the method comprising culturing the transformed host cell in medium that contains crystalline cellulose under suitable conditions for a period sufficient to allow saccharification and fermentation of the cellulose.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] Figure 1 depicts a graph showing growth of constructed strains on Avicel™ PH105 in aerobic shake flasks. Black circles are M0360, Blue diamonds are M0149, Yellow triangles are M0359, and Brown squares are M0361.

- 5a -

[0019] Figure 2 depicts a graph showing washout of cells at a dilution rate of 0.02hr^{-1} . Calculated u_{max} based on this data for strain with all secreted cellulases is 0.012 hr^{-1} , and for the strain with tethered and secreted cellulases is 0.013 hr^{-1} . Blue diamonds are M0359, and Yellow triangles are M0360.

[0020] Figure 3 depicts a graph showing ethanol production from PASC by a number of yeast strains with YNB and amino acids as media components. Initial ethanol

concentration was subtracted from all subsequent data points, and negative values indicate that ethanol concentration dropped over time.

- [0021] Figure 4 depicts a graph showing ethanol production from BMCC by a number of yeast strains with YNB and amino acids as media components. Initial ethanol concentration was subtracted from all subsequent data points, and negative values indicate that ethanol concentration dropped over time.
- [0022] Figure 5 depicts a graph showing ethanol production from PASC by a number of yeast strains with YP and amino acids as media components. Initial ethanol concentration was subtracted from all subsequent data points, and negative values indicate that ethanol concentration dropped over time.
- [0023] Figure 6 depicts a comparison of parental (M0149) and strain with added secreted CBH1 (M0360) in 50mL shake flask culture on Avicel™ PH105 as the sole carbon source. CEN.PK 113-11C is included as the negative control.
- [0024] Figure 7 depicts data from long-term adaptation of M0360 in continuous culture where cellulose and glucose were co-fed. Dilution rate and glucose feed rate were adjusted at 200 hrs and 700 hrs. Data for CO₂ evolution was averaged over 3 residence times. Condition 1: $D=0.1 \text{ hr}^{-1}$, glucose feed rate= 0.2 g/L/hr ; Condition 2: $D=0.067 \text{ hr}^{-1}$, glucose feed rate: 0.38 g/L/hr ; Condition 3: $D=0.055 \text{ hr}^{-1}$, glucose feed rate= 0.11 g/L/hr (abbreviations: D = dilution rate).
- [0025] Figure 8 depicts the calculated conversion and mass balance from continuous culture presented in Figure 2. Conversion of cellulose was calculated based on the measured concentration of cellulose fed to the reactor, and the cell mass corrected total dry weight. The mass balance averaged $104\% \pm 5\%$. Condition 1: $D=0.1 \text{ hr}^{-1}$, glucose feed rate= 0.2 g/L/hr ; Condition 2: $D=0.067 \text{ hr}^{-1}$, glucose feed rate: 0.38 g/L/hr ; Condition 3: $D=0.055 \text{ hr}^{-1}$, glucose feed rate= 0.11 g/L/hr (abbreviations: D = dilution rate).
- [0026] Figure 9 depicts the cell concentration and CO₂ evolution rate data from glucose continuous culture with M0360. Dilution rate was 0.065 hr^{-1} and glucose feed rate was 0.11 g/L/hr .
- [0027] Figure 10 depicts the increase in cell concentration as a function of residence time in the reactor for the continuous culture data presented in Figures 7 and 8. Condition 2 ($D=0.067 \text{ hr}^{-1}$, glucose feed rate: 0.38 g/L/hr) was from 200 to 700 hrs, and Condition 3

- 7 -

($D=0.055\text{ hr}^{-1}$, glucose feed rate= 0.11 g/L/hr) was from 700 to 1500 hrs. The glucose control chemostat was run at $D=0.065\text{ hr}^{-1}$ and a glucose feed rate= 0.39 g/L/hr .

[0028] Figure 11 depicts batch growth experiment of selected and original strains on AvicelTM and cellobiose in shake flask cultures. The left panel (Figure 11A) shows the cell count data for growth of a population sample from the reactor run shown in Figure 7 as compared to the original strain. The right panel (Figure 11B) shows the number of cells formed per amount of cellobiose consumed during the experiment for 3 separate cultures for the selected (1 whole population, and 2 isolated colonies) and original strains (3 separate colonies). The initial cellobiose concentration for all batches was measured by HPLC to be $10.6 \pm 0.3\text{ g/L}$.

[0029] Figure 12 depicts growth of selected and original M0360 on BMCC as the sole carbon source. A number of strains were tested for growth kinetics on BMCC and the average number of new cells (cell count minus original count) was averaged for 3 cultures of the selected strain, and 2 cultures of the original strain. For the untransformed strain histidine and uracil were added to the media to account for the strains auxotrophies.

[0030] Figure 13 depicts growth test on 2% AvicelTM PH105 plate with YNB and complete amino acid mix added. The box at the bottom is where the control strain was streaked. 30 colonies from each of original and selected strains of M0360 were picked from YNB-glucose plates and streaked onto this plate. The colonies on the right hand-side were examined microscopically and were yeast.

DETAILED DESCRIPTION OF THE INVENTION

[0031] The disclosed methods and materials are useful generally in the field of engineered yeast.

Definitions

[0032] 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, and is usually in the form of a circular double-stranded DNA molecule. Such elements may be autonomously replicating sequences, genome integrating sequences, phage or nucleotide sequences, linear, circular, or supercoiled, of a single- or double-stranded DNA or RNA, derived from any source, in which a number of

nucleotide sequences 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. Preferably, the plasmids or vectors of the present invention are stable and self-replicating.

[0033] An "expression vector" is a vector that is capable of directing the expression of genes to which it is operably associated.

[0034] 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."

[0035] 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 cellobiohydrolase (CBH) domains include the catalytic domain (CD) and the cellulose binding domain (CBD).

[0036] 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 may be single-stranded or double-stranded. DNA includes cDNA, genomic DNA, synthetic DNA, and semi-synthetic DNA.

[0037] 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 may be 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).

[0038] 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.

[0039] 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, J., Fritsch, E. F. and Maniatis, T. *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.

[0040] 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. Preferably a minimum length for a hybridizable nucleic acid is at least about 15 nucleotides; more preferably at least about 20 nucleotides; and most preferably the length is at least 30 nucleotides. Furthermore, the skilled artisan will recognize that the temperature and wash solution salt concentration may be adjusted as necessary according to factors such as length of the probe.

[0041] 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 the case may be, as determined by the match between strings of such sequences.

[0042] 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.

[0043] "Identity" and "similarity" can be readily calculated by known methods, including but not limited to those described in: Computational Molecular Biology (Lesk, A. M., ed.) Oxford University Press, NY (1988); Biocomputing: Informatics and Genome Projects (Smith, D. W., ed.) Academic Press, NY (1993); Computer Analysis of Sequence Data,

Part I (Griffin, A. M., and Griffin, H. G., eds.) Humana Press, NJ (1994); Sequence Analysis in Molecular Biology (von Heinje, G., ed.) Academic Press (1987); and Sequence Analysis Primer (Gribskov, M. and Devereux, J., eds.) Stockton Press, NY (1991). Preferred methods to determine identity are designed to give the best match between the sequences tested. Methods to determine identity and similarity are codified in publicly available computer programs. Sequence alignments and percent identity calculations may be performed using the Megalign program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, Wis.). Multiple alignments of the sequences disclosed herein were performed using the Clustal method of alignment (Higgins and Sharp (1989) CABIOS. 5:151-153) with the default parameters (GAP PENALTY=10, GAP LENGTH PENALTY=10). Default parameters for pairwise alignments using the Clustal method were KTUPLE 1, GAP PENALTY=3, WINDOW=5 and DIAGONALS SAVED=5.

[0044] 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, or at least 250 amino acids.

[0045] 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 may 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.

Cellulase Systems Combining Tethered and Secreted Enzymes

[0046] As used herein, a protein is “tethered” to an organism’s cell surface if at least one terminus of the protein is bound, covalently and/or electrostatically for example, to the cell membrane or cell wall. It will be appreciated that a tethered protein may include one or more enzymatic regions that may be joined to one or more other types of regions at the nucleic acid and/or protein levels (e.g., a promoter, a terminator, an anchoring domain, a linker, a signaling region, etc.). While the one or more enzymatic regions may not be directly bound to the cell membrane or cell wall (e.g., such as when binding occurs via an anchoring domain), the protein is nonetheless considered a “tethered enzyme” according to the present specification.

[0047] Tethering may, for example, be accomplished by incorporation of an anchoring domain into a recombinant protein that is heterologously expressed by a cell, or by prenylation, fatty acyl linkage, glycosyl phosphatidyl inositol anchors or other suitable molecular anchors which may anchor the tethered protein to the cell membrane or cell wall of the host cell. A tethered protein maybe tethered at its amino terminal end or optionally at its carboxy terminal end.

[0048] As used herein, “secreted” means released into the extracellular milieu, for example into the media. Although tethered proteins may have secretion signals as part of their immature amino acid sequence, they are maintained as attached to the cell surface, and do not fall within the scope of secreted proteins as used herein.

[0049] As used herein, “flexible linker sequence” refers to an amino acid sequence which links a cell wall anchoring amino acid sequence with an amino acid sequence that contains the desired enzymatic activity. The flexible linker sequence allows for necessary freedom for the amino acid sequence that contains the desired enzymatic activity to have

reduced steric hindrance with respect to proximity to the cell and may also facilitate proper folding of the amino acid sequence that contains the desired enzymatic activity.

[0050] The present invention provides for a cellulase system, where the cellulase system is a host cell comprising: (a) at least one heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase which, when expressed, is tethered to the cell surface; (b) at least one heterologous polynucleotide comprising a nucleic acid which encodes a cellobiohydrolase which, when expressed, is tethered to the cell surface; (c) at least one heterologous polynucleotide comprising a nucleic acid sequence which encodes a β -glucosidase which, when expressed, is tethered to the cell surface; and (d) at least one additional heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase, cellobiohydrolase, or β -glucosidase which, when expressed, is secreted by the cell. Additional embodiments are directed to host cells comprising vectors containing polynucleotides as described above, as well as polypeptides encoded by the polynucleotides described above.

[0051] In certain embodiments, the cellulase system comprises two heterologous polynucleotides comprising nucleic acids encoding a cellobiohydrolase I and a cellobiohydrolase II.

[0052] In certain embodiments of the invention, the endoglucanase(s) can be an endoglucanase I or an endoglucanase II isoform, paralogue or orthologue. An "isoform" is a protein that has the same function as another protein but which is encoded by a different gene and may have small differences in its sequence. A "paralogue" is a protein encoded by a gene related by duplication within a genome. An "orthologue" is gene from a different species that has evolved from a common ancestral gene by speciation. Normally, orthologues retain the same function in the course of evolution as the ancestral gene.

[0053] In further embodiments, the endoglucanase expressed by the host cells of the present invention can be recombinant endo-1,4- β -glucanase. In particular embodiments, the endoglucanase is an endoglucanase I from *Trichoderma reesei*. In certain other embodiments, the endoglucanase is encoded by a polynucleotide sequence at least about 70, about 80, about 90, about 95, about 96, about 97, about 98, about 99, or 100% identical to a *T. reesei eg1* polynucleotide.

- [0054] In certain embodiments, the β -glucosidase is a β -glucosidase I or a β -glucosidase II isoform, paralogue or orthologue. In certain embodiments of the present invention the β -glucosidase is derived from *Saccharomycopsis fibuligera*. In particular embodiments, the β -glucosidase is encoded by a polynucleotide sequence at least about 70, about 80, about 90, about 95, about 96, about 97, about 98, about 99, or 100% identical to an *S. fibuligera bgl1* polynucleotide..
- [0055] In certain embodiments of the invention, the cellobiohydrolase(s) can be a cellobiohydrolase I and/or an cellobiohydrolase II isoform, paralogue or orthologue. In particular embodiments of the present invention the cellobiohydrolases are cellobiohydrolase I or II from *Trichoderma reesei*. In other embodiments, one cellobiohydrolase is tethered to the cell surface and an additional cellobiohydrolase is secreted into the extra-cellular milieu. In another embodiment, the β -glucosidase is encoded by a polynucleotide sequence at least about 70, about 80, about 90, about 95, about 96, about 97, about 98, about 99, or 100% identical to a *T. reesei cbh1* or *cbh2*.
- [0056] In certain embodiments the secreted cellobiohydrolase is encoded by a polynucleotide comprising a nucleic acid encoding *T. emersonii*, *H. grisea*, *T. aurantiacus* Cbh1 or Cbh2, or domain, fragment, variant, or derivative thereof, as described further below. In particular embodiments, the secreted cellobiohydrolase is encoded by a polynucleotide comprising a nucleic acid encoding *T. emersonii* Cbh1 or a *T. emersonii* Cbh1 fused to a domain of *T. reesei* Cbh1 or Cbh2, as described further below.
- [0057] In further embodiments the secreted cellobiohydrolase is a polypeptide comprising an amino acid sequence at least about 70, about 80, about 90, about 95, about 96, about 97, about 98, about 99, or 100% identical to an amino acid sequence encoding for a cellobiohydrolase listed in Tables 3-4 or 7.
- [0058] In certain aspects, the endoglucanase, cellobiohydrolase and β -glucosidase can be any suitable endoglucanase, cellobiohydrolase and β -glucosidase derived from, for example, a fungal or bacterial source.
- [0059] In some embodiments of the present invention, the tethered cellulase enzymes are tethered by a flexible linker sequence linked to an anchoring domain. In some embodiments, the anchoring domain is of CWP2 (for carboxy terminal anchoring) or FLO1 (for amino terminal anchoring) from *S. cerevisiae*. In particular embodiments, the

anchoring domain is encoded by the CWP portion of SEQ ID NO: 48. In other embodiments, the novel flexible linker sequence is encoded by the linker sequences of any one of SEQ ID NOs: 20-21 or 48.

[0060] In some embodiments, heterologous secretion signals may be added to the expression vectors of the present invention to facilitate the extra-cellular expression of cellulase proteins. In some embodiments, the heterologous secretion signal is the secretion signal from *S. cerevisiae* Xyn2.

[0061] In some embodiments, exogenous cellulase enzymes are added to the media. These may include the cellulase enzymes also expressed by the transformed host cells of the present invention such as cellobiohydrolases, endoglucanases, and β -glucosidases. However exogenously added enzymes may also include xylanases, amylases, and ligninases (such as laccases). One skilled in the art would recognize the need for various mixtures of exogenous enzymes depending on the host cell embodiments, and the particular substrates of the present invention.

[0062] In alternative embodiments, host cells of the present invention may themselves express xylanases, amylases, and ligninases (such as laccases).

[0063] In certain aspects of the invention, the cellulase system is a host cell comprising: (a) one heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase which, when expressed, is tethered to the cell surface; (b) two heterologous polynucleotides comprising nucleic acids which encode a cellobiohydrolase I and a cellobiohydrolase II which, when expressed, are tethered to the cell surface; (c) one heterologous polynucleotide comprising a nucleic acid sequence which encodes a β -glucosidase which, when expressed, is tethered to the cell surface; and (d) one additional heterologous polynucleotide comprising a nucleic acid which encodes a cellobiohydrolase which, when expressed, is secreted by the cell. In further aspects of the invention, the tethered endoglucanase is encoded by *T. reesei* *egl*, the tethered cellobiohydrolase I and II are encoded by *T. reesei* *cbh1* and *cbh2*, the tethered β -glucosidase is encoded by *S. fibuligera* *bgl1*, and the secreted cellobiohydrolase is encoded by *T. emersonii* *cbh1* or a fusion protein comprising *T. emersonii* *cbh1*, or *T. reesei* Cbh1 or Cbh2, or domain, fragment, variant, or derivative thereof.

[0064] The *T. emersonii* *cbh1* nucleic acid sequence is available in GenBank (Accession Number AY081766), and has the following sequence:

CTCAGACTCAAACACTCCATCAGCAGCTTCGAAAGCGGTCTTTTTGCTATCATCATGCTTCGA
 CGGGCTCTTCTTCTATCCTCTTCCGCCATCCTTGCTGTCAAGGCACAGCAGGCCGGCACGGCG
 ACGGCAGAGAACCACCCGCCCTGACATGGCAGGAATGCACCGCCCCTGGGAGCTGCACCA
 CCCAGAACGGGGCGGTCTTCTTGATGCGAACTGGCGTTGGGTGCACGATGTGAACGGATAC
 ACCAACTGCTACACGGGCAATACCTGGGACCCACGTAAGTGCCTGACGACGAAACCTGCGC
 CCAGAACTGTGCGCTGGACGGCGCGGATTACGAGGGCACCTACGGCGTGACTTCGTCGGGCA
 GCTCCTTGAAACTCAATTCGTACCCGGGTGCAACGTCGGATCCCGTCTCTACCTGCTGCAGG
 ACGACTCGACCTATCAGATCTTCAAGCTTCTGAACCGCGAGTTCAGCTTTGACGTCGATGTCT
 CCAATCTTCCGTGCGGATTGAACGGCGCTCTGTACTTTGTGCGCCATGGACGCCGACGGCGGC
 GTGTCCAAGTACCCGAACAACAAGGCTGGTGCCAAGTACGGAACCGGGTATTGCGACTCCCA
 ATGCCACCGGGACCTCAAGTTCATCGACGGCGAGGCCAACGTCGAGGGCTGGCAGCCGTCTT
 CGAACAACGCCAACACCGGAATTGGCGACCACGGCTCCTGCTGTGCGGAGATGGATGTCTGG
 GAAGCAAACAGCATCTCCAATGCGGTCACTCCGCACCCGTGCGACACGCCAGGCCAGACGA
 TGTGCTCTGGAGATGACTGCGGTGGCACATACTCTAACGATCGCTACGCGGGAACCTGCGAT
 CCTGACGGCTGTGACTTCAACCCTTACCGCATGGGCAACACTTCTTTCTACGGGCCTGGCAAG
 ATCATCGATACCACCAAGCCCTTCACTGTCGTGACGCAGTTCCTCACTGATGATGGTACGGAT
 ACTGGAACCTCTCAGCGAGATCAAGCGCTTCTACATCCAGAACAGCAACGTCATTCCGCAGCC
 CAACTCGGACATCAGTGGCGTGACCGGCAACTCGATCACGACGGAGTTCGCACTGCTCAGA
 AGCAGGCCTTTGGCGACACGGACGACTTCTCTCAGCACGGTGGCCTGGCCAAGATGGGAGCG
 GCCATGCAGCAGGGTATGGTCCTGGTGATGAGTTTGTGGGACGACTACGCCGCGCAGATGCT
 GTGGTTGGATTCCGACTACCCGACGGATGCGGACCCACGACCCCTGGTATTGCCCGTGGAA
 CGTGTCCGACGGACTCGGGCGTCCCATCGGATGTCGAGTCGCAGAGCCCCAACTCCTACGTG
 ACCTACTCGAACATTAAGTTTGGTCCGATCAACTCGACCTTACCGCTTCGTGAGTCTTGGTT
 ACATTTGAAGTAGACGGAAGTAGCTCTGCGATGGAAGTGGCATATGGAGAAGACCACACAA
 AACTGCATCGAAGAAAAGAGGGGGGAAAAGAGAAAAGCAAAGTTATTTAGTTTGAAAATGA
 AACTACGCTCGTTTTTATTCTTGAAAATCGCCACTCTTGCCCTTTTTTTCTTTTTTCTTTTTATT
 TTTTTTCTTTTGAAATCTTCAATTTAAATGTACATATTGTTAAATCAAATCAAGTAAATATAC
 TTGAAAAAAAAAAAAAAAAAAAAA (SEQ ID NO: 1)

[0065] The *H. grisea cbh1* nucleic acid sequence is available in GenBank (Accession Number X17258), and has the following sequence:

GCCGTGACCTTGCGCGCTTTGGGTGGCGGTGGCGAGTCGTGGACGGTGCTTGCTGGTTCGCCG
 GCCTTCCCGGCGATCCGCGTGATGAGAGGGCCACCAACGGCGGGATGATGCTCCATGGGGA
 ACTTCCCCTATGGAGAAGAGAGAGAACTTGCGGAGCCGTGATCTGGGGAAAGATGCTCCGT
 GTCTCGTCTATAAATCGAGTCTCCCCGAGCCCTCAACACCACAGCTCTGATCTCACCATC
 CCCATCGACAATCACGCAAACACAGCAGTTGTGCGGGCCATTCTTCAGACACATCAGTCACC
 CTCCTTCAAAATGCGTACCGCCAAGTTCGCCACCCTCGCCGCCCTTGTGGCCTCGGCCGCCGC
 CCAGCAGGCGTGCACTCTCACCACCGAGAGGCACCCTTCCTCTCTTGGAACAAGTGCACCG
 CCGGCGGCCAGTGCCAGACCGTCCAGGCTTCCATCACTCTCGACTCCAAGTGGCGCTGGACT
 CACCAGGTGCTTGCTCCACCAACTGCTACACGGGCAACAAGTGGGATACTAGCATCTGCAC
 TGATGCCAAGTTCGTGCGCTCAGAACTGCTGCGTCGATGGTGGCGACTACACCAGCACCTATG
 GCATCACCACCAACGGTGATTCCCTGAGCCTCAAGTTCGTACCAAGGGCCAGCACTCGACC
 AACGTCGGCTCGCGTACCTACCTGATGGACGGCGAGGACAAGTATCAGAGTACGTTCTATCT
 TCAGCCTTCTCGCGCCTTGAATCCTGGCTAACGTTTACACTTCACAGCCTTCGAGCTCCTCGG
 CAACGAGTTCACCTTCGATGTCGATGTCTCAACATCGGCTGCGGTCTCAACGGCGCCCTGTA
 CTTCTGCTCCATGGACGCCGATGGTGGTCTCAGCCGCTATCCTGGCAACAAGGCTGGTGCCA
 AGTACGGTACCGGCTACTGCGATGCTCAGTGCCCCCGTGACATCAAGTTCATCAACGGCGAG
 GCCAACATTGAGGGCTGGACCGGCTCCACCAACGACCCCAACGCCGGCGCGGGCCGCTATG
 GTACCTGCTGCTCTGAGATGGATATCTGGGAAGCCAACAACATGGCTACTGCCTTCACTCCTC
 ACCCTTGCAACCATCATTGGCCAGAGCCGCTGCGAGGGCGACTCGTGCGGTGGCACCTACAGC
 AACGAGCGCTACGCCGGCGTCTGCGACCCCGATGGCTGCGACTTCAACTCGTACCGCCAGGG
 CAACAAGACCTTCTACGGCAAGGGCATGACCGTCGACACCACCAAGAAGATCACTGTCGTCA

CCCAGTTCCTCAAGGATGCCAACGGCGATCTCGGCGAGATCAAGCGCTTCTACGTCCAGGAT
 GGCAAGATCATCCCCAACTCCGAGTCCACCATCCCCGGCGTCGAGGGCAATTCCATCACCCA
 GGACTGGTGCACCGCCAGAAGGTTGCCTTTGGCGACATTGACGACTTCAACCGCAAGGGCG
 GCATGAAGCAGATGGGCAAGGCCCTCGCCGGCCCCATGGTCCTGGTCATGTCCATCTGGGAT
 GACCACGCCTCCAACATGCTCTGGCTCGACTCGACCTTCCCTGTCGATGCCGCTGGCAAGCCC
 GGGCGCCGAGCGCGGTGCCTGCCCGACCACCTCGGGTGTCCCTGCTGAGGTTGAGGCCGAGGC
 CCCCACAGCAACGTCGTCTTCTCCAACATCCGCTTCGGCCCCATCGGCTCGACCGTTGCTGG
 TCTCCCCGGCGCGGGCAACGGCGGCAACAACGGCGGCAACCCCCCGCCCCCACCACCACC
 ACCTCCTCGGCTCCGGCCACCACCACCACCAGCGCTGGCCCCAAGGCTGGCCGCTGGCA
 GCAGTGCGGCGGCATCGGCTTCACTGGCCCCGACCCAGTGCGAGGAGCCCTACATTTGCACCA
 AGCTCAACGACTGGTACTCTCAGTGCCTGTAAATTCTGAGTCGCTGACTCGACGATCACGGC
 CGGTTTTTGCATGAAAGGAAACAAACGACCGCGATAAAAATGGAGGGTAATGAGATGTC
 (SEQ ID NO: 2)

[0066] The *T. aurantiacus cbh1* nucleic acid sequence is available in GenBank
 (Accession Number AF478686), and has the following sequence:

GAATTCTAGACCTTTATCCTTTCATCCGACCAGACTTCCCTTTTTGACCTTGGCGCCCTGTTGA
 CTACCTACCTACCTAGGTAGTAACGTCGTCGACCCTCTTGAATGATCCTTGTCACTGCAAA
 CATCCGAAACATACGGCAAAAGATGATTGGGCATGGATGCAGGAGACATCGAATGAGGGC
 TTAGAAGGAAATGAAAACCTGGGACCAGGACGCTAGGTACGATGAAATCCGCCAATGGTGA
 AACTTTAAGTCGTGCCTACAGCACAGGCTCTGTGAAGATTGCGCTGTTTCAGACTTAATCTTCT
 CATCACAGTCCAAGTCTTTATGAAAAGGAAAAAGAGAGGGGAAGAGCGCTATTTTCGAGCTGTT
 GGCCTCATAGGGAGACAGTCGAGCATACCAGCGGTATCGACGTTAGACTCAACCAAGAATA
 ATGACGAGAATAAACACAGAAGTCAACCTTGAAGTGGATAGCAGGGTTCCAGCAGCAGATA
 GTTACTTGCATAAAGACAACCTCCCCGAGGGCTCTCTGCATACACCAGGATGTTCCGGAATTA
 TTCCTGCTCGTTTCCGACGTGGCGTCAGTGATCCGTCTCCACAGAACTCTACCTGGGAATAA
 CCCAGGGGAGGAATCTGCAAGTAAGAACTTAATACCAATCCCCGGGGCTGCCGAGGTGAAT
 CGAATCTCCCGCGGGAAATTAACCCATACGATGTTTTTGCACCACATGCATGCTTAGCACG
 ATTTCTCCGCAAGGGAGTCACAGAGAAAGACATATTTTCGCATACTACTGTGACTCTGCAGAG
 TTACATATCACTCAGGATACATTGCAGATCATTGTCCGGGCATCAAAAATGGACCTGCAGGA
 TCAACGGCCCGACAAAACACAAGTGGCTAAAGCTGGGGGATGCCCGAAACCCTCTGGTGCA
 ATATCATTTGATGGATGTTCCCCCGCATTTCTAAGACATCGACGGATCGGCCCGCATACTAA
 TCCTTTTATCAACCAAAAGTTCCACTCGACTAGAGAAAAAAAAGGCCAAGGCCACTAGTTGC
 AGTCGGATACTGGTCTTTTCGCCGTCCAACACCTTCATCCATGATCCCCTTAGCCACCAATGC
 CCCACATAATACATGTTGACATAGGTACGTAGCTCTGTTATCCAATCGGATCCGAACCTCTTT
 AACGGACCCCTCCTACACACCTTATCCTAACTTCAGAAGACTGTTGCCCATTTGGGGATTGAG
 GAGGTCCGGGTCGCAGGATGCGTTCTAGGCTAAATTCTCGGCCGGTAGCCATCTCGAATCTC
 TCGTGAAGCCTTCATCTGAACGGTTGGCGGCCCGTCAAGCCGATGACCATGGGTTCTTGATA
 GAGCTTGTGCCTGACCGGCCTTGGCGGCATAGACGAGCTGAACACATCAGGTATGAACAGAT
 CAGATATAAAGTCGGATTGAGTCCTAGTACGAAGCAATCCGCCACCACCAAAATCAAGCAAC
 GAGCGACACGAATAACAATATCAATCGAATCGCAATGTATCAGCGCGCTCTTCTCTTCTCTTT
 CTTCTCGCCGCCGCCCGCGCGCACGAGGCCGGTACCCTAACCAGCAGAGAATCACCCCTCCC
 TGACCTGGCAGCAATGCTCCAGCGGCGGTAGTTGTACCACGCAGAATGGAAAAGTCGTTATC
 GATGCGAACTGGCGTTGGGTCCATACCACCTCTGGATACACCAACTGCTACACGGGCAATAC
 GTGGGACACCAGTATCTGTCCCGACGACGTGACCTGCGCTCAGAATTGTGCCTTGGATGGAG
 CGGATTACAGTGGCACCTATGGTGTACGACCAGTGGCAACGCCCTGAGACTGAACTTTGTC
 ACCCAAAGCTCAGGGAAGAACATTGGCTCGCGCCTGTACCTGCTGCAGGACGACACCACTTA
 TCAGATCTTCAAGCTGCTGGGTGAGGAGTTTACCTTCGATGTCGACGTCTCCAATCTCCCTTG
 CGGGCTGAACGGCGCCCTCTACTTTGTGGCCATGGACGCCGACGGCAATTTGTCCAAATACC
 CTGGCAACAAGGCAGGCGCTAAGTATGGCACTGGTTACTGCGACTCTCAGTGCCCTCGGGAT
 CTCAAGTTCATCAACGGTCAGGTACGTGAGAAGTGATAACTAGCCAGCAGAGCCCATGAATC
 ATTAATAACGCTGTCAAATACAGGCCAACGTTGAAGGCTGGCAGCCGTCTGCCAACGACCC

AAATGCCGGCGTTGGTAACACGGTTCCTCGTGCGCTGAGATGGATGTCTGGGAAGCCAACA
GCATCTCTACTGCGGTGACGCCTCACCCATGCGACACCCCCGGCCAGACCATGTGCCAGGGA
GACGACTGTGGTGGAACTACTCCTCCACTCGATATGCTGGTACCTGCGACCCTGATGGCTG
CGACTTCAATCCTTACCAGCCAGGCAACCACTCGTTCTACGGCCCCGGGAAGATCGTCGACA
CTAGCTCCAAATTCACCGTCGTCACCCAGTTCATCACCGACGACGGGACACCCTCCGGCACC
CTGACGGAGATCAAACGCTTCTACGTCCAGAACGGCAAGGTGATCCCCAGTCGGAGTCGAC
GATCAGCGGCGTCACCGGCAACTCAATCACCAACGAGTATTGCACGGCCCAGAAGGCAGCCT
TCGGCGACAACACCGGCTTCTTCACGCACGGCGGGCTTCAGAAGATCAGTCAGGCTCTGGCT
CAGGGCATGGTCTCGTCATGAGCCTGTGGGACGATCACGCCGCCAACATGCTCTGGCTGGA
CAGCACCTACCCGACTGATGCGGACCCGGACACCCCTGGCGTCGCGCGCGGTACCTGCCCA
CGACCTCCGGCGTCCCGGCCGACGTTGAGTCGCAGAACCCCAATTCATATGTTATCTACTCCA
ACATCAAGGTTCGGACCCATCAACTCGACCTTCACCGCCAATAAGTAAGTAACGGGCACTCT
ACCACCGAGAGCTTCGTGAAGATACAGGGGTAGTTGGGAGATTGTCGTGTACAGGGGACAT
GCGATGCTCAAAAATCTACATCAGTTTGCCAATTGAACCATGAAGAAAAGGGGGAGATCAA
AGAAGTCTGTCAGAAGAGAGGGGCTGTGGCAGCTTAAGCCTTGTTGTAGATCGTTCAGAGAA
AAAAAAGTTTGCGTACTTATTATATTAGGTCGATCATTATCCGATTGACTCCGTGACAAGA
ATTAAAAAGAGTACTGCTTGCTTGCCATTTTAAATTGTTATATACGCCGTAGCGCTTGCGGAC
CACCCCTCACAGTATATCGGTTGCGCTCTTCTGTCTCTTCATCTCACATCACAGGTCCAGGTC
CAGCCCGGCCCGGTCCGGGTGCCATGCATGCACAGGGGGACTAATATATTAATCGTGACCCT
GTVCCCTAAGCTAGGGTCCCTGCATTTTGAACCTGTGGACGTCTG (SEQ ID NO: 3)

[0067] The *T. reesei cbh1* nucleic acid sequence is available in GenBank (Accession Number E00389), and has the following sequence:

AAGGTTAGCCAAGAACAATAGCCGATAAAGATAGCCTCATTAACCGGAATGAGCTAGTAGG
CAAAGTCAGCGAATGTGTATATATAAAGGTTTCGAGGTCCGTGCCTCCCTCATGCTCTCCCAT
CTACTCATCAACTCAGATCCTCCAGGAGACTTGTACACCATCTTTTGAGGCACAGAAACCA
ATAGTCAACCGCGGACTGGCATCATGTATCGGAAGTTGGCCGTCATCACGGCCTTCTTGCC
ACAGCTCGTGCTCAGTCGGCCTGCACTCTCCAATCGGAGACTCACCCGCCTCTGACATGGCA
GAAATGCTCGTCTGGTGGCACTTGCACTCAACAGACAGGCTCCGTGGTCATCGACGCCAACT
GGCGCTGGACTCACGCTACGAACAGCAGCACGAAGTCTACGATGGCAACACTTGAGCTC
GACCCTATGTCCTGACAACGAGACCTGCGCGAAGAACTGCTGTCTGGACGGTGCCGCCTACG
CGTCCACGTACGGAGTTACCACGAGCGGTAACAGCCTCTCCATTGGCTTTGTCACCCAGTCTG
CGCAGAAGAACGTTGGCGCTCGCCTTTACCTTATGGCGAGCGACACGACCTACCAGGAATTC
ACCCTGCTTGGAACGAGTTCTCTTTTCGATGTTGATGTTTCGCAGCTGCCGTAAGTGACTTAC
CATGAACCCCTGACGTATCTTCTTGTTGGGCTCCAGCTGACTGGCCAATTTAAGGTGCGGCTT
GAACGGAGCTCTCTACTTCGTGTCCATGGACGCGGATGGTGGCGTGAGCAAGTATCCACCA
ACAACGCTGGCGCCAAGTACGGCACGGGGTACTGTGACAGCCAGTGTCCCCGCGATCTGAA
GTTTCATCAATGGCCAGGCCAACGTTGAGGGCTGGGAGCCGTCATCCAACAACGCAAACACG
GGCATTGGAGGACACGGAAGCTGCTGCTCTGAGATGGATATCTGGGAGGCCAACTCCATCTC
CGAGGCTCTTACCCCCCACCCTTGACGACTGTGCGCCAGGAGATCTGCGAGGGTGATGGGT
GCGGCGGAACTTACTCCGATAACAGATATGGCGGCACTTGCGATCCCGATGGCTGCGACTGG
AACCCATACCGCCTGGGCAACACAGCTTCTACGGCCCTGGCTCAAGCTTTACCTCGATAC
CACCAAGAAATTGACCGTTGTACCCAGTTTCGAGACGTCGGGTGCCATCAACCGATACTATG
TCCAGAATGGCGTCACTTTCCAGCAGCCCAACGCCGAGCTTGGTAGTTACTCTGGCAACGAG
CTCAACGATGATTACTGCACAGCTGAGGAGACAGAATTCGGCGGATCTCTTTCTCAGACAAG
GGCGGCCTGACTCAGTTCAAGAAGGCTACCTCTGGCGGCATGGTTCTGGTCATGAGTCTGTG
GGATGATGTGAGTTTGATGGACAAACATGCGCGTTGACAAAGAGTCAAGCAGCTGACTGAG
ATGTTACAGTACTACGCCAACATGCTGTGGCTGGACTCCACCTACCCGACAAACGAGACCTC
CTCCACACCCGGTGCCGTGCGCGGAAGCTGCTCCACCAGCTCCGGTGTCCCTGCTCAGGTGCG
AATCTCAGTCTCCCAACGCCAAGGTACCTTCTCCAACATCAAGTTCGGACCCATTGGCAGC
ACCGGCAACCCTAGCGGCGGCAACCCTCCCGGCGGAAACCGTGGCACCACCACCCGCC
GCCAGCCACTACCACTGGAAGCTCTCCCGGACCTACCCAGTCTCACTACGGCCAGTGCGGC

GGTATTGGCTACAGCGGCCCCACGGTCTGCGCCAGCGGCACAACCTTGCCAGGTCCTGAACCC
T TACTACTCTCAGTGCCTGTAAAGCTCCGTGCGAAAGCCTGACGCACCGGTAGATTCTTGGTG
AGCCCGTATCATGACGGCGGCGGGAGCTACATGGCCCCGGGTGATTTATTTTTTTGTATCTA
CTTCTGACCCTTTTCAAATATACGGTCAACTCATCTTTCACCTGGAGATGCGGCCTGCTTGGTA
TTGCGATGTTGTCAGCTTGGCAAATTGTGGCTTTCGAAAACACAAAACGATTCCTTAGTAGCC
ATGCATTTTAAGATAACGGAATAGAAGAAAGAGGAAATTAAAAAAAAAAAAAAAAACAAAC
ATCCCGTTCATAACCCGTAGAATCGCCGCTCTTCGTGTATCCCAGTACCA (SEQ ID NO: 4)

[0068] The *T. emersonii cbh2* nucleic acid sequence is available in GenBank (Accession Number AF439936), and has the following sequence:

GACGGACCTGCACTTAGTCGGTAGGTTATGTATGTAGCTGGAGATTGGGATAGGGAAGTTAG
CTAATAGTCTACTTCGTGTGAGGGTTGATTTTGATGGTCGACAGTATTCGTTTCTTATACGCA
GCGTCATGGATCTGTGTTTCTGTACATGTCGGGTGGATGGTTCCTGGACAGCAGCACACAA
ATGGTGTTCGTAGATAGGCGATACTCGGCAGGGGATTGTGCAGGGGATTGTATCGTAGATG
GTTCTAGTAAAATAGATCCCGAGTATGGTTAGCTCTCATACCTCGAGTNGATGAAGCACAAT
ATGCTACGATATGCCAAGTAAAACCTCTATTGTATTCTGCAGCTAGCAATTGAAGAATCCGAC
ATTCCCATTTGTCATCTAATCGGGCAGACATGTGCAAAGAGGGACGATTTCGTGATCGAAGTGC
TCCAATCCATGGCGTAGGACCAGACAGCTCCATCCGATCTAGAGCTATATGGAGCTCCTCGC
AACTCCGACACTCCGCGAGACAGCTCTCACAAGCACTATAAATATGGCCAAGAACCCTGCAG
AACAGCTTCACTCTACAGCCCGTTGAGCAGAAACAAACAAATATCACTCCAGAGAGAAAGC
AACATGCGGAATCTTCTTGCTCTTGACACCGGCCGCGCTGCTTGTCGGCGCAGCGGAAGCGCA
ACAATCCCTCTGGGGACAATGTGAGCAGCTCCTAAACGTCTGTCTGAGGGATTATGTCTGAC
TGCTCAGGCGGCGGGAGTTTCGTGGACTGGCGCGACGAGCTGTGCTGCTGGAGCGACGTGCA
GCACAATCAATCCTTGTACGTCTGCTGAACGATAATCCTACATTGTTGACGTGCTAACTGCGT
AGACTACGCACAATGCGTTCCTGCAACGGCCACTCCGACCACGCTGACGACAACGACAAAA
CCAACGTCCACCGGCGGCGCTGCTCCAACGACTCCTCCTCCGACAACGACTGGAACAACGAC
ATCGCCCGTCGTACACAGGCCCCGCTGCTGCCCTCCGGCAACCCGTTTCAAGGCTACCAGCTCT
ACGCCAATCCGTACTATGCGTCGGAGGTGATTAGTTTGGCAATTCCTCGCTGAGCAGCGAG
CTGGTTCCCAAGGCGAGCGAGGTGGCCAAGGTGCCGTCTTTCGTCTGGCTGTAAGTAAATTC
CCCCAGGCTGTCAATTTCCCCTTACTGATCTTGTCCAGCGACCAAGCCGCCAAGGTGCCAGCA
TGGGCGACTATCTGAAAGACATCCAGTCGAGAACGCAGCCGGCGCAGACCCCCCGATTGC
AGGCATCTTTGTCGTCTACGACCTGCCTGACCGCGACTGCGCGGCTGCAGCCAGCAATGGCG
AGTTCTCCATCGCCAACAACGGCGTCGCCCTGTACAAGCAGTACATCGACTCGATCCGCGAG
CAGCTGACGACCTATTTCAGATGTGCACACCATCCTGGTCATCGGTAGTTCAGTCCTCTTCTG
TGATGTTGATGAAAAAATACTGACTGACTCCTGCAGAACCCGACAGCCTTGCGAACGTGGT
CACCAACCTGAACGTGCCGAAATGCGCAAATGCCCAGGACGCCTATCTCGAATGCATCAACT
ACGCCATCACCCAGCTCGATCTGCCAAACGTGGCCATGTATCTTGATGCTGGTGAGTCCTCAC
ATACAAGTGAATAAAAAATAAACTGATGCAGTGCAGGACACGCCGGATGGCTAGGCTGGCA
AGCCAACCTCGCCCCCGCCGCCAGCTGTTTGCTCGGTGTACAAAACGCCTCCTCTCCGGC
ATCCGTCCGCGGTCTCGCCACCAACGTGCGCAACTACAACGCCTGGTCGATCAGCCGGTGCC
CGTCGTACACGCAGGGCGACGCCAATTGCGACGAGGAGGATTACGTGAATGCCTTGGGGCC
GTTGTTCCAGGAACAGGGATTCCCGGCATATTTATCATTGATACATGTAAGCTTTACCCAG
AACCCTCCATAGAAGGTCAATCTAACGGTAATGTACAGCCCGCAATGGCGTCCGACCCACC
AAGCAAAGCCAATGGGGCGACTGGTGCAACGTATCGGCACGGGCTTCGGCGTCCGGCCCA
CGACCGACACCGGCAATCCTCTCGAGGACGCTTTCGTCTGGGTCAAGCCCGGTGGCGAGAGC
GATGGCACGTCCAACACGACCTCTCCGCGGTACGACTACCACTGCGGGCTGAGCGATGCGCT
GCAGCCGGCGCCGGAGGCGGGGACTTGGTTCCAGGTATGACGCGCCTTCGTATTAGCAATTA
CGATACATGTGCATGCTGACCATGCGACAGGCGTACTTTGAGCAGTTGCTCACGAATGCTAA
CCCGCTGTTCTGA (SEQ ID NO: 5)

[0069] The *T. reesei cbh2* nucleic acid sequence is available in GenBank (Accession Number M16190), and has the following sequence:

TCGAACTGACAAGTTGTTATATTGCCTGTGTACCAAGCGCGAATGTGGACAGGATTAATGCC
 AGAGTTCATTAGCCTCAAGTAGAGCCTATTTCTCGCCGAAAGTCATCTCTCTTATTGCATT
 TCTGCCCTTCCCACTAACTCAGGGTGCAGCGCAACACTACACGCAACATATACACTTTATTAG
 CCGTGCAACAAGGCTATTCTACGAAAAATGCTACACTCCACATGTTAAAGGCGCATTCAACC
 AGCTTCTTTATTGGGTAATATACAGCCAGGCGGGGATGAAGCTCATTAGCCGCCACTCAAGG
 CTATACAATGTTGCCAACTCTCCGGGCTTTATCCTGTGCTCCCGAATACCACATCGTGATGAT
 GCTTCAGCGCACGGAAGTCACAGACACCGCCTGTATAAAAGGGGGACTGTGACCCTGTATGA
 GGCAGCAACATGGTCTCACAGCAGCTCACCTGAAGAGGCTTGTAAGATCACCTCTGTGTATT
 GCACCATGATTGTCGGCATTCTCACCACGCTGGCTACGCTGGCCACACTCGCAGCTAGTGTG
 CCTCTAGAGGAGCGGCAAGCTTGCTCAAGCGTCTGGTAATTATGTGAACCCTCTCAAGAGAC
 CCAAATACTGAGATATGTCAAGGGGGCCAATGTGGTGGCCAGAATTGGTTCGGGTCCGACTTGC
 TGTGCTTCCGGAAGCACATGCGTCTACTCCAACGACTATTACTCCCAGTGTCTTCCCGGCGCT
 GCAAGCTCAAGCTCGTCCACGCGCGCCGCGTCGACGACTTCTCGAGTATCCCCACAACATC
 CCGGTCGAGCTCCGCGACGCCTCCACCTGGTTCTACTACTACCAGAGTACCTCCAGTCGGATC
 GGAACCGCTACGTATTCAGGCAACCCTTTTGTGGGGTCACTCCTTGGGCCAATGCATATTA
 CGCCTCTGAAGTTAGCAGCCTCGCTATTCCTAGCTTGACTGGAGCCATGGCCACTGCTGCAGC
 AGCTGTCGCAAAGGTTCCCTCTTTTATGTGGCTGTAGGTCCTCCCGGAACCAAGGCAATCTGT
 TACTGAAGGCTCATCATTCACTGCAGAGATACTCTTGACAAGACCCCTCTCATGGAGCAAAC
 CTTGGCCGACATCCGCACCGCCAACAAGAATGGCGGTAATATGCCGGACAGTTTGTGGTGT
 ATGACTTGCCGGATCGCGATTGCGCTGCCCTTGCCCTCGAATGGCGAATACTCTATTGCCGATG
 GTGGCGTCGCCAAATATAAGAACTATATCGACACCATTTCGTCAAATTGTCGTGGAATATTCC
 GATATCCGGACCCTCCTGGTTATTGGTGAGTTTAAACACCTGCCTCCCCCCCCCTTCCCTTC
 CTTTCCCGCCGGCATCTTGTCGTTGTGCTAACTATTGTTCCCTCTTCCAGAGCCTGACTCTCTT
 GCCAACCTGGTGACCAACCTCGGTACTCCAAAGTGTGCCAATGCTCAGTCAGCCTACCTTGA
 GTGCATCAACTACGCCGTCACACAGCTGAACCTTCCAAATGTTGCGATGTATTTGGACGCTG
 GCCATGCAGGATGGCTTGGCTGGCCGGCAAACCAAGACCCGGCCGCTCAGCTATTTGCAAAT
 GTTTACAAGAATGCATCGTCTCCGAGAGCTCTTCGCGGATTGGCAACCAATGTCGCCAACTA
 CAACGGGTGGAACATTACCAGCCCCCATCGTACACGCAAGGCAACGCTGTCTACAACGAG
 AAGCTGTACATCCACGCTATTGGACCTCTTCTTGCCAATCACGGCTGGTCCAACGCCTTCTTC
 ATCACTGATCAAGGTCGATCGGGAAAGCAGCCTACCGGACAGCAACAGTGGGGAGACTGGT
 GCAATGTGATCGGCACCGGATTTGGTATTCGCCCATCCGCAAACACTGGGGACTCGTTGCTG
 GATTCGTTTGTCTGGGTCAAGCCAGGCGGCGAGTGTGACGGCACCAGCGACAGCAGTGCGCC
 ACGATTTGACTCCCACTGTGCGCTCCAGATGCCTTGCAACCGGCGCCTCAAGCTGGTGCTTG
 GTTCCAAGCCTACTTTGTGCAGCTTCTCACAACGCAAACCCATCGTTCTGTAAAGGCTTTCG
 TGACCGGGCTTCAAACAATGATGTGCGATGGTGTGGTTCCCGGTTGGCGGAGTCTTTGTCTAC
 TTTGGTTGT (SEQ ID NO: 6)

[0070] The present invention also provides for the use of an isolated polynucleotide comprising a nucleic acid at least about 70%, 75%, or 80% identical, at least about 90% to about 95% identical, or at least about 96%, 97%, 98%, 99% or 100% identical to any of SEQ ID NOs:1-6, or fragments, variants, or derivatives thereof.

[0071] In certain aspects, the present invention relates to a polynucleotide comprising a nucleic acid encoding a functional or structural domain of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 or Cbh2. For example, the domains of *T. reesei* Cbh 1 include, without limitation: (1) a signal sequence, from amino acid 1 to 33 of SEQ ID

NO: 17; (2) a catalytic domain (CD) from about amino acid 41 to about amino acid 465 of SEQ ID NO: 17; and (3) a cellulose binding module (CBM) from about amino acid 503 to about amino acid 535 of SEQ ID NO: 17. The domains of *T. reesei* Cbh 2 include, without limitation: (1) a signal sequence, from amino acid 1 to 33 of SEQ ID NO: 18; (2) a catalytic domain (CD) from about amino acid 145 to about amino acid 458 of SEQ ID NO: 18; and (3) a cellulose binding module (CBM) from about amino acid 52 to about amino acid 83 of SEQ ID NO: 18.

[0072] The present invention also encompasses an isolated polynucleotide comprising a nucleic acid that is at least about 70%, 75%, or 80% identical, at least about 90% to about 95% identical, or at least about 96%, 97%, 98%, 99% or 100% identical to a nucleic acid encoding a *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 or Cbh2 domain, as described above.

[0073] The present invention also encompasses variants of the *cbh1* or *cbh 2* genes, as described above. Variants may contain alterations in the coding regions, non-coding regions, or both. Examples are polynucleotide variants containing alterations which produce silent substitutions, additions, or deletions, but do not alter the properties or activities of the encoded polypeptide. In certain embodiments, nucleotide variants are produced by silent substitutions due to the degeneracy of the genetic code. In further embodiments, *T. emersonii*, *H. grisea*, *T. aurantiacus*, and *T. reesei cbh1* or *cbh2* 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 *T. emersonii cbh1* mRNA to those preferred by a host such as the yeast *Saccharomyces cerevisiae*).

[0074] Codon-optimized polynucleotides of the present invention are discussed further below.

[0075] The present invention also encompasses an isolated polynucleotide encoding a fusion protein. In certain embodiments, the nucleic acid encoding a fusion protein comprises a first polynucleotide encoding for a *T. emersonii cbh1*, *H. grisea cbh1*, or *T. aurantiacus cbh1*, *T. emersonii cbh1* and a second polynucleotide encoding for the CBM domain of *T. reesei cbh1* or *T. reesei cbh2*. In particular embodiments of the nucleic acid encoding a fusion protein, the first polynucleotide is *T. emersonii cbh1* and the second polynucleotide encodes for a CBM from *T. reesei* Cbh1 or Cbh2. In further embodiments of the fusion protein, 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*. In particular embodiments of the nucleic acid encoding a fusion protein, the first polynucleotide is a codon-optimized *T. emersonii cbh1* and the second polynucleotide encodes for a codon-optimized CBM from *T. reesei* Cbh1 or Cbh2.

[0076] Also provided in the present invention are allelic variants, orthologs, and/or species homologs. Procedures known in the art can be used to obtain full-length genes, allelic variants, splice variants, full-length coding portions, orthologs, and/or species homologs of genes corresponding to any of SEQ ID NOs: 1-6, using information from the sequences disclosed herein or the clones deposited with the ATCC. For example, allelic variants and/or species homologs may be isolated and identified by making suitable probes or primers from the sequences provided herein and screening a suitable nucleic acid source for allelic variants and/or the desired homologue.

[0077] By a nucleic acid having a nucleotide sequence at least, for example, 95% "identical" to a reference nucleotide sequence of the present invention, it is intended that the nucleotide sequence of the nucleic acid is identical to the reference sequence except that the nucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence encoding the particular polypeptide. In other words, to obtain a nucleic acid having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. The query sequence may be an entire sequence shown of any of SEQ ID NOs:1-6, or any fragment or domain specified as described herein.

[0078] As a practical matter, whether any particular nucleic acid molecule or polypeptide is at least 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% identical to a nucleotide sequence or polypeptide of the present invention can be determined conventionally using known computer programs. A method for determining the best overall match between a query sequence (a sequence of the present invention) and a subject sequence, also referred

to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag *et al.* (*Comp. App. Biosci.* (1990) 6:237-245.) In a sequence alignment the query and subject sequences are both DNA sequences. An RNA sequence can be compared by converting U's to T's. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB alignment of DNA sequences to calculate percent identity are: Matrix=Unitary, k-tuple=4, Mismatch Penalty=1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score=1, Gap Penalty=5, Gap Size Penalty 0.05, Window Size=500 or the length of the subject nucleotide sequence, whichever is shorter.

[0079] 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.

[0080] 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.

[0081] Some embodiments of the invention encompass a nucleic acid molecule comprising at least 10, 20, 30, 35, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, or 800 consecutive nucleotides or more of any of SEQ ID NOs:1-6, or domains, fragments, variants, or derivatives thereof.

[0082] The polynucleotide of the present invention may be in the form of RNA or in the form of DNA, which DNA includes cDNA, genomic DNA, and synthetic DNA. The DNA may be double stranded or single-stranded, and if single stranded may be the coding strand or non-coding (anti-sense) strand. The coding sequence which encodes the mature polypeptide may be identical to the coding sequence encoding SEQ ID NO:11-14 or 17-18, or may be a different coding sequence which coding sequence, as a result of the redundancy or degeneracy of the genetic code, encodes the same mature polypeptide as the DNA of any one of SEQ ID NOs:1-6,

[0083] In certain embodiments, the present invention provides an isolated polynucleotide comprising a nucleic acid fragment which encodes at least 10, at least 20, at least 30, at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 95, or at least 100 or more contiguous amino acids of SEQ ID NOs: 11-14 and 17-18.

[0084] The polynucleotide encoding for the mature polypeptide of SEQ ID NOs: 11-14 and 17-18 or may include: only the coding sequence for the mature polypeptide; the coding sequence of any domain of the mature polypeptide; and the coding sequence for the mature polypeptide (or domain-encoding sequence) together with non-coding sequence, such as introns or non-coding sequence 5' and/or 3' of the coding sequence for the mature polypeptide.

[0085] Thus, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only sequences encoding for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequences.

[0086] In further aspects of the invention, nucleic acid molecules having sequences at least about 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleic acid sequences disclosed herein, encode a polypeptide having Cbh, Eg or Bgl functional activity. By "a polypeptide having Cbh, Eg or Bgl functional activity" is intended polypeptides exhibiting activity similar, but not necessarily identical, to a functional activity of the Cbh, Eg or Bgl polypeptides of the present invention, as measured, for example, in a particular biological assay. For example, a Cbh, Eg or Bgl functional activity can routinely be measured by determining the ability of a Cbh, Eg or Bgl polypeptide to hydrolyze cellulose, or by measuring the level of Cbh, Eg or Bgl activity.

[0087] 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 at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequence of any of SEQ ID NOs:1-6, or fragments thereof, will encode polypeptides "having Cbh, Eg or 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 Cbh, Eg or Bgl functional activity.

[0088] The polynucleotides of the present invention also comprise nucleic acids encoding a *T. emersonii*, *H. grisea*, *T. aurantiacus*, and *T. reesei* Cbh1 and/or Cbh2, or domain, fragment, variant, or derivative thereof, fused in frame to a marker sequence which allows for detection of the polypeptide of the present invention. The marker sequence may be a yeast selectable marker selected from the group consisting of *URA3*, *HIS3*, *LEU2*, *TRP1*, *LYS2* or *ADE2*. Casey, G.P. *et al.*, "A convenient dominant selection marker for gene transfer in industrial strains of *Saccharomyces* yeast: *SMR1* encoded resistance to the herbicide sulfometuron methyl," *J. Inst. Brew.* 94:93-97 (1988).

Codon Optimized Polynucleotides Encoding Secreted and Tethered Enzymes

[0089] As used herein the term "codon optimized coding region" means a nucleic acid coding region that has been adapted for expression in the cells of a given vertebrate by replacing at least one, or more than one, or a significant number, of codons with one or more codons that are more frequently used in the genes of that vertebrate.

[0090] 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.

[0091] The CAI of codon optimized sequences of the present invention corresponds to between about 0.8 and 1.0, between about 0.8 and 0.9, or about 1.0. A codon optimized sequence may be further modified for expression in a particular organism, depending on that organism's biological constraints. For example, large runs of "As" or "Ts" (e.g., runs greater than 4, 4, 5, 6, 7, 8, 9, or 10 consecutive bases) can be removed from the sequences if these are known to effect transcription negatively. Furthermore, specific restriction enzyme sites may be removed for molecular cloning purposes. Examples of such restriction enzyme sites include PacI, AscI, BamHI, BglII, EcoRI and XhoI. Additionally, the DNA sequence can be checked for direct repeats, inverted repeats and mirror repeats with lengths of ten bases or longer, which can be modified manually by replacing codons with "second best" codons, *i.e.*, codons that occur at the second highest frequency within the particular organism for which the sequence is being optimized.

[0092] 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 "

[0093] 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.

[0094] Given the large number of gene sequences available for a wide variety of animal, plant and microbial species, it is possible to calculate the relative frequencies of codon usage. Codon usage tables are readily available, for example, at <http://phenotype.biosci.umbc.edu/codon/sgd/index.php> (visited May 7, 2008) or at <http://www.kazusa.or.jp/codon/> (visited March 20, 2008), and these tables can be adapted

in a number of ways. See Nakamura, Y., *et al.* "Codon usage tabulated from the international DNA sequence databases: status for the year 2000" *Nucl. Acids Res.* 28:292 (2000). Codon usage tables for yeast, calculated from GenBank Release 128.0 [15 February 2002], are reproduced below as Table 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
Total			
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
Total			
Ile	AUU	196893	30.1
Ile	AUC	112176	17.2
Ile	AUA	116254	17.8
Total			
Met	AUG	136805	20.9
Total			
Val	GUU	144243	22.1
Val	GUC	76947	11.8
Val	GUA	76927	11.8
Val	GUG	70337	10.8
Total			
Ser	UCU	153557	23.5
Ser	UCC	92923	14.2
Ser	UCA	122028	18.7

Amino Acid	Codon	Number	Frequency per hundred
Ser	UCG	55951	8.6
Ser	AGU	92466	14.2
Ser	AGC	63726	9.8
Total			
Pro	CCU	88263	13.5
Pro	CCC	44309	6.8
Pro	CCA	119641	18.3
Pro	CCG	34597	5.3
Total			
Thr	ACU	132522	20.3
Thr	ACC	83207	12.7
Thr	ACA	116084	17.8
Thr	ACG	52045	8.0
Total			
Ala	GCU	138358	21.2
Ala	GCC	82357	12.6
Ala	GCA	105910	16.2
Ala	GCG	40358	6.2
Total			
Tyr	UAU	122728	18.8
Tyr	UAC	96596	14.8
Total			
His	CAU	89007	13.6
His	CAC	50785	7.8
Total			
Gln	CAA	178251	27.3
Gln	CAG	79121	12.1
Total			
Asn	AAU	233124	35.7
Asn	AAC	162199	24.8
Total			
Lys	AAA	273618	41.9
Lys	AAG	201361	30.8
Total			

Amino Acid	Codon	Number	Frequency per hundred
Asp	GAU	245641	37.6
Asp	GAC	132048	20.2
Total			
Glu	GAA	297944	45.6
Glu	GAG	125717	19.2
Total			
Cys	UGU	52903	8.1
Cys	UGC	31095	4.8
Total			
Trp	UGG	67789	10.4
Total			
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
Total			
Gly	GGU	156109	23.9
Gly	GGC	63903	9.8
Gly	GGA	71216	10.9
Gly	GGG	39359	6.0
Total			
Stop	UAA	6913	1.1
Stop	UAG	3312	0.5
Stop	UGA	4447	0.7

[0095] 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.

[0096] 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.

[0097] 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.

[0098] 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.

[0099] Codon-optimized sequences of the present invention include those as set forth in Tables 3 and 4 below:

Table 3. Synthetic DNA constructs for expression vector backbone and tethered cellulase constructs, provided on plasmids.

Name	Abbreviation	Sequence (lower-case denotes restriction sites)	Native sequence of DNA fragment (Accession Number listed at top)	Amino Acid Sequence (from optimized sequence, including linkers and/or anchors from synthetic constructs)
YDRWdelt a23	δ	agtcggtaccTGTTGGAATAAAAAATCCACTATCGTCT ATCAACTAATAGTTATATTATCAATATATTATC ATATACGGTGTTAAGATGATGACATAAGTTATG AGAAGCTGTCATCGATGTTAGAGGAAGCTGAA ACGCAAGGATTGATAATGTAATAGGATCAATG AATATAAACATATAAAACGGAATGAGGAATAA TCGTAATATTAGTATGTAGAAATATAGATTCCA TTTTGAGGATTCCTATATCCTCGAGGAGAACTT CTAGTATATTCTGTATACCTAATATTATAGCCTT	NC_001136 (<i>S. cerevisiae</i> Chromosome IV—sequence contained in) Same sequence	None

			TATCAACAATGGAATCCCAACAATTATCTAATT ACCCACATATATCTCAgggccgcgc (SEQ ID NO: 19)		
Endo- glucanase I from <i>Trichoderm a reesei</i> (TrEGI) and short linker	C2	gagtcgccgggCAACAACCAGGAACATCAACACCAG AAGTCCATCCAAAGTTAACAACCTATAAATGTA CTAAGAGTGGAGGGTGTGTAGCGCAGGACACA AGTGTGGTCTTAGACTGGAATTATCGTTGGATG CATGATGCCAATTATAATTCCTGTACTGTTAAC GGCGGTGTTAACTACGTTATGCCCGGATGAA GCGACTTGTGGTAAGAATTGTTTTATTGAAGGG GTTGACTACGCCGCTAGTGGTGTACGACGAGT GGGTACTCCTTGACGATGAATCAATACATGCCT TCTTCTAGTGGTGGGTATTCTCTGTGTCTCCAA GGCTGTATTTATTGGATTCCGATGGGGAATATG TTATGTTAAAATTAATGGGCAAGAACTGAGTT TTGATGTGGATCTATCTGCATTACCTTGTGGAG AAAATGGTAGTCTTTATTTATCACAAATGGACG AAAACGGCGGAGCCAATCAGTACAATACAGCT GGTGCTAATTATGGTTCAGGCTATTGTGATGCT CAATGTCCAGTGCAGACTTGGAGGAATGGCAC CTTAAACACATCACATCAAGGATTTTGCTGTAA CGAAATGGACATATTAGAAGGTAATTCAAGAG CTAATGCACTAACTCCGCACTCTTGTACTGCGA CCGCATGTGATTCTGCCGGTTGTGGTTTCAACC CTTATGGTICTGGTTATAAGAGTTACTACGGTC CGGGAGACACCGTGGATACGTCAAAGACCTTC ACTATAATCACTCAGTTTAAACACAGATAACGGA TCTCCGAGTGGTAATTTGGTGAGTATTACTAGG AAATATCAGCAGAACGGTGTGATATTCGGTCC GCGCAGCCAGGCGGTGACACTATATCTAGCTGT CCTTCCGCCAGTGCCTATGGCGGACTTGTCTACA ATGGGTAAGGCATTGTCTCAGGTATGGTCTTA GTATTTTCTATTTGGAATGATAATTCACAATAC ATGAATTGGCTGGATTCTGGTAATGCAGGCCCT TGCTCCTCTACAGAAGGTAACCAAGCAATATA CTAGCTAATAACCCAAATACATCATGTTGTCTTT AGTAATATTAGATGGGGCGATATAGGTAGCAC TACGAACAGTACCGCACCTCCTCCTCCACCTGC TAGCTCCACGACATTTTCCACTACTAGAAGGTC CAGCACTACCAGCTCATCACCATCTTGTACTCA AACCCATTGGGGACAGTGTGGTGGTATAGGTTA CAGCGGTTGCAAACTTGCACATCTGGTACTAC ATGCCAATACAGTAATGACTATTACTACAATG TTTACCAGGTGCTGCGTCAAGTTCAAGTAGTgga tcc (SEQ ID NO: 20)	AB003694 (<i>Trichoderma reesei</i> EGI with secretion signal) Tggccaaatcgtgatcgtgatactcgtatcataag atggcacagatcgtactcttgattcacagacatccgta gccctcaagccgtttgcaagtcacaaacacaagcac aagcatagcgtcgaatgaagttccttcaagtcctcct gccctcataccggccgccctggcccaaacagctgt gaccagtgggcaacctcactggcaacggctacaca gtcagcaacaacctttggggagcatcagccggctctg gatttgctgctgacggcggtatcgtcagcggcg ggcctcctggcagcagactggcagtggtccggcg ccagaacaacgtcaagtcgtaccagaactctcagatt gccattccccagaagaggaccgtcaacagcatcagc agcatgccaccactgccagctggagctacagcggg agcaacatccgcgctaattgttgcgtatgactgttcacc gcagccaaccgaatcatgtcacgtactcgggagact acgaactcatgatctgtaagccataagaagtgcacct ccttgatagtttcgactaacaacatgtcttgaggcttg caaatacggcgataattgggccgattgggtcctcacag ggaacagtcaacgtcgggtggcagagctggacgctc tactatggctacaacggagccatgcaagtcattccttt gtggcccgagaccaactaccaactacagcggagat gtcaagaactcttcaattatctccgagacaataaagga tacaacgtcgcaggccaatatgttcttagtaagtcaccc tcactgtgactgggtcgtgattgttgcacgttgcataac aaaacctctgataggtaccgaattgttaccgagccct tcacggcgagtggaactctgaacgtcgcacctggac cgcatctatacaactaaaacctggaacgtgagatgtgg tgggcatacgttattgagcgagggaaaaaagcattg gatcc (SEQ ID NO: 29)	Qqpgtstpevh pkltykctksgg cvaqdsrvld wnyrwmhdan ynsctvngvnt tlcpeatcgkn cfiegvdyaasg vttgssltmnq ympsssgyys vsprlyllsdge yvmklngqels fdvdlalpcge ngslylsqmd ngganqyntag anygsydaq cpvqtfwmgtln tshqgfcnem dilegnsranalt phsctatacdsa gcfnpypsgy ksyypgdtvdt sktfiitqfntdn gspsgnlvstrk yqngvdiapsa qpggdttsscps asayglatmg kalssgmvlvfs iwndnsqymn wldsgnagpc stegnpsnilann pnthvfvsnirw gdigstnstapp pppasstfstrr sstsspsctqth wgqcggyysg ckctsgttcys ndyysqclpgaa sssss (SEQ ID NO: 39)	
Cellobiohyd rolase I from <i>Trichoderm a reesei</i> (TrCBHI); Linker 2; CWP2— reoptimized	C3_L2 _A1a	gagtcgccgggCAATCCGCTTGTACCCTACAATCCGA AACTACCCACCATTTGACCTGGCAAAAGTGTTT TAGCGGTGGAACCTTGTACTCAACAACTGGTTT TGTTGTTATCGACGCTAACTGGAGATGGACaCA CGCCACTAACTCTTCTACCAACTGTTACGACGG TAACACTTGGTCTTCCACTTTATGTCCAGATAA CGAACTTGTGCTAAGAATTGCTGTTTGGACGG TGCCGCTACGCTTCTACCTACGGTGTACCAC CTCCGGTAACTCCTTGTCTATTGGTTTCGTCAC CAATCCGCTCAAAAGAACGTTGGTGCTAGATTG TACTTGATGGCTTCTGACACTACTTATCAAGAA TTTACTTTGTTGGGTAACGAATTTTCTTTCGATG TTGACGTTTCCCAATTGCCATTTGGCTTGAACG GTGCTTTGTACTTTGTCTCTATGGATGCTGACG GTGGTGTCTTAAGTACCAACTAACACTGCCG GTGCTAAGTACGGTACTGGTACTGTGATTCTC AATGTCCACGTGACTTGAAGTTTCAATTAACGGTC AAGCCAACGTCGAAGGTTGGGAACCATCTCTC AACAAACGCTAACACGGTATCGGTGGTCAACGG TTCTGTGTTTCCGAAATGGACATCTGGGAAGC	X69976 (<i>T. reesei</i> CBHI; includes non-coding regions) gacattcaaggagatttagccagggtgcttgagt tatcgtgaaggaggtttgtctccgatacagcaat actgtatagtcactctgatgaagtggtccatattgaa atgtaagtcggcactgaacagggcaaaagattgagt gaaactgcctaagatctcggccctcggccctcgg ccttgggtgtacatgttctgctccggcaaatgca aagtgtgtaggatcgaacacactgctccttacc agcagctgagggtatgtgataggcaaatgttcagg gccactgcatggttgcgaatagaagagaagcttag ccaagaacaatagccgataaagatagcctcattaaa cggaatgagctagtaggcaaatgagcaatgtgt atatataaagggttcgaggtccgtcctcctcatgt ctccccatctactcatcaactcagatcctccaggaga ctgtacaccatcttttaggcacagaaccaatag tcaaccgaggactggcatcatgtatcggaagtggc cgatcatcggccctcttgccacagctcgtgctcag tcggcctgcactcctcaatcgagagactcaccgcct ctgacatggcagaatgctcgtctgtggcacgtgc	Qsactlqsethp pltwqkcssggt ctqqtgsvvida nwrwthansst ncydntwsstl cpdnetcakncc ldgaayastygv ttsngslsigfvq saqknvgarlyl masdtyqefill gnefsfdvdsq lpcngalyfvs mdadggyvsky ptntagakytg ydsqcpdrdkf ngqanvegwe pssnnantgigg hgscsemdiw eansisealtph cttvgeicegd	

	TAACAGTATTTCTGAAGCTTTGACACCACACCC ATGCACCACTGTGGTCAAGAAATTTGTGAAGG TGATGGATGTGGTGGAACCTACTCTGATAACAG ATACGGTGGTACTTTGTGACCCAGACGGTTGTGA CTGGAACCCATACAGATTGGGTAACACTTCTTT CTATGGTCCAGGTTCTTCTTTCACCTTGGATACC ACCAAGAAGTTGACTGTTGTTACCCAATTCGAA ACTTCTGGTGCTATCAACAGATACTACGTTCAA AACGGTGTACCTTCCAACAACCAAAACGCTGA ATTGGGTTCTTACTCTGGTAATGAATTGAACGA CGACTACTGTACCGCTGAAGAAGCTGAATTTGG TGGTTCCTCTTCTCCGACAAGGGTGGTTTGAC CCAATTCAAGAAGGCTACCTCCGGTGGTATGGT TTTGGTTATGTCCTTGTGGGATGATTACTACGC AAACATGTTATGGTTAGACAGTACTTACCCAAC TAACGAAACCTCCTCTACTCCAGGTGCTGTCAG AGGTTCCTGTTCTACCTCTTCTGGTGTTCAGCT CAAGTTGAATCTCAATCTCCAACGCTAAGGTC ACTTTCTCCAACATCAAGTTCCGTTCCAATCGT TCCACTGTGAATCCATCTGGTGGAACCCCTCCA GGTGGTAACAGAGGTACTACCACTACTCGTAG GCCAGCTACTACAAGTGGTTCTTCCCAGGCC AACCCAATCCCCTACTACGGTCAATGTGGTGGTAT CGGTACTCTGGTCCAACCGTCTGTGCTTCTGG TACTACCTGTCAAGTTTTAAACCCATACTACTC TCAATGTTTGCCTGGTGTGCTTCCAGTTCATCT AGTggatccGGTGGCGGTGGATCTGGAGGAGGCG GTTCTTGGTCTCACCCACAATTTGAAAAGGGTG GAGAAAACCTTGTACTTTCAAGGCGGTGGTGA GGTCTGGCGGAGGTGGCTCCggtcagctATCTCT CAAAACACCGACGGTCAAAATCCAAGCCACTAC CACAGCTACCACTGAAGCTACAACCTACCGCTGC TCCTTCATCTACTGTTGAACTGTTTCTCCATCT TCCACCGAAACCATCTCTCAACAAACCGAAAA CGGTGCTGCTAAGGCTGCTGTTGGTATGGGTGC ggcgcgcc (SEQ ID NO: 21)	actcaacagacaggctccgtggtcatcgacccaa ctggcgctggactcacgctacgaacagcagcaga actgctacgatggcaacacttggagctcgacctat gtcctgacaacgagactgctgcgaagaactgctgt ctggacgggtcccctacgctccacgtacggagt taccacgagcggtaacagcctctccattggcttgc acccagctcgcgagaagaacttggcgctcgctt taccttatggcgagcgacacgacctaccaggaattc accctgcttggcaacgagttcttcttgatgtgatgt tcgcagctgccgaagtacttaccatgaaccctg acgctatcttcttggctcccagctgactggccaat tcaaggtgctgcttgaacggagctcttacttctgt ccatggacgcggatgggtggcgtagcaagtatccc accaaacaccgctggcgcaagtagcgacgggt actgtgacagccagtgtcccgcgactgaagtcat caatggccaggccaacgttggggctgggagcgg tcatccaacaacgcgaacacggcgattggagaca cgggaagctgctgctgagatggatctgggagggc caactccatctccgaggtcttaccctccaccttgc acgactgtcggccgagagatctcgaggggtgatgg gtgcggcggaacttactccgataacagatatggcg gcacttgcgacccgatggctgcgactggaaccat accgcctggcgcaacacagcttctacggccctggc tcaagctttaccctcgataccaccaagaattgaccg ttgtaccaggttcgagacgtcgggtgccatcaacc gatactatgtccagaatggcgctacatttccagcgc caacgccgagcttggatgtacttgcgaacgagct caacgatgattactcacagctgagggagcagaatt cggcggtatctcttctcagacaaggcgccctgac tcagttcaagaaggtacctctgcgcgcatggttctg gtcatgagctgtgggatgatgtgatttgatggaca aacatgcgcttgacaagagtcaagcagctgact gagatgttacagtactacccaacatgctgtgctg gactccacctaccgacaacagagacctctccac acccgggtgccgtgcggaagctgctccaccagct ccggtgctccctgctcaggtcgaatctcagttcccaa cgccaaggtcaccttctccaacatcaagttcgacc cattggcagcaccggcaaccctagcggcggaac cctccggcggaaccgtggcaccaccaccacc ggcgccagccactactcactggaagctctcccgga cctaccagctcactacggccagtgccggcgtatt ggctacagcgccccacggtctgcgcagcgga caacttgcaggtctgaaccttactactctcagtg cctgtaagctcctgtgcgaaagcctgacgcaccgg tagattcttggtagccgtatcatgacggcgcg gagctacatggccccgggtgatttttttgaicta cttctgaccttttaataataggtcaactcatcttca ctggagatgcggcctgcttggattgcatgtgtga gcttggcaaatgtgcttgcgaacacacaaacgat tccttagtagccatgcattttaagataacggaataga gaaagaggaaataaaaaaaaaaaaaaaaaaca tccggtcataaccgtagaatcgccgctctctgta tccagtagccagcgcaaggtatttcatgatcgttca atgttgatattgtcccgccagtatggtccaccccc atctccgcaatctcttctcgaacgcgtgtggc gcgcaattggtaatgacccataggagacaac agcataatagcaacagtggaaattagtgcgcaata attgagaacacagtgaaccatagctggcgccgtg gaaagcactgttgagaccaactgtccgttgcgag gccaaactgcatgctgcaagacgatgacaacgta gccgagggacgtcacaaggacgcaagttgtcg cgatgaggtctccgtatggcatagccggcaat ccgagagtagcctcacaaggtggccttttgaaa ccggtaaacctgttcagacgtcctagccgagctc accgtaccagatcgaggattgacggcagaatagc	gcggtysdnryg gtdcpdgcdown pyrlgntsfygp gssfldtkkltv vtqfetsgainry yvqngvtfqpp naelgsysnel nddyctaeaeaf ggssfsdkggl qfkatstgsmvl vmslwddyya nmlwldstyptn etsstpgavrgsc stssgvpaqves qspnakvtfnsi kfgpigstgnps ggnpaggngrrt ttrpattgsspg ptqshyqcggi gysgptvcasgtt cqvlnpysqcl pgaasssssgsg ggsgsgsgsww hpqfekggenly fqggggsgsgg gsgsaisqitdgq iqattttatetta apsstvetvspss tetisqqtengaa kaavmgagal aaaamll (SEQ ID NO: 40)
--	---	--	--

			agtggtctccaggattgactggacaaaattcca gtattcccaggtcacagtgtctggcagaagtccttc tcggtgcgagtcgaaagtcgctatagtcgcaatg agagcacagtaggagaataggaacccgcgagcac attgtcaatccacatgaattggatgactgtgggc agaatgtgctgcctccaaaatcctgcgtccacaga tactctggcaggggttcagatgaatgcctctgggc ccccagataagatgcagctctggattctcggtacga tgatata (SEQ ID NO: 30) >YKL096W-A Chr 11 (CWP2 from Saccharomyces Genome Database) ATGCAATTCTCTACTGTCGCTTC CGTTGCTTTTCGTGCTTTGGCTA ACTTTGTTGCGCTGAATCCGCTG CCGCCATTTCTCAAACTACTGAC GGTCAAATCCAAGCTACTACCA CTGCTACCACCGAAGCTACCAC CACTGCTGCCCATCTTCCACCG TTGAAACTGTTTCTCCATCCAGC ACCGAACTATCTCTCAACAAA CTGAAAATGGTGCTGCTAAGGC CGCTGTCGGTATGGGTGCCGGT GCTCTAGCTGCTGCTATGTT GTTATAA (SEQ ID NO: 31)	
Cellobiohydrolase II from <i>Trichoderma reesei</i> (TrCBHII); Linker 3	C4_L3	gagtcggggGTCCATTAGAAGAAAGACAAGCCTG CTCCTCTGTTTGGGGTCAATGTGGTGGTCAAAA CTGGTCTGGTCCAACCTGTTGTGCTTCCGGTCT ACCTGTGTTTACTCCAACGACTACTATTCCCAA TGTTTGCCAGGTGCTGCTTCTCTTCTCTTCAA CTAGAGCTGCTTCTACAACTTCTAGGGTCTCCC CAACCACTTCCAGATCCTCTTCTGCTACTCCAC CACCAGGTCTACTACCACTAGAGTTCCACCAG TCGGTTCGGTACTGCTACTTACTCTGGTAACC CTTTCGTGCGGTGTTACTCCATGGGCTAACGCTT ACTACGCTTCTGAAGTATTCTTCTTTGGCTATCCC ATCTTTGACTGGTACTATGGCTACCGCTGCTGC TGCTGTCGCCAAAGTTCATCCTTCATGTGGTT GGACACCTTGGACAAAACCTCCATTAATGGAAC AAACCTTGGCAGACATAAGGACTGCTAACAAAG AACGGCGGTAACCTACGCTGGTCAATTTGTTGTG TACGACTTGGCAGACAGAGACTGTGCTGCTTTG GCTTCCAACGGTGAATACTCCATCGCTGACGGT GGTGTGCGCAAGTACAAGAACTACATTGATACC ATTAGACAAATCGTTGTGCAATACTCTGACATC AGAACCTTGTAGTCATCGAACCAAGATTCTTTA GCCAATTTAGTCAACCACTTGGGTACTCCAAAG TGTGCTAACGCTCAATCTGCCTACTTAGAATGT ATCAATTATGCAGTTACCAATTGAACCTTGCCA AACGTTGCTATGTACTTGGACGCTGGTCACGCC GGTTGGTTGGGTTGGCCAGCTAACCAAGACCCA GCCGCTCAATTATTCGCCAACGTTTACAAGAAT GCCTCTCTCCTAGAGCCTTGGTGGTTTGGCT ACTAACGTCGCTAACTACAACGGTTGGAACATC ACTTCTCCACCATCTTACACCCAAGGTAACGCT GTTTACAACGAAAAGTTGTACATTACGCTATC GGTCCATTATTGGCTAACCATGGTTGGTCTAAC GCCTTCTCATCACCACCAAGGTAGATCCGGT AAACAACCAACTGGTCAACAACAATGGGGTGA TTGGTGTAAACGTCATCGGTACTGGTTTCGGTAT CAGACCATCCGCTAACACTGGTGATTCTTGTG GGATTCTTCTGCTGGGTAAAGCCAGGTGGTGA ATGTGATGGCACCTCTGATTCTCTGCTCCAAG ATTCTGATTCCCACTGCGCTTCCAGACGCTTT GCAACCAGCCCCACAAGCTGGTGCATGGTTCCA	M16190 (T. reesei CBHII; includes non-coding regions) tcgaactgacaagttgtatattgctgtgtaccaagc gcgaatgtggacaggattaatgcagagttcattag cctcaagtagagcctatttctcgcggaagtcate tctcttattgcaatttctgccttcccaactcagggt gcagcgcacaactacagcaacatacactttatta gccgtgcaacaaggtattctacgaaatgctaca ctccacatgttaaaggcgcattcaaccagctctttatt gggtaataacagccagcgggggaatgaagtcatt agccggcactcaaggctatacaatgttccaactctc cgggctttatcctgtgtctccgaataccacgtga tgatgcttcagcgcaggaagtcacagacaccgcc tgtataaaaggggactgtgacctgtatgaggcgc aacatgtctcacagcagctcacctgaagaggcttg taagatcacctctgtgtattgacatgattgtcgcg attctaccacgtggtcagcgtggccacactcga gctagtgtgctctagaggagcggcaagctgtcga agcgtctggaattatgtgaacctctcaagagacc aaatactgagatatgtcaagggccaatgtgtgga cagaattggtcgggtccgactgtgtgtctccgga gcacatgctgtactccaacgactattactccagtg tcttcccgctgctgcaagctcaagctcgtccacgcg cgccgctgcagcactctcagtatccccacaac atcccggtcagctcgcgacgctccacctggttc tactactaccagagattcctcagtcggatcggaac cgctacgtattcaggcaacctttgttgggtcactc cttggccaatgcatattacgctctgaagttagcag cctcgtattcctagcttgactggagccatggccact gtgcagcagctgtcgaaggttccctctttatgtg gtgtaggtcctcccggaaccaaggcaatctgttact gaaggtcatcattcactgcagagatactcttgaca gacctctcatgagcaaaccttggccgacatccg caccgccaacaagaatggcggttaactatgccggac agtttgtgtgtatgacttgcggatcgcgattcgct gccttgcctcgaatggcgaatactctattgccgatg gtggcgtgcgcaaatataagaactatcgcaccca ttcgtcaaatgtctggaatattccgatatccggacc ctcgtgtattgtgtggttaaacactgcctcccc cccccttcccttcttcccgccgcatctgtctgt	Vpleerqacssv wgqcgqgnws gptccasgstcv ysndyysqclpg aassssstraast srvspttsrssat pppgsttrvpp vsgtatysgnp fvytpwanay yasevsslaipsll gamataaaava kvpsfmwldtl dktplmeqtlad irtanknggny gqfvvydlpdrd caalasngeysia dggvakyknyi dtrjqvveysdi rllviedpslanl vtnltpkcana qsaylecinayv qnlplnvamyl daghagwlgw panqdpaaqlfa nvyknasspral rglatnvanyng wnitsppsytqg navynekiyiha igpilanhwgsn affitdgrsgkq ptgqqqwgdw cnvigtfgirps antgdsldsfv wvkgpgecdgt sdssapfrdshc alpdalqpapqa gawfqayfvql tnanpsflsgsg gsgsggshhh hhhggenlyfq

		AGCTTACTTTGTCCAATTGTTGACCAACGCTAA CCCATCTTTCTTGggtaccGGTGGCGGTGGATCTG GTGGAGGCGGTTCTCATCACCACCATCATCAG GTGGCGAAAACTTGTACTTTCAAGGCGGCGGTG GAGGTAGTGGAGGAGGTGGCTCCggctcagct (SEQ ID NO: 22)	gctaactattgtccctctccagagcctgactctctg ccaacctggtagcaaacctcggtagctcaaatgtg ccaatgctcagtcagcctaccttgagtgcatcaacta cgccgtcacacagctgaaccttccaatgttgcgat gtatttgacgttgccatgcaggatggctggctg gccggcaaaccaagaccggccgctcagctatttg caaatgtttacaagaatgcatcgtctccgagagctctt cgcgattggcaaccaatgtcgcaactacaacgg gtggaacattaccagcccccatcgtacacgcaag gcaacgctgtctacaacgagaagctgtacatccacg ctattggacctcttctgccaatcacggctgtccaac gccttctcatcactgatcaaggtcgatcgggaaagc agcctaccggacagcaacagtgaggagactgtg caatgtgatcgccagcgatttgattcgccatcc gcaaacactggggactgtgtggtgattctgttct gggtcaagccagcgccgagtgtagcgccaccag cgacagcagtgccacgatttgactccactgtgc gtcccagatgccttgaacggcgccctcaagtg gtgtgtgtccaagcctacttctgacgttctcaca aacgcaaacccatcgttctgtaaggcttctgtgacc gggttcaaacatgatgtgcgatgggtgtgttccc gggtggcggagcttctgtacttctgtgt (SEQ ID NO: 32)	ggggsgsgsgs gsa (SEQ ID NO: 41)
Linker 1; CWP2 original optimization	L1_A I	ggatccGGAGGTGGTTCAGGAGGTGGTGGGTCTGC TTGGCATCCACAATTTGGAGGAGGCGGTGGTG AAAATCTGTATTTCCAGGGAGGCGGAGGTGATT ACAAGGATGACGACAAAGGAGGTGGTGGATCA GGAGGTGGTGGCTCCggctcagctATTAGCCAAATA ACTGATGGTCAAAATACAAGCAACTACAACAGC AACAACCGAAGCTACTACCACAGCCGCGCCTTC TTCAACTGTTGAGACTGTTAGTCCTTCCCTCCAC GGAAACGATTTCTCAACAGACTGAAAAACGGTG CAGCCAAAGCAGCAGTCGGCATGGGTGCCGA GCCCTAGCAGCTGCAGCAATGCTTTTGTAAggcg cgcc	>YKL096W-A Chr 11 (CWP2 from Saccharomyces Genome Database) ATGCAATTCTCTACTGTCTGCTTC CGTTGCTTTTCGTCGCTTTGGCTA ACTTTGTTGCGCTGAAATCCGCTG CCGCCATTTCTCAAACTACTGAC GGTCAAATCCAAGCTACTACCA CTGCTACCACCGAAGCTACCAC CACTGCTGCCCCATCTTCCACCG TTGAAACTGTTTCTCCATCCAGC ACCGAAACTATCTCTCAACAAA CTGAAAATGGTGCTGCTAAGGC CGCTGTCGGTATGGGTGCCGGT GCTCTAGCTGCTGCTGCTATGTT GTTATAA (SEQ ID NO: 33)	Gsggsgsgsgs awhpqfggggg enlyfqqgggd ykdddkggggs gggsgsaisqit dgqiatatttea ttaapsstvetvs psstetisqten gaakaavgmga galaaaamll (SEQ ID NO: 42)
Xyn2 secretion signal + spacer	S06	gaattcttaattaaAAACAAAATGGTCTCCTTCACCTCC CTGCTGGCCGGCGTTGCCGCTATCTCTGGTGTG CTAGCAGCCCCTGCCGCGAAGTTGAACCTGTG GCAGTTGAGAAACGTGAGGCCGAAGCAGAAGC Tcccgggactc (SEQ ID NO: 23)	U24191 (T. reesei endo-beta-1,4- xylanase) Caacatggtctcctcactcctcctcgcggcgct cgccgccatctcggcgctctggccgctcccgccg ccgaggtcgaaccctgtgctgtgagaagcgc (SEQ ID NO: 34)	Mvsfslilagva aisgvlaapae vepvavekreae aea (SEQ ID NO: 43)
Mfalpha pre/pro secretion signal + spacer	S04	gaattcttaattaaAAACAAAATGAGATTTCCATCAATA TTTACAGCAGTTTTGTTTGGCGCGAGTTCAGCC CTTGCAGCACCCGTCAATACCACGACGGAGGA TGAGACAGCCCAGATCCCAGCAGAGGCTGTGA TAGGATATTTAGACCTGGAAGGCGATTTTGATG TGGCCGTATTACCGTTTTCTAACTCTACGAATA ATGGATTGTTATTTATTAATACTACAATTGCCTC TATAGCCGCAAAGGAAGAAGGGGTGTCTTTAG ATAAGAGAGAAGCTGAGGCTGAAGCCcccgggact c (SEQ ID NO: 24)	>YPL187W Chr 16 (MF alpha from Saccharomyces genome database) ATGAGATTTCTTCAATTTTAC TGCAGTTTTTATTCGCAGCATCCT CCGCATTAGCTGCTCCAGTCAA CACTACAACAGAAGATGAAACG GCACAAATCCCGCTGAAGCTG TCATCGTTTACTTAGATTTAGAA GGGATTTTCGATGTTGCTGTTTT GCCATTTTCCAACAGCACAAAT AACGGGTTATTGTTTATAAATA CTACTATTGCCAGCATTGCTGCT AAAGAAGAAGGGGTATCTTTGG ATAAAAGAGAGGCTGAAGCTTG GCATTGGTTGCAACTAAAACCT GGCCAACCAATGTACAAGAGAG AAGCCGAAGCTGAAGCTTGGCA TTGGCTGCAACTAAAGCCTGGC CAACCAATGTACAAAAGAGAAG CCGACGCTGAAGCTTGGCATTG GCTGCAACTAAAGCCTGGCCAA	Efliknkmrfps iftavlfaassala apvntttedetaq ipaeavigyldle gdfdvavlpfsn stnngllfintias iaakeegvsldk reaeaeapgt (SEQ ID NO: 44)

			CCAATGTACAAAAGAGAAGCCG ACGCTGAAGCTTGGCATTGGTT GCAGTTAAAACCCGGCCAACCA ATGTACTAA (SEQ ID NO: 35)	
Hybrid killer toxin/hIL- 1b + consensus kex2 and spacer	S16	gaattctaattaaAAACAAAATGAATATATTTTATATT TTCTATTTCCTTTATCATTTGTGCAGGGATCAT TAAATTGTACATTAAGAGATTACAACAAAAGT CTTTAGTAATGTCAGGTCCATATGAATTTAAAG CATCCCTTGATAAAAGGGAAGCCGAAGCCGAA GCTcccgggactc (SEQ ID NO: 25)	HIL-1B (accession #E11934; full sequence) cttattacag tggcaatgag gatgactgt tctttgaagc tgaaggccct aaacagatga agtgtctctt ccaggacctg gacctgtgcc ctctggatgg cggcatccag ctacgaatct ccgaccacca ctacagcaag ggcttcaggc aggccgcgtc agttgtgtg gccatggaca agctgaggaa gatgtgtgt ccttgcacac agacctcca ggagaatgac ctgagcacct tcttccctt catcttgaa gaagaacct tcttctga cacatggat aacgaggctt atgtgcacga tgcacctga cgtactga actgcacgt cgggactca cagcaaaaa gcttggatgt gtctgttcca tatgaactga aagctctcca cctccaggga caggatatgg agcaacaagt ggtgttctc atgtccttg tacaaggaga agaaagtaac gacaaatac ctgtggcctt gggcctcaag gaaagaaac tgtacctgc ctgcgtgtg aaagatgata agcccaactc acagctggag agttagatc ccaaaaatta ccaaaagaag aagatggaaa agcgattgt cttcaacaag atagaatca ataacaagct ggaattgag tctgccagt tccccaaact gtacatcagc acctctcaag cagaaaaacat gccgtcttc ctggaggga ccaaaggcgg ccaggatata actgactca ccatgcaatt tgtgtctcc taaagagagc tgtaccaga gagtctgtg ctgaatgtg actcaatccc taggctgagc agaaaggga cagaagggtt ttgagtacgg ctatagcctg gacttctctg ttgtctacac caatgccc ctgcctgct tagggtagt ctaagacgat ctctgtcca tcagccagga cagtcagtc tctcttca gggccaatcc cagcccttt gttgagccag gcctctctc cactctctc actcactaa agccgcctg acagaaacca ggccacattt tggttctaag aaacctctc ctgtcattcg ctccacatt ctgatgagca accgctccc tattattta ttatttgt ttgtgttt gattcattgg tctaattat tcaaggggg caagaagtag cagtgtctg aaaagagcct acttttatt agctatgaa tcaattcaat ttggactggt gtgtctctt taaatcaagt cctttaatta agactgaaaa tatataagct cagattattt aaatgggaat attataaat gagcaaatat catactgttc aatgttctc aaataaactt cactaaaaa aaaaaaaaaa aaaaaaaaa aaaa (SEQ ID NO: 36) K. lactis killer toxin (accession #M26799; full sequence) ataaaatgaa tatattttac atattttgt tttgtctg attcgttcaa ggtttggagc atactcatg aagaggctcc ttagtcaaaa gagcagtatg ttatgacact gatcaagtc cacttaatat tttcttgggt cctccagata agaagaaaag agattac (SEQ ID NO: 37)	Mnifyiflflsfv qgslnctlrdsq kslvmsgpyel kasldkreaeae a (SEQ ID NO: 45)
Kjeldsen synthetic + spacer	S17	gaattctaattaaAAACAAAATGAAGTTGAAGACTGTT AGGTCAGCCGTTTTGAGTAGTTTATTTGCCCTCTC AAGTCTTGGGTCAACCAATTGATGATACGGAA	Kjeldsen (1997)—synthetic construct	Mklktvrsavls slfasqvlqpid dtesntsvnlm

[illegible]

	caggaacaaacgggttgccaacagacgagactgttattgtgtttaaactccaa ccactgcaatlagccttctcattgagttccagttctagtggtcaaatfactagcagc ggatcc (SEQ ID NO: 27)	aattgaccacagtcaccggtagcaaatggttgccaa ctgatgagacatcattgtcatcagaacaccaacaa cagccactactgcatgactacaactcagccatgga acgacacttttacccttacttctaccgaattgaccaca gtcaccggtaccaatggttgccaactgatgagacc atcattgtcatcagaacaccaacacagccactact gcatgactacaactcagccatggaacgacactttta cctctacttctaccgaattgaccacagtcaccggtac caatggttgccaactgatgagacatcattgtcatca gaacaccaacacagccactactgcatgactaca actcagccatggaacgacacttttaccctctacatcca ctgaaatcaccaccgtcaccggtaccaatggtttgc caactgatgagaccatcattgtcatcagaacaccaa caacagccactactgcatgactacacctcagccat ggaacgacacttttaccctctacatcactgaaatgac caccgtcaccggtaccaacgggttgccaactgatga aacatcattgtcatcagaacaccaacacagccac tactgcatcactacaactgagccatggaacagcac ttttaccctctacatcactgaaatgaccaccgtcaccg gtaccaacgggttgccaactgatgaaacatcattgt catcagaacaccaacacagccactactgcatcaca ctacaactcagccatggaacgacacttttaccctac atccactgaaatgaccaccgtcaccggtaccaacg gtttgccaactgatgaaacatcattgtcatcagaac accaacacagccactactgcatgactacaactca gcatggaacgacacttttaccctctacatcactgaa atcaccaccgtcaccggtaccaccgggttgccaact gatgagaccatcattgtcatcagaacaccaacaa gccactactgcatgactacaactcagccatggaac gacacttttaccctctacatcactgaaatgaccaccgt caccggtaccaacggcgttccaactgacgaaacg tcattgtcatcagaactccaactagtgaaggctaatc agcaccaccactgaaacatggactggtactttacct ctacatcactgagatgaccaccgtcaccggtacta acgggtcaaccaactgacgaaacgggtattgtatca gaactccaaccagtgaagggttggttacaaccacca ctgaaccatggactggtacttttactctacatcactg aatgaccaccatcattggaaccaacggcgttccaa ctgacgaaacggatggtcatcagaactccaacca gtgaagggtcattcagcaccaccactgaaccatgga ctggtacttttactctacatcactgaaatgaccacca ttactggaaccaatggtcaaccaactgacgaaacgg ttattgtatcagaactccaactagtgaagggtcattc agcactacaacggaaacatggaccggtactttcact tctacatcactgaaatgacgcacgtcaccggtacca acggcgttccaactgacgaaacggatgtcatca gaactccaaccagtgaagggtcattcagcaccacc actgaaccatggactggcactttcacttcgactcca ctgaggttaccaccatcactggaaccaacgggtcaac caactgacgaaactgtgattgtatcagaactccaac cagtgaaagggtcattcagcaccaccactgaaccatg gactggtactttcactctacatcactgaaatgacca ccgtcaccggtactaacgggtcaaccaactgacgaa accgtgattgttatcagaactccaaccagtgaagggt tggttacaaccaccactgaaccatgactggtactttt acttcgacttcactgaaatgtctactgtcactggaac caatggcttgccaactgatgaaactgtcattgttgc aaactccaactactgcatcctcactcagttgtcatca tcatctcaggacaaatcaccagctctacatcgtcttc gcgtccaattattacccattctatcctagcaatggaa cttctgtgatttctcctcagtaatttctcctcagtcact tcttcttattcacttctcctcagtcatttctcctcagtc atttcttcttacaacaacctccacttctatatttctga atcatcacaatcaccgtcattccaaccagttagttcca cctctggtcttctgagagcgaaacgagttcagctgg	tstemttvtgtn qptdetvivrpt seglvtttpepw gtfststemttit gtngvptdetvi virtptseglstt epwtgtfstste mttittgngqptd etvivrptsegl isttpewtgtfst stemthvtgtn vptdetvivrpt seglsttpepw gtfststevttitg tngqptdetvivi rptseglsttpe wtgtfststemtt vtgtnqptdet vivrptseglvt ttpewtgtfstst emstvtgtnlpt detvivvktptta isslsssssgqit ssgs (SEQ ID NO: 47)
--	---	---	--

[illegible]

Table 4: Synthetic cellobiohydrolase (CBH) genes constructed

Donor organism/ Gene	DNA sequence used	Accession number and amino acid sequence
<i>Humicola grisea cbh1</i>	GAATTCATGAGAACCGCTAAGTTCGCTACCTTGGTGCCTTGGTTCCTCTGCGTCTCAACAAAGCCTGTTCTGACTACTGAACGTCACCCATCTTGTCTTG GAACAAAGTGTACTGCTGGTGGTCAATGTCAAACTGTCCAAGCCTCCATCACTT TGGACTCTAATTGGAGATGGACCCCAAGTCTCTGGTAGTAACTGTTAC ACCGGTAATAAGTGGACACTTCTATTGTACTGACGCTAAGTCTTGTGCTCA AAATTGTTGTTGATGGTCTGATTACACCTCCACTTATGGTATTACCAACA ACGGTGACTCTTTGCTTGAAGTTCGTTACTAAAGGTCAACATTCCACCAAC GTCGGTCTAGAACCTACTTAATGGACGGTGAAGACAAGTACCAACCTTCG AATTGTTGGTAATGAATTAACCTTCGATGTCGATGTCTAAACATCGGTTGT GGTTGAACGGTCTTTATCTTCGTTTCTATGGACCGCGACGGTGGTTGTCT CGTTACCCAGGTAATAAGGCTGGTGCCAAAGTATGGTACCGGTACTGTGATGC TCAATGCCCAAGAGACATTAAGTTCTCAACGGTGAAGTAAACATTGAAGGT TGGACTGGTCTACCAACGACCCCAACCGTGGCGCGGTAGATACGGTACCT GTTGTTCCGAAATGGACATTTGGGAAGCCCAACAACATGGCTACTGCTTTTACT CCACACCATGTACCATCATTTGGTCAATCCAGATGTGAAGGTGACTCCTGTGG CGGTACCTACTCCAAAGAAAGATACGCTGGTGTGTTGTATGCCAGACGGTTGTG ACTTCAACTCCTACAGACAAGGTAACAAAGACTTTCTATGGTAAGGATGACT GTCGATACCCAAAGAGATCACCGTGTGTCACCCCAATCTTGAAGGACGCTA ACGGTGATTTAGGTGAAATTAAGAGATTCTACGTTGTGAAGTAATGCCATCACTCAAG CCCAAACTCTGAATCAATCCAGTTGCTTGGTGTGATATTGACGACTTCAACAG ACTGGTGTGACAGACAAAGTTGCTTGGTGAAGGTTTGGCCGGTCCAAATGGTCTTG AAAGGTGGTATGAAGCAATGGGTAAAGGTTTGGCCGGTCCAAATGGTCTTG GTTATGTCTATTGGGACGATCACGCTTCCAAACATGTTGTGGTTGGACTCCAC CTTCCAGTTGATGCTGTGTAAGCCAGGTGCCGAAAGAGGTGCTTGTCCAA CTACTCCGGTGTCCAGCTGAAGTTGAAGCCGAAAGCTCCAAATCTTAAACGTT GTCTTCTCTAACATCAGATTCCGGTCCAATCGGTTCCACAGTCTGTTTGGC AGGTGCTGGTAATGGTGGTAATAACGGTGGTAACCCACCAACCAACCACT ACCACTTCTTGTGCCCCAGTACTACCAACCGCTTCTGCTGTGTCCTCAAGGC TGGTAGATGGCAACAATGTGGTGGTATTGGTTTCAACCGTCCAAACCAATGTG	Accession No.: CAA35159 MRTAKFATLAALVASAA AQACSLTTERHPSLSWN KCTAGGCQTVQASITLD SNWRWTHQVSGSTNCYT GNKWDTSICTDAKSCAQ NCCVDGADYTSTYGITIN GDSLSLKFVTKGQHSINV GSRTYLMGDEKDYQTFEL LGNEFTFDVDVSNIGCGL NGALYFVSMDADGGLSR YPGNKAGAKYGTGYCDA QCPRDIKFINGEANEIWT GSTNDPNAGAGRYGTCCS EMDIWEANNMATAFTPH PCTIIGQSRCEGDS CGGT SNERYAGVCDPDGCDPNS YRQGNKTFYKGKMTVDY TKKITVVTQFLKDA NGDL GEIKRFVYQDGKIIPNSES TIPGVEGNSITQDWCDRQ KVAFGDIDDFNRKGGMK QMGKALAGPMVLVMSIW DDHASNMLWLDSTFPVD AAGKPGAERGACPTTSGV PAEVEAEAPNSNVVFSNIR FGPIGSTVAGLPGAGNGG NNGGNPPPTTTTSSAPAT

	AAGAACCATACATCTGTACCAAGTTGAACGACTGGTACTCTCAATGTTTATAA CTCGAG (SEQ ID NO: 7)	TTTASAGPKAGRWWQQG GIGFTGPTQCEEPYICTKL NDWYSQCL (SEQ ID NO: 11)
<i>Thermoascus aurantiacus cbh1</i>	GAATTCATGTACCAAGAGCTCTATTTGTTCTCTTCTTCTTGGCCGCGCTAG AGCTCATGAAGCCGGTACTGTACCGCCGAAACACACCCATCCTTGACTTGGC AACAAATGTTCTCTGTTGTTCTGTACTACTCAAAACGGGAAAGGTTGTTAT GACGCTAACTGGAGATGGGTTACACACTACTCCGGTTACACCAACTGTTACAC TGGTAACACATTGGGATACCTTCCATCTGTCCAGACGACGTTACTGTGCTCAAA ACTGTGCTTTGGACGGTGTGACTACTCCGGTACTCCGGTGTCACTACCTCT GGCAACGGTTGAGATTGAACTTGTCAACCACTCTTCTGGTAAGAACATCGG TTCTAGATTGTAATTGTTGCAAGACGATACCTTACCAATCTTCAAGTTGT GGTCAAGAGTTCACTTTCGACGTTGATGTTTCCAACTTGCCTTGTGGTTGA ACGGTCTTTGTACTTCTGCTATGGACGGCAGGTAACCTATCCAAAGTAC CCAGGTAAACAGCCGGTCCAAAGTACGGTACCGGTTACTGTGATCTCAAT GTCCAAGAGACCTAAATTCATTAAACGGTCAAGCTAACGTCGAAGGTTGGCA ACCATCTGCTAACGATCCAAACGGCGGTGTCGTAATCACGGTTCCTCCTGTG CTGAAATGGACGTTTGGGAAGCTAACTCTATCTCCACCGCGTCACTCCACAT CCATGTGATACCCAGGTCAAAACCATGTGTCAAGGTGATGATTGTGGTGATC CTACTCTCCACTAGATACGCTGTACCTGTGACACCGACGTTGTGATTCTCA ACCCATACCAACGAGTAACCACTTTCTACGTCAGGTAAGATTGTGAT ACTTCTTCAAGTTCACTGTTGTCACTCAATTCATTACCGACGATGTTACCCCA TCTGTTACCTTAAGTAAATTAAGAGATTCTACGTCGTAACCGTAAAGTCAAT TCCACAATCCGAAAGCACCATTTCCGGTGTACCGGTAACTCCATCACCACCTG AATACTGTACCGTCAAAAGCGCGCTTTGACAAACCGGTTTCTTCAACCCAT GGTGGTTTGCAAAAGATTCTCAAGCCTTGGCTCAAGGTATGGTTTGTGTCAT GTCCTTGTGGGATGACCAACGCTGTCAACATGTTGTGGTTGGAATCTACTACC CAACTGACGCTGATCCAGACACCCAGGTGTGTGCTAGAGTACTGTCCAAACC ACTTCTGGTGTCCAGCTGACGTCGAACTCAAAACCCCTAACTCTTACGTTAT CTACTCTAACATCAAGGTGGTCCAAATTAACCTCCACCTTCACTGCTAACTAAC TCGAG (SEQ ID NO: 8)	Accession No.: AAL83303AAL16941 MYQRALLFSFFLAARAHA EAGTVTAENHPSLTWQQ CSSGSCCTTQNGKVVDA NWRWVHTTSGYTNCTG NTWDTSCPDVVTCAQNC ALDGADYSGTGVTTS NALRLNFVTQSSGKNIGS RLYLLQDDTTYQIFKLLG QEFTFDVDSNLPCLNG ALYFVAMDADGNLSKYP GNKAGAKYGTGYCDSQC PRDLKFNQANVEGWQP SANDPNAGVGNHGSSCA EMDVWEANSISTAVTPHP CDTPGQTMCOGDDCGGT YSSTRYAGTCDTIDGCDNF PYQPGNHSFYGPGKIVDT SSKFTVVTQFITDDGTSPG TLTEIKRFYVQNGKVIPQS ESTISGVTGNSITTEYCTA QKAAFDNTGFFTHGGLQ KISQALAQGMVLVMSLW DDHAANMLWLDSTYPTD ADPDTPGVARGTCPTTSG VPADVESQNPNSYVIYSNI KVGPINSTFTAN (SEQ ID NO: 12)
<i>Talaromyces emersonii cbh1</i>	GAATTCATGCTAAGAAAGAGCTTACTATTGAGCTCTTCTGCTATCTTGGCCGT TAAGGCTCAACAAGCCGGTACCGCTACTGCTGAAACACACCCCTCCATTGACCT GGCAAGAAATGTACCGCTCCAGGTTCTTGTACCAACCCAAACCGGTGCTGCTC TTGGACGCTAACTGGAGATGGGTCCACGACGTCACCGTTACACTAACTGTT ACACCGGTAAACACCTGGGACCCCACTTACTGTCCAGACGACGAACTTGGCGC	Accession No.: AAL89553 MLRRALLSSSAILAVKA QQAGTATAENHPPLTWQ ECTAPGSCCTQNGAVVLD

	TCAAACTGTGCTTGACGGTGTGCTGACTACGAAGTACTTACGGTGTACCT CCTCTGTTCTTCTTGAAGTTGAAGTTCGTCACTGTTCTTAACGTCGGTTCCA GATTGATTTGTTGCAAGATGACTCCACTTACCAAACTTCAAGTTGTGAAC AGAGAAATTTCTTTCGACGTCGATGTGTCCAACTTGCTTGTGGTTTGAACGG TGCTCTATACCTCGTTGCTATGGACGCTGATGGTGGTGTTCACCAAGTACCCAA ACAACAAGGCTGGTGCCAAATACGGTACTGGTACTGTGACTCTCAATGTCCA CGTGACTTGAAATTTATGATGGTGAAGCTAATGTGCAAGTTGGCAACCATC TTCTAACACGCTAACACTGGCATCGGTGACCAAGGTTCTTGTCTGTGCCGAAA TGGACGTTTGGGAAGCCAACTCCATTTCCAAACGCCGTCACCTCCACACCCATGT GACACTCCAGGTCAAACTATGTGTTCCGGCGATGACTGTGGTGGTACTTACTC TAACGATAGATACGCTGGTACCTGTGATCCAGACGTTGCGACTTCAATCCAT ACAGAAATGGGTAACACTTCTTTACGGTCCAGGCAAGATCATCGACACTACT AAGCCATTCACTGTTGTCAACCCAACTTCTTGACCGACGATGGTACTGATACCGG TACTTTGTCCGAAATCAAGAGATTTCTACATCCAAACTCTAACGTCATCCAC AACCAATTCGACATCTCTGGTGTCACTGGTAACCTCCATTACCAACCGAATTT TGTAACGCCCAAAAGCAAGCTTTCGGTGACACCGACGACTTCTCTCAACACG GTGGTTTGGCTAAGATGGGTGCTGCTATGCAACAAAGGTATGGTTTGGTCATG TCTTTGTGGACGACTACGCTGCTCAAAATGTTGTGGTGGACTCCGATTACCC AACCGATGCCGACCCAAACCCCTGGTATCGCTAGAGTACCTGTCCAACT GACTCTGGTGTTCATCTGACGTGCAATCCCAATCTCCAAACTCCTACGTCAC TTACTCCAAACATTAATTT GGTCCAAATCAACTCCACTTTTCACTGCTTCTTAACTCGAG (SEQ ID NO: 9)	ANWRWVHDVNGYTNCY TGNTWDPTYCPDDETC QNCALDGDADYEGTYGVT SSGSLKLNFTVSGSNVGS RLYLLQDDSTYQIFKLLN REFSFDVDSNLPFCGLNG ALYFVAMDADGGVSKYP NNKAGAKYGTGYCDSQC PRDLKFIDGEANVEGWQP SSNNANTGIGDHGSCCAE MDVWEANSISNAVTPHPC DTPGQTMCSGDDCGGTY SNDRYAGTCDPDGCDENP YRMGNTSFYGPCKIIDTT KPFTVVVTQFLTDDGTDTG TLSEIKRFYIQNSNVIPQPN SDISGVTGNSITTEFCTAQ KQAFGDTDDFSQHGGLA KMGAAMQQGMVLVMSL WDDYAAQMLWLDSDYP TDADPTTPGIARGTCPTDS GVPSDVESQSPNSYVTYS NIKFGPINSTFTAS (SEQ ID NO: 13)
<i>Talaromyces emersonii cbh2</i>	Accession No.: AAL78165 MRNLLALAPAALLVGAA EAQQLWGQCGGSSWTG ATSCAAGATCSTINPYA QCVPATATPTTLTTTKPT STGGAAPTTPPTTTGTTT SPVTRPASASGNPFEGY QLYANPYAYASEVISLAIPS LSSELVPKASEVAKVPSFV WLDQAAKVPMSGDYLLKD IQSQNAAGADPPIAGFVV YDLPRDCAAAASNGEFS IANNGVALYKQYIDSIREQ LTTYSDVHTILVIEPDSLA	

	TACCCAATTGGACTTGCCAAACGTTGCTATGTACTTGGACGCTGGTCAACGCCG GTTGGTTGGTTGGCAAGCCAACTTGGCCCCAGCTGCTCAATTAATTCGCTTCT GTTTACAAGAACGCTCTTCCCCAGCTCTGTAGAGGTTGGCTACCAACGT GGCTAACTACAACGCTGGTCCATTCTAGATGTCCATCCTACACTCAAGGTG ACGCTAACTGTGATGAAGAAGATTACGTTAACGCTTTGGGTCCATTGTTCCAA GAACAAGGTTTCCAGCTTACTTTCATCATCGACACTTCCCGTAAACGGTGTACG ACCAACTAAGCAATCTCAATGGGTGACTGGTGTAACTGTTATTGGTACCGGTT TCGGTTAGACCAACACCGACACTGGTAAACCCATTGGAAGACGCTTTCGTT TGGTCAAGCCAGGTGGTGAATCCGACGGTACTCCCAACACTACTAGCCAC GTTACGATTACCACTGTGTTTGTCTGACGCTTGTGCAACCAAGCTCCAGAAAGCT GGTACCTGGTTCGAAGCCTACTTCGAACAAATTGTTGACTAAGGCCAACCCATT GTTCTAACTCGAG (SEQ ID NO: 10)	NVVTNLNVPKCANAQDA YLECINYAITQLDLPNVA MYLDAGHAGWLGWQAN LAPAAQLFASVYKNASSP ASVRGLATNVANYNAWS ISRCPSYTOGDANDEED YVNALGPLFQEQGFAYF IIDTSRNGVRPTKQSQWG DWCNVIGTGFVRPTTDT GNPLEDAFVWVKPGGES DGTSNTTSPRYDYHCGLS DALQPAPEAGTWFAQYFE QLLTNANPLF (SEQ ID NO: 14)
<i>Trichoderma reesei cbh1</i>	ATGTTCTCTTCACTCCCTGCTGGCCGGGCTTGCCTATCTCTGGTGTCTTA GCAGCCCCTGCCGCAAGTTGAACCTGTGCGCAGTTGAGAAACGTGAGGCCG AAGCAGAAAGCTCAATCCGCTTGTACCTCAATCCGAAACTCACCCACCATTG ACCTGGCAAAAGTGTCTACGGTGGAACTTGTACTCAACAAACTGGTCTGT TGTTATCGACGCTAACTGGAGATGGACACACGCCCACTAACTCTTCTACCAACT GTTACGACGGTAACACTTGGTCTTCCACTTATGTCCAGATAACGAAACTTGT GCTAAGAAATTGCTGTTGGACGGTGGCCCTACGCTTCTACCTACGGTGTAC CACCTCCGGTAACTCCTTGTCTATTGTTTGTCTGCTCACTCAATCCGCTCAAGA ACGTTGGTGTAGATTGTACTTGTGCTTCTGCTGCTGCTGCTGCTGCTGCTGCT ACTTGTGGGTAAACGAAATTTCTTCTCGATGTTGACGTTTCCCAATTGCCATGT GGCTTGAAACGGTGTCTGCTTGTCTTCTGCTGCTGCTGCTGCTGCTGCTGCT AAGTACCCAACTAACACTGCCGTGCTAAGTACGGTACTGTTACTGTGATTC TCAATGTCCACGTGACTTGAAGTTCAATTAACGGTCAAGCCAAACGTCGAAAGTT GGGAACCATCTCCAAACGCTAACACCGGTATCGGTGGTCAAGGTTCCCTGT TGTTCCGAAATGGACATCTGGGAAGCTAACAGTATTTCTGAAGCTTTGACACC ACACCCATGCACCACTGTCCGTCAAGAAATTTGTGAAGGTGATGGATGTGT GGAACCTACTCTGATAACAGATACGGTGGTACTTGTGACCCAGAGGTTGTG ACTGGAAACCCATACAGATTGGGTAACTTCTTCTATGTTACCCAAATTCGAAACTTC TTACCTTGGATACCAACCAAGATTGACTGTGTTGTTACCCAAATTCGAAACTTC TGGTGTATCAACAGATACTACGTTCAAAACGGTGTCACTTCCCAACCAACCAA ACGCTGAAATTGGGTCTTACTCTGTTAATGAATTGAACGACGACTACTGTACC GCTGAAGAAGCTGAATTTGGTGGTCTCTTCTCCGACAAAGGTGGTTTGAC CCAATTCAAGAAGGCTACCTCCGGTGGTATGTTTGTGTTATGTCCTTGTGGG ATGATTACTACGCAAAACATGTTATGGTTAGACAGTACTTACCCCAACTAACGAA ACCTCCTCTACTCCAGGTGCTGTGTCAGAGGTTCTGTCTTCTACTCTTCTGTGTT	Accession No.: CAA49596 MVSFTSLLAGVAAISGVL AAPAAEVEPVAVEKREAE AEAQSACTLQSETHPLT WQKCSSGGTCTQQTGSV VIDANWRWTHATNSSIN CYDGNWSSTLCPDNETC AKNCCLDGAAAYASTYGV TTSGNLSIGFVTSQAQKN VGARLYLMASDTTYQEFT LLGNEFSFDVDVSQLPCG LNGALYFVSMADAGGVS KYPTNTAGAKYGTGYCD SQCPRLKFFINGQANVEG WEPSSNNANTGIGHGSC CSEMDIWEANSISEALTPH PCTTVGQEICEGDGCGGT YSDNRYGGTCDPDGCDW NPYRLGNTSFYGPSSFTL DTTKKLTVVTOFETSGAI NRYYYQNGVTFQQPNAE LGSYSGNELNDDYCTAEE AEFGSSFSFKGGLTQFK KATSGGMVLVMSLWDD

	CCAGCTCAAGTTGAATCTCAATCTCCAAACGCTAAGTCACTTTCTCCAAACAT CAAGTTCGGTCCAAATCGTTCCACTGGTAATCCATCTGGTGGAACCCCTCCAG GTGTAACAGAGGTACTACCACTACTCGTAGGCCAGCTACTACAACTGGTTCT TCCCAGGCCCAACCAATCCCACTACCGGTCAATGTGGTGATCGGTTACTC TGTTCCAAACCGTCTGTGCTTCTGGTACTACCTGTCAAAGTTTAAACCCATACT <u>ACTCTCAATGTTTGTA (SEQ ID NO: 15)</u>	YYANMLWLDSTYPTNET SSTPGA VRGSCSTSSGVPA QVESQSPNAKVTFNSIKFG PIGSTGNPSGPNPPGGNR GTTTTRRPATTTGSSPGPT QSHYQCGGIGYSGPTVC ASGTTCCVLPNPPYSQCL (SEQ ID NO: 17) [Secretion signal : 1 – 33 catalytic domain : 41 – 465 cellulose-binding domain : 503 – 535]
<i>Trichoderma reesei cbh2</i>	ATGGTCTCCTTACCTCCCTGCTGGCCGGCGTTGCCGCTATCTCTGGTGCTCTA GCAGCCCTGCCGAGAGTTGAACCTGTGCGCAGTTGAGAAACGTGAGGCCG AAGCAGAAAGCTGCCATTAGAAAGAAAGACAAAGCCTGCTCTCTGTTGGG TCAATGTGGTGGTCAAACTGGTCTGGTCCAACTGTTGTGCTTCCGGTTCTA CCTGTGTTTACTCCACGACTACTATCCCAATGTTGCCAGGTGCTGCTTCTC CTTCTCTTCAACTAGAGTGTCTTCAAACTTCAAGTCTAGGCTCCCCAACCACTT CCAGATCCTCTTCTGCTACTCCACCACGAGTTCTACTACCACTAGAGTTCCA CCAGTCGGTTCGGTACTGCTACTTACTCTGTAACCTTTCGTCGGTGTACT CCATGGGCTAACGCTTACTACGCTTCTGAAGTTCTTCTTGGCTATCCCATCT TTGACTGGTGTATGGCTACCGCTGCTGCTGTGCGCAAGTTCCATCCTT CATGTGGTTGGACACCTTGGACAAACTCCATTAATGGAACAAACCTTGGCA GACATAAGGACTGTCTAACAAAGACGGCGGTAACTACGCTGGTCAATTTGTG TGTAAGACTTGCCAGACAGAGACTGTGCTGCTTTGGCTTCCAAACGGTGAATAC TCCATCGCTGACGGTGGTGTGCGCAAGTACAAAGAACTACATTGATACCATTAG ACAAATCGTTGTGCAATACTCTGACATCAGAACCTTGTAGTCATCGAACCCAG ATTCCTTAGCCAAATTTAGTCAACCACTTGGGTACTCCAAAGTGTGCTAACGCT CAATCTGCCCTACTTAGAATGATCAATATGCAATACCAATGAACTTGCC AAAGTTGCTATGTAATTTGACGCTGTGTCACGCGGTTGGTGGGTGGCCAG CTAACCAAGACCCAGCCGCTCAATTAATTCGCCAACGTTTACAAAGATGCTCT TCTCTAGAGCCTTGGTGGTGGTACTAACGTCGCTAACTACAAACGGTTG GAACATCACTTCTCCACCATCTTACCCCAAGGTAACGCTGTTTACAAACGAAA AGTTGTACATTACGCTATCGGTCCATTATTGGCTAACCATGGTGGTCTAAC GCCTTCTTCAATCACCGCAAGGTAGATCCGGTAAACCAACCACTGGTCAAC AACAAATGGGGTGATTGGTGAACGTCATCGGTACTGTTTCGGTATCAGACCA	Accession No.: AAA72922AAA34210 MIVGILTTLATLATLAASV PLEERQACSSVWQCGG QNWSGPTCCASGSTCVYS NDYYSQCLPGAASSST RAASTTSRVSPTTSRSSA TPPPGSTTTRVPPVSGSTA TYSGNPFVGVTPWANAY YASEVSSLAIPSLTGAMAT AAAAVAKVPSEMWLDTL DKTPLMEQTLADIRTK NGGNYAGQFVVYDLPDR DCAALASNGEYSLADGGV AKYKNYDITRQIVVEYSD IRTLVIEPDSLNLVTNL GTPKCANAQSAYLECIN AVTQLNLPNVAMYLDAG HAGWLGWPNQDPAQAQ LFANVYKNASSPRALRGL ATNVANYNGWNITSPSY TQGNVYNEKLYHAIGR LLANHGWSNAFFITDQGR

	TCCGCTAACACTGGTGATTCCCTTGTGATTCCCTTCGCTCTGGTTAAGCCAGG TGGTGAAATGTGATGGCACCTCTGATTCCCTCTGCTCCAAGATTTCGATTCCCACT GCGCCTTGCCAGACGCTTTGCAACCAAGCCCAAGCTGGTGCAATGGTTCCAA GCTTACTTTGTCCCAATTGTTGACCAACGCTAACCCCATCTTCTTGTA (SEQ ID NO: 16)	SGKQPTGQQQWGDWCN VIGTGFGRPSANTGDSLL DSFVWVKPGGECDDGTSDS SAPRFDSHCALPDALQPA AQAGAWFQAYFVQLLTN ANPSFL (SEQ ID NO: 18)
--	---	--

[0100] An amino acid sequence corresponding to optimized linker 1 according to the invention is a flexible linker - strep tag - TEV site - FLAG - flexible linker fusion and corresponds to GGGGSGGGGS AWHPQFGG ENLYFQG DYKDDDK GGGGSGGGGS (SEQ ID NO; 48).

[0101] The DNA sequence: is as follows:
GGAGGAGGTGGTTCAGGAGGTGGTGGGTCTGCTTGGCATCCACAATTTGGAG
GAGGCGGTGGTGAAAATCTGTATTTCCAGGGAGGCGGAGGTGATTACAAGGA
TGACGACAAAGGAGGTGGTGGATCAGGAGGTGGTGGCTCC (SEQ ID NO: 49).

[0102] An amino acid sequence corresponding to optimized linker 2 is a flexible linker - strep tag -linker- TEV site - flexible linker and corresponds to GGGGSGGGGS WSHPQFEK GG ENLYFQG GGGGSGGGGS (SEQ ID NO:50). The DNA sequence is as follows:
Ggtggcggtgatctggaggagggcgggtcttggctcaccacaaattgaaaagggtggagaaaacttgactttcaaggcgggt
ggaggaggttctggcgaggtggctccggctca (SEQ ID NO:51).

[0103] When using the methods above, the term "about" is used precisely to account for fractional percentages of codon frequencies for a given amino acid. As used herein, "about" is defined as one amino acid more or one amino acid less than the value given. The whole number value of amino acids is rounded up if the fractional frequency of usage is 0.50 or greater, and is rounded down if the fractional frequency of use is 0.49 or less. Using again the example of the frequency of usage of leucine in human genes for a hypothetical polypeptide having 62 leucine residues, the fractional frequency of codon usage would be calculated by multiplying 62 by the frequencies for the various codons. Thus, 7.28 percent of 62 equals 4.51 UUA codons, or "about 5," i.e., 4, 5, or 6 UUA codons, 12.66 percent of 62 equals 7.85 UUG codons or "about 8," i.e., 7, 8, or 9 UUG codons, 12.87 percent of 62 equals 7.98 CUU codons, or "about 8," i.e., 7, 8, or 9 CUU codons, 19.56 percent of 62 equals 12.13 CUC codons or "about 12," i.e., 11, 12, or 13 CUC codons, 7.00 percent of 62 equals 4.34 CUA codons or "about 4," i.e., 3, 4, or 5 CUA codons, and 40.62 percent of 62 equals 25.19 CUG codons, or "about 25," i.e., 24, 25, or 26 CUG codons.

[0104] 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 April 15, 2008) and the "backtranseq" function available at <http://bioinfo.pbi.nrc.ca:8090/EMBOSS/index.html> (visited July 9, 2002). 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.

[0105] 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.

[0106] 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.

[0107] 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.

[0108] The codon-optimized coding regions can be versions encoding a Cbh1, Cbh2, Egl, or Bgl1 from *T. emersonii*, *H. grisea*, *T. aurantiacus*, or *T. reesei*, or domains, fragments, variants, or derivatives thereof.

[0109] Codon optimization is carried out for a particular vertebrate species by methods described herein, for example, in certain embodiments codon-optimized coding regions encoding polypeptides of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 or Cbh2, or domains, fragments, variants, or derivatives 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 *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 or Cbh2, or domains, variants, or derivatives 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 polypeptides of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 or Cbh2, or domains, fragments, variants, or derivatives thereof, and various methods of using such polynucleotides, vectors and other expression constructs.

[0110] In certain embodiments described herein, a codon-optimized coding region encoding any of SEQ ID NOs:11-14 or 17-18, or domain, fragment, variant, or derivative thereof, is optimized according to codon usage in yeast (*Saccharomyces cerevisiae*). Alternatively, a codon-optimized coding region encoding any of SEQ ID NOs:11-14 or 17-18 may be optimized according to codon usage in any plant, animal, or microbial species.

Polypeptides of the Invention

[0111] The present invention further relates to the expression of tethered or secreted *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Eg1, Bgl1, Cbh1 or Cbh2 polypeptides in a host cell, such as *Saccharomyces cerevisiae*. The sequences of *T. reesei* Eg1 and Bgl1 are set forth above and the sequences of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1 and/or Cbh2 are set forth in the table below:

Organism and Protein	SEQ ID NO:
<i>H. grisea</i> Cbh1	11
<i>T. aurantiacus</i> Cbh1	12
<i>T. emersonii</i> Cbh1	13
<i>T. emersonii</i> Cbh2	14
<i>T. reesei</i> Cbh1	17
<i>T. reesei</i> Cbh2	18

[0112] The present invention further encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is at least about 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% identical to, for example, any of the polypeptide sequences of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4, and/or domains, fragments, variants, or derivative thereof, of any of these polypeptides (*e.g.*, those fragments described herein, or domains of any of the polypeptide sequences of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4.

[0113] 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 may 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 may be inserted, deleted, (indels) or substituted with another amino acid. These alterations of the reference sequence may 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.

[0114] As a practical matter, whether any particular polypeptide is at least 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% identical to, for instance, the amino acid sequence of SEQ ID NO:3 or to the amino acid sequence encoded by the deposited clone 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 may be made to the results in certain instances.

[0115] In certain embodiments, the polypeptide of the present invention encompasses a fusion protein comprising a first polypeptide, where the first polypeptide is a *T. emersonii* Cbh1, *H. grisea* Cbh1, or *T. aurantiacusi* Cbh1, *T. emersonii* Cbh2, *T. reesei* Cbh1, *T. reesei* Cbh2, or domain, fragment, variant, or derivative thereof, and a second polypeptide, where the second polypeptide is a *T. emersonii* Cbh1, *H. grisea* Cbh1, or *T. aurantiacusi* Cbh1, *T. emersonii* Cbh2, *T. reesei* Cbh1 or *T. reesei* Cbh2, or domain, fragment, variant, or derivative thereof. In particular embodiments the first polypeptide is *T. emersonii* Cbh1 and the second polynucleotide is a CBM from *T. reesei* Cbh1 or Cbh2. In further embodiments of the fusion protein, the first and second polypeptides 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*. In particular embodiments, the first polynucleotide is a codon-optimized *T. emersonii cbh1* and the second polynucleotide encodes for a codon-optimized CBM from *T. reesei* Cbh1 or Cbh2. In certain other embodiments, the first polypeptide and the second polypeptide are fused via a linker sequence.

[0116] 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.

[0117] The present invention also encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is at least about 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% similar to the polypeptide of any of SEQ ID NOs: 11-14 or 17-18, and to portions of such polypeptide with such portion of the polypeptide generally containing at least 30 amino acids and more preferably at least 50 amino acids.

[0118] 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.

- [0119] The present invention further relates to a domain, fragment, variant, derivative, or analog of the polypeptide of any of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4.
- [0120] Fragments or portions of the polypeptides of the present invention may be employed for producing the corresponding full-length polypeptide by peptide synthesis, therefore, the fragments may be employed as intermediates for producing the full-length polypeptides.
- [0121] Fragments of Cbh, Eg1, or Bgl1 polypeptides of the present invention encompass domains, proteolytic fragments, deletion fragments and in particular, fragments of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 polypeptides which retain any specific biological activity of the Cbh1, Cbh2, Eg1 or Bgl1 proteins. Polypeptide fragments further include any portion of the polypeptide which comprises a catalytic activity of the Cbh1, Cbh2, Eg1 or Bgl1 proteins.
- [0122] The variant, derivative or analog of the polypeptide of any of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4, or that encoded by the deposited clone, may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the mature polypeptide for purification of the polypeptide 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.
- [0123] 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.

[0124] 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 may be produced using art-known mutagenesis techniques. Allelic variants, though possessing a slightly different amino acid sequence than those recited above, will still have the same or similar biological functions associated with the *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1, or Bgl1 protein.

[0125] The allelic variants, the conservative substitution variants, and members of the endoglucanase, cellobiohydrolase or β -glucosidase protein families, will have an amino acid sequence having at least 75%, at least 80%, at least 90%, at least 95% amino acid sequence identity with a *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 amino acid sequence set forth in any one of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4. 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.

[0126] Thus, the proteins and peptides of the present invention include molecules comprising the amino acid sequence of SEQ ID NOs: 11-14 or 17-18 or of Tables 3 or 4 or fragments thereof having a consecutive sequence of at least about 3, 4, 5, 6, 10, 15, 20, 25, 30, 35 or more amino acid residues of the *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1, or Bgl1 polypeptide sequences; 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 animal species, including but not limited to bacterial, fungal, insect, rabbit, rat, porcine, bovine, ovine, equine and non-human primate species, 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).

[0127] Using known methods of protein engineering and recombinant DNA technology, variants may be generated to improve or alter the characteristics of the CBH 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.

[0128] Thus, the invention further includes *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 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.

[0129] The skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly effect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid), as further described below.

[0130] For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science* 247:1306-1310 (1990), wherein the authors indicate that there are two main strategies for studying the tolerance of an amino acid sequence to change.

[0131] The first strategy exploits the tolerance of amino acid substitutions by natural selection during the process of evolution. By comparing amino acid sequences in different species, conserved amino acids can be identified. These conserved amino acids are likely important for protein function. In contrast, the amino acid positions where substitutions have been tolerated by natural selection indicates that these positions are not critical for protein function. Thus, positions tolerating amino acid substitution could be modified while still maintaining biological activity of the protein.

[0132] The second strategy uses genetic engineering to introduce amino acid changes at specific positions of a cloned gene to identify regions critical for protein function. For example, site directed mutagenesis or alanine-scanning mutagenesis (introduction of

single alanine mutations at every residue in the molecule) can be used. (Cunningham and Wells, *Science* 244:1081-1085 (1989).) The resulting mutant molecules can then be tested for biological activity.

[0133] 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.

[0134] The terms "derivative" and "analog" refer to a polypeptide differing from the *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 polypeptide, but retaining essential properties thereof. Generally, derivatives and analogs are overall closely similar, and, in many regions, identical to the *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 polypeptides. The term "derivative" and "analog" when referring to *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 polypeptides of the present invention include any polypeptides which retain at least some of the activity of the corresponding native polypeptide, *e.g.*, the exoglucanase activity, or the activity of the its catalytic domain.

[0135] Derivatives of *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 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.

[0136] An analog is another form of a *T. emersonii*, *H. grisea*, *T. aurantiacus* or *T. reesei* Cbh1, Cbh2, Eg1 or Bgl1 polypeptide of the present invention. An "analog" also retains

substantially the same biological function or activity as the polypeptide of interest, *e.g.*, functions as a cellobiohydrolase. An analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

[0137] The polypeptide of the present invention may be a recombinant polypeptide, a natural polypeptide or a synthetic polypeptide, preferably a recombinant polypeptide.

Vectors and Host Cells

[0138] 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.

[0139] Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

[0140] The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences, *e.g.*, derivatives of SV40; bacterial plasmids; and yeast plasmids. However, any other vector may be used as long as it is replicable and viable in the host.

[0141] The appropriate DNA sequence may 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.

[0142] 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:

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

[0143] Additionally, promoter sequences from stress and starvation response genes are useful in the present invention. In some embodiments, promoter regions from the *S. cerevisiae* genes *GAC1*, *GET3*, *GLC7*, *GSH1*, *GSH2*, *HSF1*, *HSP12*, *LCB5*, *LRE1*, *LSP1*, *NBP2*, *PIL1*, *PIM1*, *SGT2*, *SLG1*, *WHI2*, *WSC2*, *WSC3*, *WSC4*, *YAP1*, *YDC1*, *HSP104*, *HSP26*, *ENA1*, *MSN2*, *MSN4*, *SIP2*, *SIP4*, *SIP5*, *DPL1*, *IRS4*, *KOG1*, *PEP4*, *HAP4*, *PRB1*, *TAX4*, *ZPR1*, *ATG1*, *ATG2*, *ATG10*, *ATG11*, *ATG12*, *ATG13*, *ATG14*, *ATG15*, *ATG16*, *ATG17*, *ATG18*, and *ATG19* may be used. Any suitable promoter to drive gene expression in the host cells of the invention may be used.

[0144] Additionally, the *E. coli*, lac or trp, and other promoters known to control expression of genes in prokaryotic or lower eukaryotic cells. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression, or may include additional regulatory regions.

[0145] In addition, the expression vectors may contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as *URA3*, *HIS3*, *LEU2*, *TRP1*, *LYS2* or *ADE2*, dihydrofolate reductase, neomycin (G418) resistance or zeocin resistance for eukaryotic cell culture, or tetracycline or ampicillin resistance in *E. coli*.

[0146] The vector containing the appropriate DNA sequence as herein, as well as an appropriate promoter or control sequence, may be employed to transform an appropriate host to permit the host to express the protein.

[0147] 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.

[0148] As representative examples of appropriate hosts, there may be mentioned: 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.

[0149] Appropriate fungal hosts include yeast. In certain aspects of the invention the yeast is *Saccharomyces cerevisiae*, *Kluyveromyces lactus*, *Schizosaccharomyces pombe*, *Candida albicans*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Debaryomyces polymorphus*, *Schwanniomyces occidentalis*, *Issatchenkia orientalis*, and *Kluyveromyces marxianus*.

[0150] In particular embodiments, the vector of the present invention is a plasmid of Table 5 below. Table 6 lists primer sequences utilized to construct various plasmids of the invention.

Table 5. Plasmids in this study.

*P/T = Promoter/Terminator; **BGL1 = β -glucosidase 1 from *Saccharomycopsis fibuligera*, Van Rooyen *et al.* (2005)

#	Name of Plasmid	Used for/Genes carried	Reference/ accession #
1	pBluescript II SK+	Expression vector backbone for assembling expression cassettes	X52328
2	pTEF1-zeo	TEF1/Zeo marker	Invitrogen
3	M4297	KanMX marker	Prof. David Stillman
3a	ySFI	BGLI	Van Rooyen (2005)
4	pBK	pBluescript; KanMX marker	This study
5	pBZ	pBluescript; TEF1/Zeo marker	This study
6	pBK_1	pBK + PGK P/T*	This study
7	pBK_2	pBK + ENO1 P/T*	This study
8	pBZ_1	pBZ + PGK P/T*	This study
9	pBZ_2	pBZ + ENO1 P/T*	This study
10	pBKD1_1	pBK_1 + 1 δ sequence	This study
11	pBKD1_2	pBK_2 + 1 δ sequence	This study
12	pBZD1_1	pBZ_1 + 1 δ sequence	This study
13	pBZD1_2	pBZ_2 + 1 δ sequence	This study
14	pBKD_1	pBK_1 + 2 δ sequences	This study
15	pBZD_1	pBK_2 + 2 δ sequences	This study
16	pBKD_2	pBZ_1 + 2 δ sequences	This study
17	pBZD_2	pBZ_2 + 2 δ sequences	This study

18	pBKD_10001	pBKD_1 + L1_A1 (original optimization)	This study
19	pBKD_20001	pBKD_2 + L1_A1 (original optimization)	This study
20	pBZD_20001a	pBZD_2 + L2_A1a (re-optimized)	This study
21	pBKD_20511	pBKD_20001 + BGL1	This study
22	pBKD_11621	pBKD_10001 + S16 + C2	This study
23	pBKD_10621	pBKD_10001 + S06 + C2	This study
24	pBKD_10621_20511	pBKD_10621 + 20511 (i.e. only the cellulase construct)	This study
25	pBKD_11621_20511	pBKD_11621 + 20511 (i.e. only the cellulase construct)	This study
26	pBZD_11631	pBZD_1 + S16 + C3_L2_A1	This study
27	pBZD_20641	pBZD_20001a + C4_L3	This study
28	pBZD_11631_20641	pBZD_11631 + 20641 (i.e. only the cellulase construct)	This study
29	pBZD_10511	pBZD_1 + 0511 (BGLI with xyn sec)	This study
30	pBKD non-anchored EGI	pBKD_10001 with EGI PCR product	This study
31	pBKD non-anchored CBHI	pBKD_10001 with CBHI PCR product	This study
32	pBKD Flo1_EGI	pBKD with Flo1 anchor and EGI	This Study
33	pBKD Flo1_CBHI	pBKD with flo1 anchor and CBHI	This Study
34	pBKD_Flo1_EGI_20511	N-terminally anchored EGI with BGLI for co-expression	This Study
35	pBZD_Flo1_CBHI_20641	N-terminally anchored CBHI with CBH2 for co-expression	This Study

Table 6. Primers used for constructs.

#	Name	Sequence 5' → 3'
1	ENO1f	TATATGGGCCCCACTAGTCTTCTAGGCGGGTTATCTACTGATCC (SEQ ID NO: 52)
2	ENO1overlap	GGACTAGAAGGCTTAATCAAAAGCGGCGCGCCGGATCCTTAATTAAGT GTGTTGATAAGCAGTTGCTTGGTT (SEQ ID NO: 53)
3	ENO1r	GCTACGAATTCGCGGCCGCCGTCGAACAACGTTCTATTAGGA (SEQ ID NO: 54)
4	PGKf	TATATGGGCCCTCCCTCCTTCTTGAATTGATGT (SEQ ID NO: 55)
5	PGKoverlap	GATCTATCGATTTC AATTCAATTCAATGGCGCGCCGGATCCTTAATTAA TGTA AAAAGTAGATAATTACTTCTTGATG (SEQ ID NO: 56)
6	PGKr	CTTAGGAATTCTTTCGAAACGCAGAATTTTC (SEQ ID NO: 57)

7	kanMXf	GATCCGAATTCGTTTAGCTTGCCTCGTCCC (SEQ ID NO: 58)
8	kanMXr	CAGTCGACTAGTTTTTCGACACTGGATGGCG (SEQ ID NO: 59)
9	Zeof	GCGCTAGAATTCCCCACACACCATAGCTTCAA (SEQ ID NO: 60)
10	Zeor	CCGCATACTAGTAATTCAGCTTGCAAATTAAAGCCTTCGAG (SEQ ID NO: 61)
11	BGL1f	GCCGCCTTAATTAAAAACAAAATGGTCTCCTTCACCTCCCT (SEQ ID NO: 62)
12	BGL1r	CGGTTGGATCCAATAGTAAACAGGACAGATGTCTTGAT (SEQ ID NO: 63)
13	Delta2f	AGTCGCGGCCGCTGTTGGAATAAAAATCCACTATCGT (SEQ ID NO: 64)
14	Delta2r	GCGCCCGCGGTGAGATATATGTGGGTAATTAGATAATTGT (SEQ ID NO: 65)
15	M13f	TCCAGTCACGACGTCGT (SEQ ID NO: 66)
16	M13r	GGAAACAGCTATGACCATG (SEQ ID NO: 67)
17	PGKseqf	TCTTTTCTCTTTTACAGATCATCA (SEQ ID NO: 68)
18	ENO1seqf	TCCTTCTAGCTATTTTCATAAAAAAC (SEQ ID NO: 69)
19	EGI_detect_F	CGCTAGTGGTGTACGACGA (SEQ ID NO: 70)
20	EGI_detect_R	CTCCAAGTCTGCACTGGACA (SEQ ID NO: 71)
21	BGL1_detect_F	GAGCCCGCATTATTATCCAA (SEQ ID NO: 72)
22	BGL1_detect_R	CAAAGTCAGCGAATCGAACA (SEQ ID NO: 73)
23	CBHI_detect_F	AGACGGTTGTGACTGGAACC (SEQ ID NO: 74)
24	CBHI_detect_R	CAACTTGAGCTGGAACACCA (SEQ ID NO: 75)
25	CBHII_detect_F	CAGAGACTGTGCTGCTTTGG (SEQ ID NO: 76)
26	CBHII_detect_R	GGATCTACCTTGGTCGGTGA (SEQ ID NO: 77)
27	EGI N term F	ACTGGGCTCAGCTCAACAACCAGGAACATCAACAC (SEQ ID NO: 78)
28	EGI N term R	AGCTGGCGCGCCTTATAAACATTGTGAGTAATAGTCATTA CTGT (SEQ ID NO: 79)
29	CBHI N term F	ACTGGGCTCAGCTCAATCCGCTTGTACCCTACAA (SEQ ID NO: 80)
30	CBHI N term R	AGCTGGCGCGCCTTACAAACATTGAGAGTAGTATGGGTTT AA (SEQ ID NO: 81)
31	His3 F	TATTGTGAGGGTCAGTTATT (SEQ ID NO: 82)
32	His3 R	TAAAAGGAGCCTTGAGACTC (SEQ ID NO: 83)
33	Trp1 F	TACTATTAGCTGAATTGCCA (SEQ ID NO: 84)
34	Trp1 R	GGAACGTTTGTATTCATACT (SEQ ID NO: 85)
35	Leu2 F	ACATCGAGACCAAGAAGAAC (SEQ ID NO: 86)
36	Leu2 R	CGAGATTGATGAAGAAAGAA (SEQ ID NO: 87)
37	Ura3 F	AGCTTTTCAATTCAATTCAT (SEQ ID NO: 88)

38	Ura3 R	CCGGGTAATAACTGATATAA (SEQ ID NO: 89)
----	--------	--------------------------------------

[0151] Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

[0152] The polynucleotides of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, non-chromosomal and synthetic DNA sequences, *e.g.*, derivatives of SV40; bacterial plasmids; and yeast plasmids. However, any other vector may be used as long as it is replicable and viable in the host.

[0153] The appropriate DNA sequence may 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.

[0154] The DNA sequence in the expression vector is operatively associated with an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. As representative examples of such promoters, there may be mentioned: *ENO1*, *PGK1*, *TEF1*, *GPD1*, *ADH1* and the *E. coli*, *lac* or *trp*, and other promoters known to control expression of genes in prokaryotic or lower eukaryotic cells. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression, or may include additional regulatory regions.

[0155] "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.

- [0156] "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 may 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 may 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 may 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.
- [0157] 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.
- [0158] "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.
- [0159] 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.

[0160] 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 may also refer to translation of mRNA into a polypeptide.

Selection Methods

[0161] As used herein, "selection methods" or "selection protocol(s)" refers to methods for putting pressure on (or challenging) a given strain to adapt to new conditions. The selection methods favor sporadic "variants" of the original strain wherein the variants undergo some genetic or epigenetic change that confers a reproductive and thereby growth advantage in the culture conditions of the embodiment. Using the methods of the present invention, it is thereby possible to apply continuous selective pressure on strains of the invention, causing the variant strains with genetic or epigenetic changes that confer a reproductive and growth advantage, to eventually dominate the culture. Thereby one can continue to improve the performance of the organism with respect to its ability to grow in certain conditions, for example on cellulosic material.

[0162] In some embodiments, a small quantity of a favored carbon source (such as glucose) may be added to the selection media to allow for a slight increase in growth rate of the cells. Adding small quantities of a favored carbon source enables the cells undergoing selection to reproduce more rapidly, allowing for more generations of cells per unit time. This in turn allows for more opportunities to undergo a genetic or epigenetic change that confers a reproductive, and thereby growth advantage, in the culture condition. Additionally, small quantities of glucose or other sugars might be useful to drive gene expression if polynucleotides of the present invention are operably linked to promoters influenced by carbon source.

[0163] Favored carbon sources can differ by host cell, but are generally well known to a person of ordinary skill in the art. Favored carbon sources generally are mono and di saccharides such as glucose, galactose, maltose, fructose, as well as soluble or insoluble oligmers of glucuse. For example, cellulose chains from 3 up to to 30 or 40 glucose units in length would provide high reactivity, but still require some cellulase activity.

[0164] In some embodiments, the selection methods are carried out using a semi-continuous culture. In some embodiments, the semi-continuous culture comprises: (a) a residence chamber, wherein host cells of the invention are grown; (b) a fresh media

chamber, in controlled, fluid communication with the residence chamber; and, (c) a waste chamber, in controlled fluid communication with the residence chamber. In some embodiments, the fresh media from the media chamber is pumped into the residence chamber, and at the same or similar rate, the spent media is pumped from the residence chamber into the waste chamber. In these embodiments, culture conditions are kept largely constant or in a "substantially steady state," meaning the media and culture conditions are stable. Optionally, in some embodiments, a fourth chamber is used to separately regulate levels of a media ingredient separately, for example glucose. Thereby the levels of the separate media ingredient can be altered while keeping the levels of other media components constant. In some embodiments, transport of fluids between the residence chamber and the other chambers may be accomplished, for example, by a peristaltic, or other suitable pump.

[0165] Under conditions described above, cell numbers in the residence chamber remain constant if the rate of cell division equals the wash out rate of cells from the residence chamber into the waste chamber. However, if cell division is faster than the cell washout rate, cell numbers in the residence chamber increase. Conversely, if cell division in the residence chamber is slower than the washout rate, the cell numbers in the residence chamber decrease. Therefore, by modulating the washout rate and media conditions, methods of the present invention allow for the selection of cells with ever increasing ability to grow and divide in various culture conditions.

[0166] In some embodiments of the present invention, the selection methods produce variant cells that are able to grow to cell densities of at least about 1.2, at least about 1.5, at least about 2, at least about 4, at least about 8, at least about 10, or at least about 50 fold greater than the pre-selected parental strain in culture conditions of the invention. In other embodiments, the selected cells are able to grow at least about 1.2, at least about 1.5, at least about 2, at least about 4, at least about 8, at least about 10, or at least about 50 times faster than the pre-selected parental strain in the culture conditions of the invention. In still other embodiments, the selection methods produce variant cells that are able to ferment cellulosic material to produce ethanol at least about 1.2, at least about 1.5, at least about 2, at least about 4, at least about 8, at least about 10, or at least about 50 fold in excess of the pre-selected parental strain in the culture conditions of the invention.

- [0167] Methods of measuring cell density are well known in the art and include optical density measurements of cell cultures or direct counting of cells by hemocytometer. Monitoring cultures over a period of time by one of these measurements will enable a person of ordinary skill in the art to calculate growth rate of the cells of the invention.
- [0168] Although the results reported herein are for *Saccharomyces cerevisiae*, the methods and materials also apply to other types of yeast including, for example, *Schizosaccharomyces pombe*, *Candida albicans*, *Kluyveromyces lactis*, *Pichia pastoris*, *Pichia stipitis*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Debaryomyces hansenii*, *Issatchenkia orientalis*, *Debaryomyces polymorphus* and *Schwanniomyces occidentalis*.
- [0169] The yeast may be selected, for example, from the genera *Saccharomyces*, *Kluyveromyces*, *Candida*, *Pichia*, *Schizosaccharomyces*, *Hansenula*, *Kloeckera*, *Schwanniomyces*, and *Yarrowia*. Yeast species as host cells may include, for example, *S. cerevisiae*, *S. bulderi*, *S. barnetti*, *S. exiguus*, *S. uvarum*, *S. diastaticus*, *K. lactis*, *K. marxianus*, and *K. fragilis*.
- [0170] The disclosed recombinant yeast strains have the potential to contribute significant savings in the lignocellulosic biomass to ethanol conversion. For example, the disclosed recombinant yeast strains may be suitable for a consolidated bioprocessing co-culture fermentation where they would convert cellulose to ethanol, and hemicellulose would be degraded by a pentose-utilizing organism, such as *Saccharomyces cerevisiae* RWB218, disclosed by Kuyper, M. *et al.*, "Metabolic engineering of a xylose-isomerase-expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation", *FEMS Yeast Research*, 5: 399-409, (2005).
- [0171] Substrates for cellulose activity assays can be divided into two categories, soluble and insoluble, based on their solubility in water. Soluble substrates include cellodextrins or derivatives, carboxymethyl cellulose (CMC), or hydroxyethyl cellulose (HEC). Insoluble substrates include crystalline cellulose, microcrystalline cellulose (AvicelTM), amorphous cellulose, such as phosphoric acid swollen cellulose (PASC), dyed or fluorescent cellulose, and pretreated lignocellulosic biomass. These substrates are generally highly ordered cellulosic material and thus only sparingly soluble.
- [0172] It will be appreciated that suitable lignocellulosic material may be any feedstock that contains soluble and/or insoluble cellulose, where the insoluble cellulose may be in a

crystalline or non-crystalline form. In various embodiments, the lignocellulosic biomass comprises, for example, wood, corn, corn stover, sawdust, bark, leaves, agricultural and forestry residues, grasses such as switchgrass, ruminant digestion products, municipal wastes, paper mill effluent, newspaper, cardboard or combinations thereof.

[0173] The following embodiments of the invention will now be described in more detail by way of these non-limiting examples.

Examples

Example 1: Expression of tethered and secreted cellulases by yeast

[0174] In order to combine the possible benefits of secreted and tethered cellulases systems, *S. cerevisiae* strains containing both types of constructs were created.

Materials and Methods

Strain Construction

[0175] Molecular and transformation methods were used as described in the previous examples. M0144 was transformed with pRDH105 as well as a PCR product for the His3 gene. Transformants were selected on YNB media + glucose with no amino acids present. This ensured that the 2 micron plasmid for TeCBHI expression, which was selectable by the URA3 gene would be present in this strain background. The newly created strain was called M0360.

Results

[0176] Growth results for this strain can be found in later examples.

Example 2: Growth of recombinant, cellulolytic yeast strains on cellulosic substrates

[0177] CBP conversion of cellulose to ethanol requires a biocatalyst to grow on cellulosic material. This growth allows the catalyst to propagate from an initial small inoculum and consume the cellulosic substrate producing ethanol simultaneously. This example demonstrates the ability of a number of recombinant yeast strains to grow on a variety of types of insoluble cellulosic substrates, and demonstrates their cellulolytic capability.

Materials and methods

Cellulosic Substrates

- [0178] Bacterial microcrystalline cellulose (BMCC) was a gift from CP Kelco company. BMCC as received was stirred O/N at 4C in water. After the substrate was rehydrated, it was washed 6 times with water and resuspended in water. The dry weight of the substrate was measured by drying samples at 105C until constant weight was obtained.
- [0179] AvicelTM PH105 (FMC Biopolymers) was used as provided by the manufacturer.
- [0180] Pretreated mixed hardwoods were generated by autohydrolysis of the substrate at 160 PSI for 10 minutes. Pretreated material was washed 5 times to remove inhibitors and soluble sugars and resuspended in distilled water. Samples were dried overnight at 105C to determine the dry weight. Analysis of sugar content by quantitative saccharification showed a 50% glucan content.

Growth media and cultivation conditions

- [0181] Growth media with cellulose substrates as the sole carbon source were made using the non-glucose components of synthetic complete medium for yeast including, yeast nitrogen base without amino acids (YNB)—6.7 g/L, and in some cases supplementing with amino acids. In some cases, Yeast Extract (10 g/L) and Peptone (20 g/L) were used instead of YNB for the non-carbon components of the media. Cultivation conditions included aerobic and microaerobic conditions. Aerobic conditions were maintained by using 250 mL shake flasks with AvicelTM containing media. Microaerobic conditions were maintained by growing strains on BMCC in sealed hungate tubes with an air atmosphere.
- [0182] Washout experiments using semi-continuous culture of *Saccharomyces cerevisiae* strains were carried out in 3L (total volume) Sartorius bioreactors. AvicelTM (~20 g/L; PH105 from FMC Biopolymer) was added to synthetic complete medium for yeast (Yeast nitrogen base without amino acids 6.7 g/L) lacking a carbon source. AvicelTM containing media was stirred in a 20L carboy and intermittently pumped into reactors with working volumes of ~900 mL. Media was pumped out in an intermittent fashion. Conditions in the reactors were maintained at pH ~5.8 by addition of new media (growth was not enough to change the pH of the media), stirring at 400 rpm, an aeration rate of 200 mL/min, and a temperature of 35 C. The dilution rate was maintained at ~0.02 hr⁻¹,

which was verified by measuring the volume of the media accumulated in a waste carboy. Cells were quantified by direct counts via haemocytometer.

Strains for cellulose conversion

[0183] Strains expressing tethered and secreted cellulase enzymes described in the previous three examples were used for the conversion experiments. Pre-cultures were grown for 1-2 days, and cells were inoculated into cellulose containing media.

Results

[0184] Figures 1 and 2 show a variety of growth demonstrations for the strains created. Strains were also created where all auxotrophies were corrected, allowing the strains to be grown on media not containing any supplemental amino acids—i.e. Yeast Nitrogen Base without amino acids (YNB, DIFCO) as the non-carbon components. In these cases the carbon available to the yeast strains that is not added as a carbohydrate source is ~2 mg/L. Therefore, cell growth can be attributed entirely to cellulose utilization. Figure 1 shows the results of an aerobic growth experiment in shake flasks using YNB + 2% Avicel™ PH105, where a 20% inoculum was used. As can be seen, strains expressing tethered, secreted, and a combination of tethered and secreted cellulases can grow on Avicel™.

[0185] Figure 2 shows the results of a different type of growth test using prototrophic strains of *S. cerevisiae* expressing either all secreted cellulases or a combination of secreted and tethered cellulases. In this case a continuous culture with Avicel™ PH105 was run at a dilution rate of 0.02 hr⁻¹. This culture condition eliminates the effect of inoculation size in determining ability to grow on cellulose. The cells wash out of the culture, but not as quickly as would be predicted by the dilution effect of adding new media. The difference between these measurements and the predicted washout rate can be used to measure the maximum specific growth rate on the carbon source (Wang, P. *et al.*, "Kinetic analyses of desulfurization of dibenzothiophene by *Rhodococcus erythropolis* in continuous culture," *Appl. Env. Micro.* 62: 3066-3068 (1996). This technique was used to measure a maximum specific growth rate of ~0.01 hr⁻¹.

Example 3: Recombinant Yeast strains and yeast strain co-cultures fermenting cellulose to ethanol (PASC, BMCC, Avicel™) without added cellulase

- [0186] A CBP process requires strains capable of producing ethanol with reduced cellulase loading, and in the ultimate configuration, with no exogenously added cellulase. This example demonstrates the ability of the recombinant yeast strains to produce ethanol directly from cellulose without exogenously added cellulase enzymes.

Materials and Methods

Cellulosic Substrates

- [0187] Phosphoric acid swollen cellulose (PASC) was prepared as in Zhang and Lynd, "Determination of the number average degree of polymerization of cellodextrins and cellulose with application to enzymatic hydrolysis," *Biomacromolecules* 6:1510-1515. (2005), with only slight modifications. Avicel™ PH105 (10 g) was wetted with 100 mL of distilled water in a 4L flask. 800 mL of 86.2% phosphoric acid was added slowly to the flask with a first addition of 300mL followed by mixing and subsequent additions of 50mL aliquots. The transparent solution was kept at 4°C for 1 hour to allow complete solubilization of the cellulose, until no lumps remained in the reaction mixture. Next, 2L of ice-cooled distilled water was added in 500mL aliquots with mixing between additions. 300 mL aliquots of the mixture were centrifuged at 5,000 rpm for 20 minutes at 2°C and the supernatant removed. Addition of 300 mL cold distilled water and subsequent centrifugation was repeated 4X. 4.2 mL of 2M sodium carbonate and 300 mL of water were added to the cellulose, followed by 2 or 3 washes with distilled water, until the final pH was ~6. Samples were dried to constant weight in a 105°C oven to measure the dry weight.
- [0188] Bacterial microcrystalline cellulose (BMCC) was a gift from CP Kelco company. BMCC as received was stirred O/N at 4C in water. After the substrate was rehydrated, it was washed 6 times with water and resuspended in water. The dry weight of the substrate was measured by drying samples at 105C until constant weight was obtained.
- [0189] Avicel™ PH105 (FMC Biopolymers) was used as provided by the manufacturer.
- [0190] Pretreated mixed hardwoods were generated by autohydrolysis of the substrate at 160 PSI for 10 minutes. Pretreated material was washed 5 times to remove inhibitors and soluble sugars and resuspended in distilled water. Samples were dried overnight at 105C

to determine the dry weight. Analysis of sugar content by quantitative saccharification showed a 50% glucan content.

Growth media and cultivation conditions

[0191] Growth media with cellulose substrates as the sole carbon source were made using the non-glucose components of synthetic complete medium for yeast including, yeast nitrogen base without amino acids—6.7 g/L, and supplemented with a complete amino acid mix (complete supplemental mixture). In some cases yeast extract (10 g/L) and peptone (20 g/L) (YP) were used as supplements in growth experiments. Cultivation conditions were anaerobic and were maintained by flushing sealed glass bottles with N₂ after carbon source addition and before autoclaving. Non-carbon media components were added as 10X solutions by filter sterilizing after autoclaving. Inoculation into PASC and BMCC cultures was done at 10% by volume, whereas inoculation into Avicel™ cultures was done at 20% by volume.

Ethanol quantification

[0192] Quantification of ethanol in fermentation samples was carried out by HPLC analysis, and initial ethanol concentrations in bottles (from precultures) was subtracted from all subsequent data points.

Results

[0193] Results from anaerobic fermentation experiments using PASC or BMCC as substrates for ethanol production are shown in Figures 3, 4, and 5. Figures 3 and 4 show ethanol production from PASC with YNB and amino acids added (except for M0360, where no amino acids were added) as media components. In both experiments, M0360 performs the best, and M0284 is the second best strain. However, M0360 is the only strain to show ethanol accumulation above the starting concentration during cultivation on BMCC. Figure 5 shows the results when YP is used as the media source for fermentation of PASC by these strains. In this case, strain M0284 performs the best, while M0244 and M0286 perform slightly better than the controls. M0360 does not have selective pressure to retain the *T. emersonii* CBH1 in YP media, and that is the likely reason that it did not show the same performance as on YNB.

Example 4: Construction of protrophic yeast strains

Materials and Methods

- [0194] **Molecular methods, strains and plasmids.** Standard protocols were followed for DNA manipulations (Sambrook, J., E. F. Fritsch, and T. Maniatis. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y (1989)). PCR was performed using Advantage Polymerase (Clontech) for PCR of genes for correcting auxotrophies. Manufacturers guidelines were followed as supplied.
- [0195] The 2 μ plasmid with URA3 marker for secreted *Talaromyces emersonii* CBH1 expression was obtained from the University of Stellenbosch, and is named pRDH105, and was built from YEpENO-BBH. The yeast expression vector YEpENO-BBH was created to facilitate heterologous expression under control of the *S. cerevisiae* enolase 1 (*ENO1*) gene promoter and terminator and to ease combination of gene cassettes as the expression cassette from this vector could be excised with a *Bam*HI, *Bgl*II digest. YEpENO1 (Den Haan R, *et al.*, "Hydrolysis and fermentation of amorphous cellulose by recombinant *saccharomyces cerevisiae*," *Metab Eng* 9:87-94 (2007).) contains the YEp352 backbone with the *ENO1* gene promoter and terminator sequences cloned into the *Bam*HI and *Hind*III sites. This plasmid was digested with *Bam*HI and the overhang filled in with Klenow polymerase and dNTPs to remove the *Bam*HI site. The plasmid was re-ligated to generate YEpENO-B. Using the same method, the *Bgl*II and then the *Hind*III sites were subsequently destroyed to create YEpENO-BBHtemplate. YEpENO-BBHtemplate was used as template for a PCR reaction with primers ENOBB-left (5'-GATCGGATCCCAATTAATGTGAGTTACCTCA-3' (SEQ ID NO: 90)) and ENOBB-right (5'-GTACAAGCTTAGATCTCCTATGCGGTGTGAAATA-3' (SEQ ID NO: 91)) in which the *ENO1* cassette was amplified together with a 150 bp flanking region upstream and 220bp downstream. This product was digested with *Bam*HI and *Hind*III and the overhangs filled in by treatment with Klenow polymerase and dNTPs and cloned between the two *Pvu*II sites on yENO1 effectively replacing the original ENO1 cassette and generating YEpENO-BBH.
- [0196] *Talaromyces emersonii cbh1* was designed and a synthetic gene ordered from GenScript Corporation (Piscataway, NJ, USA)—Table 7 contains the codon optimized sequence. The synthetic chb gene was designed for optimal expression in *S. cerevisiae* using --"synthetic gene designer"(<http://phenotype.biosci.umbc.edu/codon/sgd/index.php>)

The synthetic *cbh* encoding gene received from GenScript Corporation was cloned in to the plasmid pUC57, subsequently digested with *EcoRI* and *XhoI* to excise the *cbh* gene, and finally cloned into a *EcoRI* and *XhoI* digested YEpENO-BBH. This created the plasmids pRDH105, with *Tecbh1* placed under transcriptional control of the *ENO1* promoter and terminator.

[0197] **PCR products for creating prototrophic yeast strains.** When yeast strains without auxotrophies was desired, PCR reactions using primers HIS F and HIS R or URA F and URA R (Table 6) were used to carry out reactions as appropriate. Genomic DNA purified from prototrophic industrial yeast strain, D5A was used as a template. PCR products were gel-purified and used in yeast transformations as described below.

Table 7. Amino acid and DNA sequence for *T. emersonii cbh1*.

Gene name	Amino Acid Sequence	DNA sequence
<i>Talaromyces emersonii cbh1</i>	MLRRALLSSSAIL AVKAQQAGTATA ENHPPLTWQECTA PGSCTTQNGAVVL DANWRWVHDVN GYTNCYTGNTWD PTYCPDDETC AQN CALDGADYEGTY GVTSSGSSLKLN VTGSNVGSRLYLL QDDSTYQIFKLLN REFSFDVDVSNLP CGLNGALYFVAM DADGVSKYPNN KAGAKYGTGYCD SQCPRLKFIDGE ANVEGWQPSSNN ANTGIGDHGSCCA EMDVWEANSISN AVTPHPCDTPGQT MCSGDDCGGTYS NDRYAGTCDPDG CDFNPYRMGNTSF YGPBKIIDTTKPFT VVTQFLTDDGTD GTLSEIKRFYIQNS NVIPQPNSDISGVT GNSITTEFCTAQK QAFGDTDDFSQHG GLAKMGAAMQQ GMVLVMSLWDD YAAQMLWLDSYD PTDADPTTPIAR GTCPTDSGVPSDV ESQSPNSYVTYSNI KFGPINSTFTAS (SEQ ID NO:	GAATTCATGCTAAGAAGAGCTTTACTATTGAGCTCTTCT GCTATCTTGGCCGTTAAGGCTCAACAAGCCGGTACCGC TACTGCTGAAAACCACTCCATTGACCTGGCAAGAAT GTACCGCTCCAGGTTCTTGTACCACCAAAACGGTGCT GTCGCTTGGACGCTAACTGGAGATGGGTCCACGACGT CAACGGTTACACTAACTGTTACACCGGTAACACCTGGG ACCAACTTACTGTCCAGACGACGAAACTTGCCTCAA AACTGTGCCTTGGACGGTGCTGACTACGAAGGTACTTA CGGTGTTACCTCCTCTGGTTCTTCTTGAAGTTGAAGTT CGTCACTGGTTCTAACGTCGGTTCAGATTGTATTTGTT GCAAGATGACTCCACTTACCAAACTTCAAGTTGTTGA ACAGAGAATTTCTTTTCGACGTCGATGTGTCCAAGTTGC CTTGTGGTTTGAACGGTGCTCTATACTTCGTTGCTATGG ACGCTGATGGTGGTGGTTTCCAAGTACCAACAACAAG GCTGGTGCCAAATACGGTACTGGTACTGTGACTCTCA ATGTCCACGTGACTTGAAGTTTATTGATGGTGAAGCTA ATGTCAAGGTTGGCAACCATCTTCTAACAACGCTAAC ACTGGCATCGGTGACCACGGTCTTGTGCTGCGGAAAT GGACGTTTGGGAAGCCAACTCCATTTCACGCGCTCA CTCCACACCCATGTGACACTCCAGGTCAAATATGTGT TCCGGCGATGACTGTGGTGGTACTTACTTAACGATAG ATACGCTGGTACCTGTGATCCAGACGGTTGCGACTTCA ATCCATACAGAATGGGTAACACTTCCTTTTACGGTCCA GGCAAGATCATCGACACTACTAAGCCATTCACTGTTGT CACCAATTCTTGACCGACGATGGTACTGATACCGGTA CTTGTCCGAAATCAAGAGATTCTACATCCAAAACCTCT AACGTCATCCCAACAACCAATTCGACATCTCTGGTGT CACTGGTAACTCCATTACCACCGAATTTGTACCGCCCA AAAGCAAGCTTTTCGGTGACACCGACGACTTCTCTCAA ACGGTGGTTTGGCTAAGATGGGTGCTGCTATGCAACAA GGTATGGTTTTGGTCATGTCTTTGTGGGACGACTACGCT GCTCAAATGTTGTGGTGGACTCCGATTACCAACCGA TGCCGACCAACCAACCCCTGGTATCGCTAGAGGTACCT GTCCAAGTACTCTGGTGTTCATCTGACGTCGAATCCC AATCTCCAACTCCTACGTCACCTTACTCCAACATTAAAT TCGGTCCAATCAACTCCACTTTCACTGCTTCTTAAC7CG

	92)	AG (SEQ ID NO: 93)
--	-----	--------------------

- [0198] **Yeast transformation.** A protocol for electrotransformation of yeast was developed based on Cho KM *et al.*, "Delta-integration of endo/exo-glucanase and beta-glucosidase genes into the yeast chromosomes for direct conversion of cellulose to ethanol," *Enzyme Microb Technol* 25:23-30 (1999) and on Ausubel *et al.*, Current protocols in molecular biology. USA: John Wiley and Sons, Inc.(1994). Yeast cells for transformation were prepared by growing to saturation in 5mL YPD cultures. 4 mL of the culture was sampled, washed 2X with cold distilled water, and resuspended in 640 μ L cold distilled water. 80 μ L of 100mM Tris-HCl, 10mM EDTA, pH 7.5 (10X TE buffer—filter sterilized) and 80 μ L of 1M lithium acetate, pH 7.5 (10X LiAc—filter sterilized) were added and the cell suspension was incubated at 30°C for 45 min. with gentle shaking. 20 μ L of 1M DTT was added and incubation continued for 15 min. 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 μ L electroporation buffer.
- [0199] For electroporation, 100 ng of plasmid DNA (pRDH105) was combined with ~100ng of His3 PCR product and added to 50 μ L of the cell suspension in a sterile 1.5 mL microcentrifuge tube. A control strain was built by using 100ng each of the Ura3 and His3 PCR products. The mixture was then transferred to a 0.2 cm electroporation cuvette, and a pulse of 1.4 kV (200 Ω , 25 μ F) was applied to the sample using the Biorad Gene Pulser device. 1mL of cold 1M sorbitol adjusted to was placed in the cuvette and the cells were spread on Yeast nitrogen base media (Difco) with glucose, and not supplemented with amino acids.
- [0200] **Growth media and batch cultivation conditions.** Growth media with cellulose substrates as the sole carbon source were made using the non-glucose components of synthetic complete medium for yeast including, yeast nitrogen base without amino acids (YNB)—6.7 g/L, and in some cases supplementing with amino acids. Cultivation conditions included aerobic and microaerobic conditions. Aerobic conditions were maintained by using 250 mL shake flasks with Avicel™ (5%) containing media. For Avicel™ batch experiments with no added glucose, large inoculum sizes were used (20% by

volume) to speed the analysis of the strains. Batch shake flasks where cellobiose was added (at 10 g/L) were inoculated with 1mL of preculture (50mL total volume).

[0201] Microaerobic conditions were maintained by growing strains on BMCC in plastic tubes with an air atmosphere, with limited mixing due to the viscosity of the BMCC substrate. Inocula for the BMCC experiments were limited to 5% by volume. Solid media containing AvicelTM was generated as above, except that 1.5% agar was added. Plates were poured when the media was as cool as possible to prevent settling of the AvicelTM. 1.5% agar was also used to generate glucose plates, which contained YNB and 2% glucose.

Example 5: Selection of improved protrophic yeast strain

[0202] **Semi-continuous culture conditions.** Selection experiments using semi-continuous culture were carried out in 3L (total volume) Sartorius bioreactors. AvicelTM (~20 g/L; PH105 from FMC Biopolymer) was added to synthetic complete medium for yeast (Yeast nitrogen base without amino acids 6.7 g/L) lacking a carbon source. AvicelTM containing media was stirred in a 20L carboy and intermittently pumped into reactors with working volumes of ~900 mL. Media was pumped out in an intermittent fashion. For the control selection with glucose only, the AvicelTM component of the media was left out of the feed. Conditions in the reactors were maintained at pH ~5.5 by, using 2M KOH, stirring at 400 rpm, an aeration rate of 200 mL/min, and a temperature of 35 C. Glucose was fed to the culture continuously via a separate pump. Glucose addition rate, and overall dilution rate were quantified by measuring weight loss of the glucose feed tank, and accumulation in the waste tank respectively. Pumps were calibrated prior to use, and the feeding system was verified for consistency of the AvicelTM feed by running control experiments with AvicelTM and water only, and measuring the cellulose concentration over time.

[0203] **Quantification of cells and dry weight.** Cell concentration was measured by counting cells with a haemocytometer. Dry weight measurements were done by filtering a known amount of sample (determined by weighing tube before and after applying to the filter) and drying the filter at 105° C overnight to constant weight. Samples from BMCC cultures were taken using pipette tips that had been cut off to get consistent samples. Control cultures were run to generate a relationship between cell dry weight and cell

counts, which was then used to correct the total dry weight for the presence of cells. This relationship was determined to be 3×10^7 cells/mL = 0.37 g cells/L.

Results

[0204] Batch comparison of new strains on Avicel™. The newly created strain of *S. cerevisiae* made by transforming in the additional *T. emersonii* CBH1 expressed from a 2μ plasmid, and the His3 PCR product was named M0360. This CBH1 has previously been shown to be very highly expressed by *S. cerevisiae* strains (Unpublished data from Riaan Den Haan). It was compared directly to a prototrophic version of its parental strain called M0149, in batch shake flask cultivations on Avicel™ PH105 to see if improvements in the strain had been made. Figure 6 shows the results of that comparison. M0360 showed improved ability to utilize Avicel™ compared to M0149.

[0205] Long-term adaptation of M0360 in glucose/Avicel™ fed semi-continuous culture. Figure 7 presents data taken during a 100 generation adaptation of M0360. At 200 hours the dilution rate was slowed and the glucose feed increased as it was determined that these parameters might allow better observation of changes taking place with respect to cellulose utilization. Of importance is that the cell concentration in the reactor was increasing between 200 and 700 hours, even though the glucose feed rate was measured as slightly decreasing over this period of time (the decrease was ~5% of the feed rate—likely due to stretching of the tubing). Additionally, over the period of time from 200 to 700 hours there was an increase in the average evolved CO₂, as well as a slight decrease in the overall dry weight, and dry weight corrected for cell concentration over that period.

[0206] At 700 hours the dilution rate was slightly slowed, and the glucose feed rate was decreased. This was done to make conditions more favorable for strains able to metabolize cellulose more effectively. As can be observed in Figure 7, a similar adaptation as over the first period occurred. Cell concentration increased, CO₂ evolution rate increased, and dry weight decreased. The data in Figure 7 can also be used to calculate conversion of cellulose by this culture (shown in Figure 8). When conversion is calculated a dramatic rise in conversion over the course of the culture is observed in both

the period between 200 to 700 hours and 700 hours to 1500 hours. This indicates adaptation to use cellulose more effectively.

[0207] Figure 9 shows the cell count and CO₂ evolution data for a control experiment where only glucose was fed to the continuous culture. This was done to determine whether increases in cell counts and CO₂ evolution could be expected when only glucose was fed, or whether the AvicelTM feed was acting as a selective pressure. As can be seen, only modest increases in cell concentration were observed. Figure 10 shows a comparison in the relative increase in cell concentration for the glucose control as well as the “condition 2” and “condition 3” periods during the selective adaptation on AvicelTM. The “condition 2” period showed a modest increase in cell concentration relative to the starting concentration. The “condition 3” period showed a much more dramatic increase in cell concentration, indicating cells adapted to growth on cellulose have become predominant in the culture.

Example 6: Characterization of the improved strain M0360

[0208] **Batch comparison of selected and original strains on AvicelTM/cellobiose media and BMCC media.** A number of strains were isolated from the continuous reactor by dilution plating of a sample taken at the 50 generation point on YNB + glucose media. Figure 11 shows cell count data for batch shake flask cultures of selected and original strains on AvicelTM and cellobiose containing media. Cellobiose was chosen as a soluble sugar additive in order to replicate the conditions in the continuous culture as well as possible—namely a slow release of glucose to the cells over time. The left hand panel of Figure 11 demonstrates that a whole culture isolate (a streak of yeast colonies from the reactor) was able to grow significantly faster and to a greater extent on AvicelTM and cellobiose media, as well as on cellobiose media alone. To account for the fact that this could be due to increase cell yield on glucose, the data for AvicelTM and cellobiose cultures can be compared directly the cellobiose alone data. For the selected strain, the cell counts are higher when AvicelTM is present, whereas for the original strain the cell counts are similar on both media, although slightly lower when AvicelTM is present. The right hand panel shows the number of cells formed on both types of media per gram of cellobiose fed. This accounts for possible differences in the amount of cellobiose added to the batches, and also includes data gathered for a number of individually isolated colonies for

both the selected and original strains. These “cell yields” are higher for the selected strains compared to the original strains, and are higher on the AvicelTM containing media for the selected strains.

[0209] Figure 12 shows the results of a growth experiment with only BMCC present in the media. This experiment was run to confirm that cellulose utilization had indeed improved. As can be observed in the figure, the cell counts for the selected strains were repeatedly higher than those for the original strains.

[0210] **Growth test on 2% AvicelTM plates.** A test for growth on AvicelTM plates was also conducted for 30 isolates from both the original strain and the selected version of M0360. Figure 13 shows a picture of the 2% AvicelTM plate that the strains were streaked on. After 2 weeks of incubation at 30°C, the streaks of the selected colonies started to appear more substantially, than those of the original strain. The control CEN.PK strain from which the strains were originally constructed also did not show any biomass accumulation on these plates. Streaks from the selected colonies were examined under the microscope and were yeast cells—not contaminants.

Discussion

[0211] The present invention describes the construction of an improved cellulolytic *S. cerevisiae* strain, as well as a method for improving this strain via selection. A highly expressed and secreted CBH1 cellulase from *T. emersonii* was added to CP1_A1_C1#1, and the strain was also made prototrophic. This construction created a strain that was able to outgrow the parental strain, which was also made prototrophic, on AvicelTM PH105, when there were no amino acids present in the media. Given these results, media for selection could be formulated where the only carbon present was added as AvicelTM or glucose.

[0212] Long-term selection experiments were co-fed AvicelTM and glucose. A glucose alone control was also run. This method increased the cell number and growth rate in the system, while still allowing the cellulose to act as a selective agent. This allowed the (relatively) rapid passage of the strain through 100 generations of adaptation, and also maintained a relatively large number of cells in the reactor, increasing the genetic diversity in the system. Measurements of cell concentration and total dry weight in these systems indicated that improvements in the ability to utilize cellulose were potentially

happening. Comparison of cell increases to selection on only glucose indicated that these selective improvements were due to the presence of AvicelTM. Batch growth experiments with both AvicelTM and cellobiose present showed that the selected population was superior at utilizing cellulose under these conditions. Additionally, batch experiments where BMCC was supplied as the sole carbon source also showed a marked improvement in the ability of the selected strains to grow on cellulose. Averaged data for the selected strains showed that they grew up to a cell concentration of 1×10^7 cells/mL in only 40 hours, whereas for CP1_A1_C1#1 this took ~450 hours—a very remarkable improvement.

[0213] Colonies isolated from the reactor also showed the ability to form biomass on solid media containing 2% AvicelTM PH105 over a 2 week interval. This was not the case for the original strain, and represents a new level of cellulolytic capability not observed previously for *S. cerevisiae* strains.

[0214] The present invention presents a number of important steps forward for creating a yeast capable of consolidated bioprocessing. It describes improved cellulolytic yeast created by combining features of tethered and secreted cellulase systems. Additionally, it demonstrates the utility of selection-based techniques for improving cellulose utilization by recombinant strains. It further demonstrates selection-based improvements when growth is dependent on extracytoplasmic enzymes. The present invention demonstrates for the first time, the utility of long-term, well-mixed continuous cultures for improving the cellulose utilization of recombinant cellulolytic organisms.

[0215] These examples illustrate possible embodiments of the present invention. The scope of the claims should not be limited by the preferred embodiments set forth in the examples, but should be given the broadest interpretation consistent with the description as a whole.

WHAT IS CLAIMED IS:

1. A transformed host cell comprising:
 - (a) at least one heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase which when expressed is tethered to the cell surface;
 - (b) at least one heterologous polynucleotide comprising a nucleic acid which encodes a cellobiohydrolase which when expressed is tethered to the cell surface;
 - (c) at least one heterologous polynucleotide comprising a nucleic acid which encodes a β -glucosidase which when expressed is tethered to the cell surface; and
 - (d) at least one additional heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase, cellobiohydrolase, or β -glucosidase which when expressed, is secreted by the cell.
2. The host cell of claim 1, wherein the host cell is an organism selected from the group consisting of *Saccharomyces cerevisiae* (*S. cerevisiae*), *Kluyveromyces lactis*, *Kluyveromyces marxianus*, *Pichia pastoris*, *Yarrowia lipolytica*, *Hansenula polymorpha*, *Phaffia rhodozyma*, *Candida utilis*, *Arxula adeninivorans*, *Pichia stipitis*, *Debaryomyces hansenii*, *Debaryomyces polymorphus*, *Schizosaccharomyces pombe*, *Candida albicans*, and *Schwanniomyces occidentalis*.
3. The host cell of claim 1, wherein the at least one tethered endoglucanase is an endoglucanase I.
4. The host cell of claim 3, wherein the polynucleotide of (a) comprises a nucleic acid sequence that encodes endoglucanase I from *Trichoderma reesei* (*T. reesei*).
5. The host cell of claim 4, wherein the polynucleotide of (a) further comprises a linker sequence.

6. The host cell of claim 5, wherein the polynucleotide of (a) comprises the nucleotide sequence of SEQ ID NO: 20.
7. The host cell of claim 1, wherein the at least one tethered cellobiohydrolase is a cellobiohydrolase I.
8. The host cell of claim 7, wherein the polynucleotide of (b) comprises a nucleic acid sequence that encodes cellobiohydrolase I from *Trichoderma reesei* (*T. reesei*).
9. The host cell of claim 8, wherein the polynucleotide of (b) further comprises a linker sequence.
10. The host cell of claim 8 or 9, wherein the polynucleotide of (b) further comprises a cell wall anchoring sequence.
11. The host cell of claim 10, wherein the polynucleotide of (b) comprises the nucleotide sequence of SEQ ID NO: 21.
12. The host cell of claim 1, wherein the at least one tethered cellobiohydrolase is a cellobiohydrolase II.
13. The host cell of claim 12, wherein the polynucleotide of (b) comprises a nucleic acid sequence that encodes cellobiohydrolase II from *Trichoderma reesei* (*T. reesei*).
14. The host cell of claim 13, wherein the polynucleotide of (b) further comprises a linker sequence.
15. The host cell of claim 14, wherein the polynucleotide of (b) comprises the nucleotide sequence of SEQ ID NO: 22.
16. The host cell of claim 1, wherein the at least one tethered β -glucosidase is a β -glucosidase I (Bgl1) from *Saccharomycopsis fibuligera* (*S. fibuligera*).

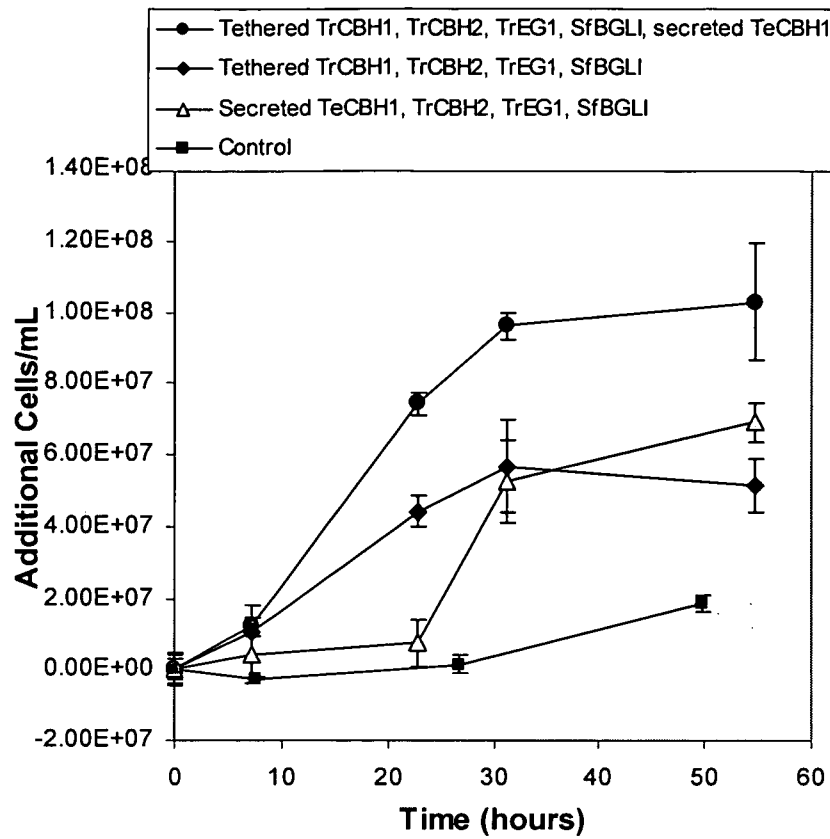
17. The host cell of claim 12, wherein the cellobiohydrolase II is from *Talaromyces emersonii* (*T. emersonii*).
18. The host cell of claim 1, wherein the at least one additional polynucleotide comprises a nucleic acid which encodes a cellobiohydrolase.
19. The host cell of claim 18, wherein the cellobiohydrolase, which when expressed, is secreted by the cell, is a cellobiohydrolase I.
20. The host cell of claim 19, wherein the polynucleotide of (d) comprises a nucleic acid sequence that encodes *Talaromyces emersonii* (*T. emersonii*) cellobiohydrolase I.
21. The host cell of claim 20, wherein the polynucleotide of (d) comprises the nucleotide sequence of SEQ ID NO: 1, SEQ ID NO: 9, or SEQ ID NO: 93.
22. The host cell of claim 19, wherein the polynucleotide of (d) comprises a nucleic acid sequence that encodes a fusion protein comprising *Talaromyces emersonii* (*T. emersonii*) Cbh1.
23. The host cell of claim 22, wherein the fusion protein comprises *T. emersonii* Cbh1 and the cellulose binding module (CBM) of *Trichoderma reesei* (*T. reesei*) Cbh2.
24. The host cell of claim 22, wherein the fusion protein comprises *T. emersonii* Cbh1 and the cellulose binding module (CBM) of *Trichoderma reesei* (*T. reesei*) Cbh1.
25. The host cell of claim 23 or 24, wherein the CBM is fused to the *T. emersonii* Cbh1 via a linker sequence.
26. The host cell of claim 1, wherein the at least one additional heterologous polynucleotide comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 1-10.
27. The host cell of claim 2, wherein the host cell is *S. cerevisiae*.

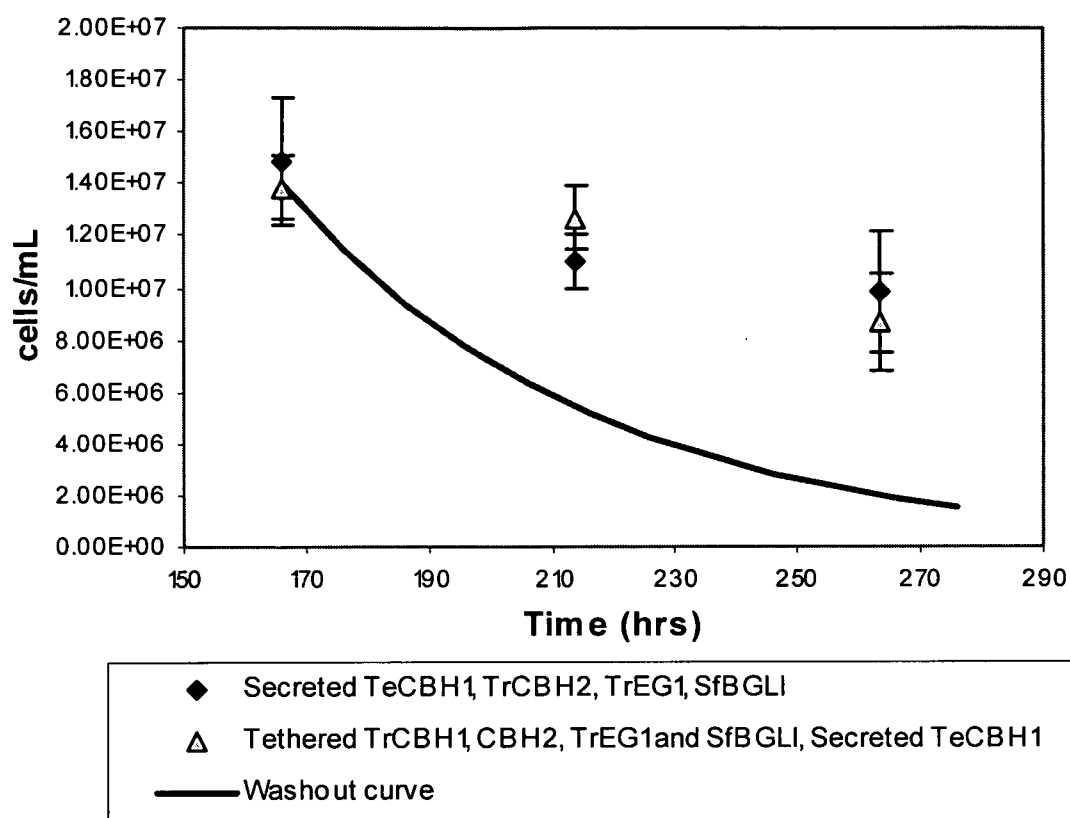
28. The host cell of claim 1, wherein the at least one additional polynucleotide comprises a nucleic acid which encodes an endoglucanase.
29. The host cell of claim 28, wherein the endoglucanase, which when expressed, is secreted by the cell, is an endoglucanase I from *Trichoderma reesei* (*T. reesei*).
30. The host cell of claim 29, wherein the nucleic acid encoding the endoglucanase I further comprises a linker sequence.
31. The host cell of claim 30, wherein the nucleic acid encoding the endoglucanase I comprises the nucleotide sequence of SEQ ID NO: 20.
32. The host cell of claim 1, wherein the at least one additional polynucleotide comprises a nucleic acid which encodes an endoglucanase II.
33. The host cell of claim 32, wherein the endoglucanase II is from *Trichoderma reesei* (*T. reesei*).
34. The host cell of claim 18, wherein the cellobiohydrolase, which when expressed, is secreted by the cell, is a cellobiohydrolase II.
35. The host cell of claim 34, wherein the polynucleotide of (d) comprises a nucleic acid sequence that encodes *Talaromyces emersonii* (*T. emersonii*) cellobiohydrolase II.
36. The host cell of claim 35, wherein the polynucleotide of (d) comprises the nucleotide sequence of SEQ ID NO: 5 or SEQ ID NO: 10.
37. The host cell of claim 1, comprising:
 - (a) one heterologous polynucleotide comprising a nucleic acid which encodes an endoglucanase I which when expressed is tethered to the cell surface;

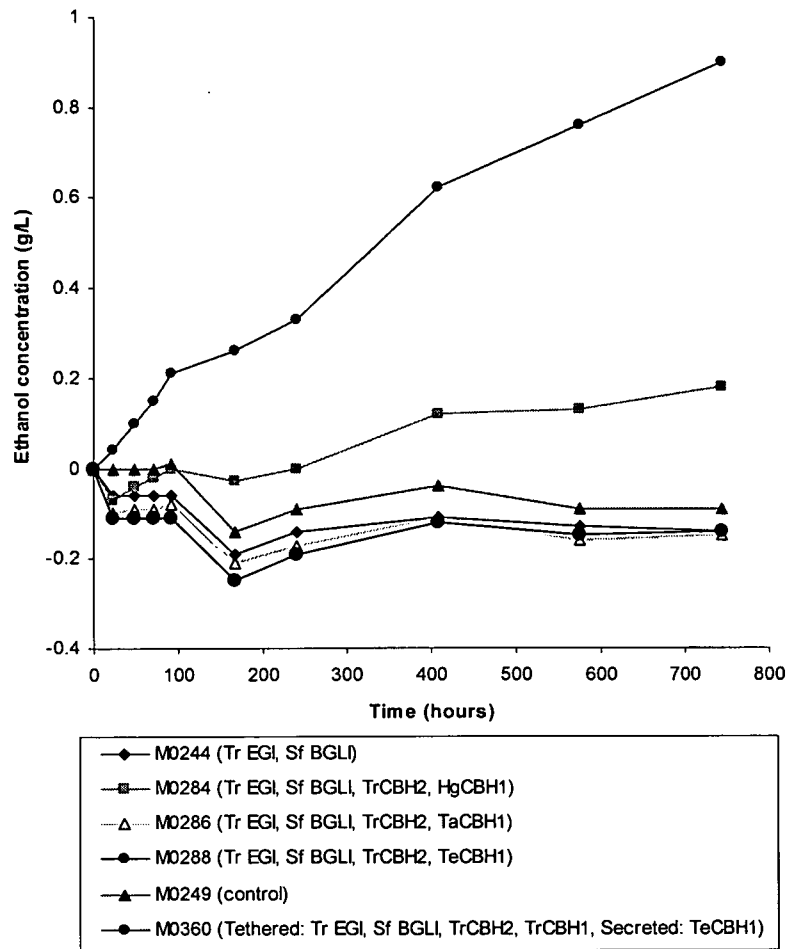
- (b) two heterologous polynucleotides comprising nucleic acids which encode a cellobiohydrolase I and a cellobiohydrolase II which when expressed are tethered to the cell surface;
 - (c) one heterologous polynucleotide comprising a nucleic acid which encodes a β -glucosidase which when expressed is tethered to the cell surface; and
 - (d) one additional heterologous polynucleotide comprising a nucleic acid which encodes a cellobiohydrolase which, when expressed, is secreted by the cell.
38. The host cell of claim 37, wherein the nucleic acid which encodes a secreted cellobiohydrolase encodes a cellobiohydrolase I comprising a nucleotide sequence selected from the group consisting of SEQ ID NOs: 1-4 and 7-9.
39. The host cell of claim 37, wherein the nucleic acid which encodes the cellobiohydrolase I comprises the nucleotide sequence of SEQ ID NO: 9.
40. The host cell of claim 37, wherein the tethered endoglucanase is *Trichoderma reesei* (*T. reesei*) Eg1, the tethered cellobiohydrolase I and cellobiohydrolase II are *T. reesei* Cbh1 and Cbh2, the tethered β -glucosidase is *Saccharomycopsis fibuligera* (*S. fibuligera*) Bgl1, and the secreted cellobiohydrolase is *Talaromyces emersonii* (*T. emersonii*) Cbh1 or a fusion protein comprising *T. emersonii* Cbh1.
41. The host cell of claim 40, wherein the fusion protein comprises *T. emersonii* Cbh1 and the cellulose binding module (CBM) of *T. reesei* Cbh2.
42. The host cell of claim 40, wherein the fusion protein comprises *T. emersonii* Cbh1 and the cellulose binding module (CBM) of *T. reesei* Cbh1.
43. The host cell of claim 41 or 42, wherein the CBM is fused to the *T. emersonii* Cbh1 via a linker sequence.

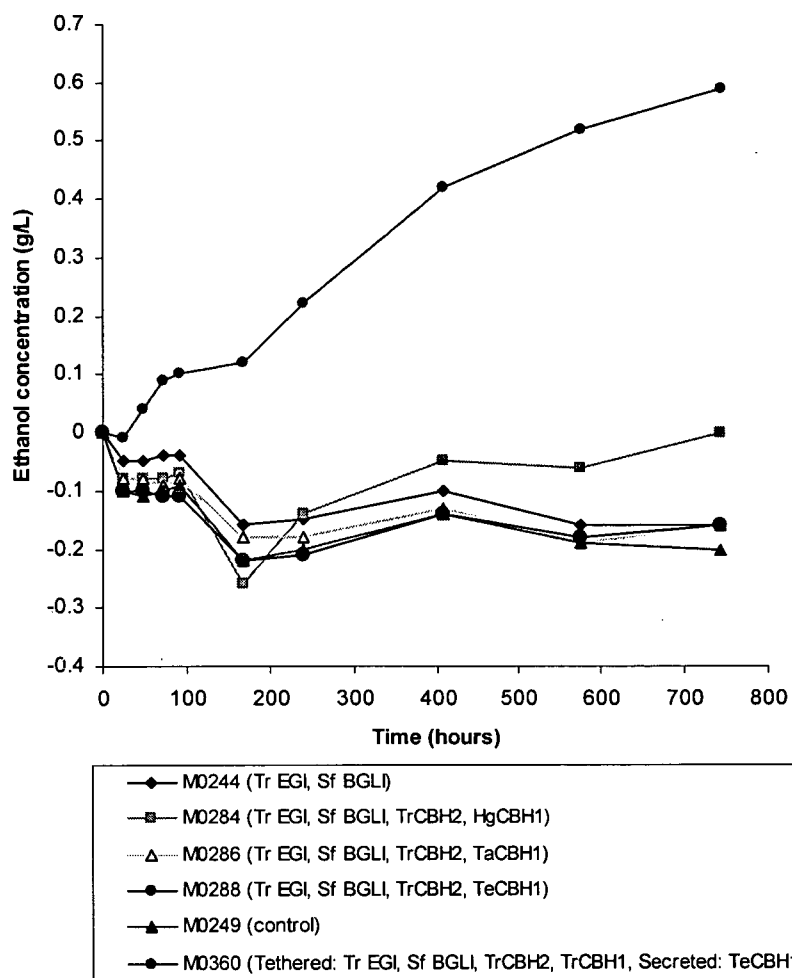
44. The host cell of claim 7 or 19, wherein the cellobiohydrolase I comprises an amino acid sequence at least 95% identical to an amino acid sequence selected from any one of the cellobiohydrolases of the group consisting of SEQ ID NOs: 11-13, 17 and 40.
45. The host cell of any one of claims 1-44, wherein the host cell has the ability to saccharify crystalline cellulose.
46. The host cell of claim 45, wherein the host cell has the ability to ferment the crystalline cellulose.
47. A method of fermenting cellulose using the host cell of any one of claims 1-46, the method comprising culturing the transformed host cell in medium that contains crystalline cellulose under suitable conditions for a period sufficient to allow saccharification and fermentation of the cellulose.
48. The method of claim 47, wherein the host cell produces ethanol.
49. The host cell of any one of claims 1-46, wherein the host cell is producing ethanol when grown using insoluble cellulose as a carbon source.
50. The host cell of any one of claims 1-46, wherein the host cell is producing ethanol when grown using insoluble cellulose as the sole carbon source.
51. The host cell of claim 49 or 50, wherein the insoluble cellulose is crystalline.
52. The host cell of any one of claims 49-51, wherein the host cell is *Saccharomyces cerevisiae* (*S. cerevisiae*).
53. A method of improving the ability of the host cell of any one of claims 1-46 to use cellulose as a carbon source comprising:
 - (a) culturing the host cell in media containing cellulose;

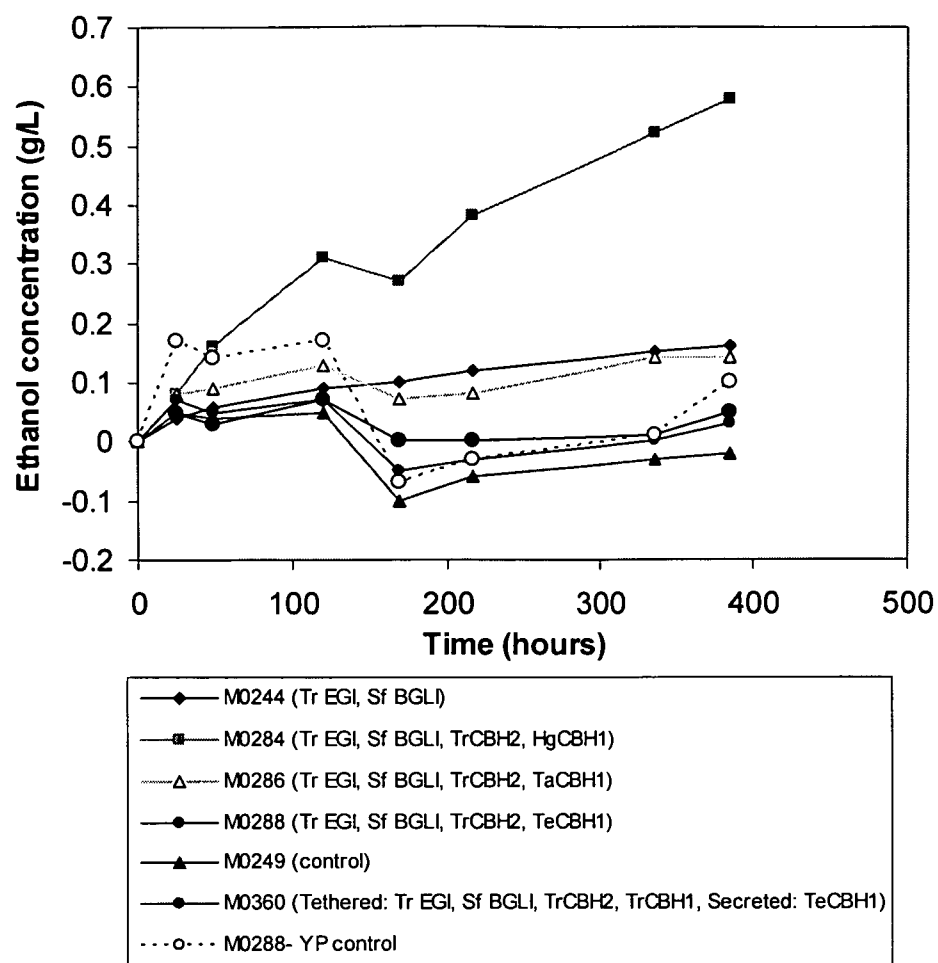
- (b) maintaining the culture conditions in a substantially steady state whereby variant progeny cells acquire a selective advantage in the culture according to the ability of the variant progeny to display increased reproductive capacity in the cellulose-containing media;
 - (c) continuously growing the variant progeny cells of step (b) on media containing cellulose; and,
 - (d) repeating steps (b) and (c) until a variant of the host cell is produced, wherein the variant acquires the ability to grow to a cell density of at least 2 fold greater than the host cell on cellulose as a sole carbon source.
54. The method of claim 53, wherein the sole cellulose carbon source is crystalline cellulose.
55. The method of claim 54, wherein the sole cellulose carbon source is selected from the group consisting of bacterial microcrystalline crystalline cellulose (BMCC) and phosphoric acid swollen cellulose (PASC).
56. The host cell of claim 20, wherein the polynucleotide of (d) comprises a nucleotide sequence at least 90% identical to SEQ ID NO: 1, SEQ ID NO: 9, or SEQ ID NO: 93.
57. The host cell of claim 20, wherein the polynucleotide of (d) comprises a nucleotide sequence at least 95% identical to SEQ ID NO: 1, SEQ ID NO: 9, or SEQ ID NO: 93.

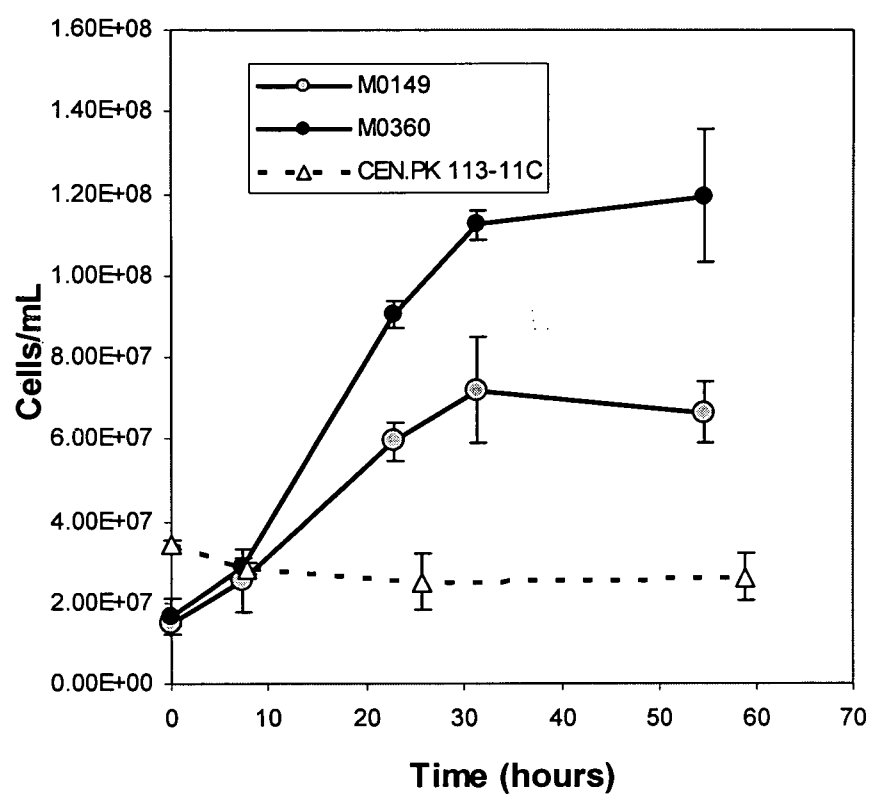
**Figure 1**

**Figure 2**

**Figure 3**

**Figure 4**

**Figure 5**

**Figure 6**

7/12

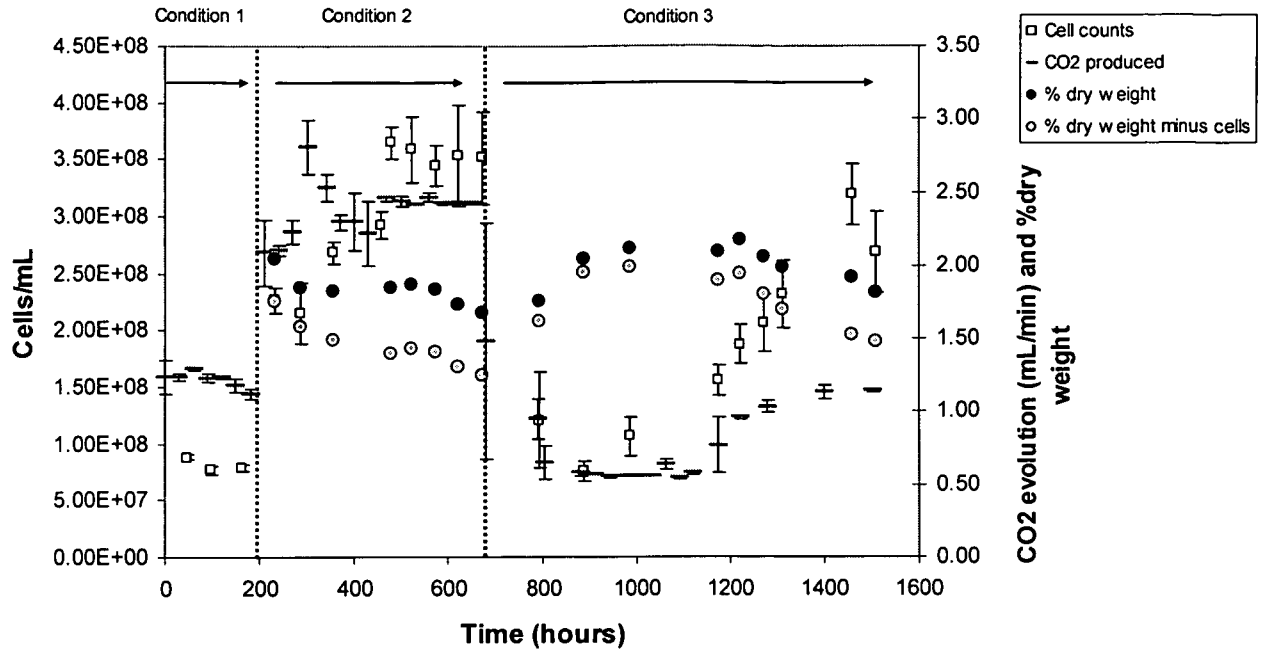


Figure 7

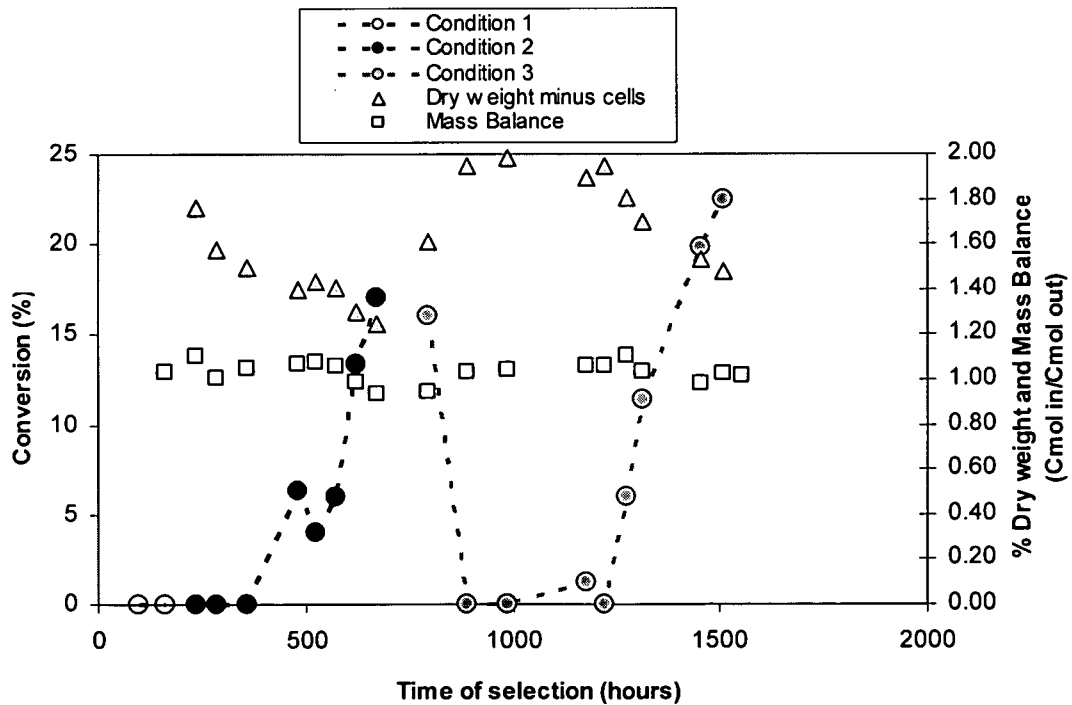
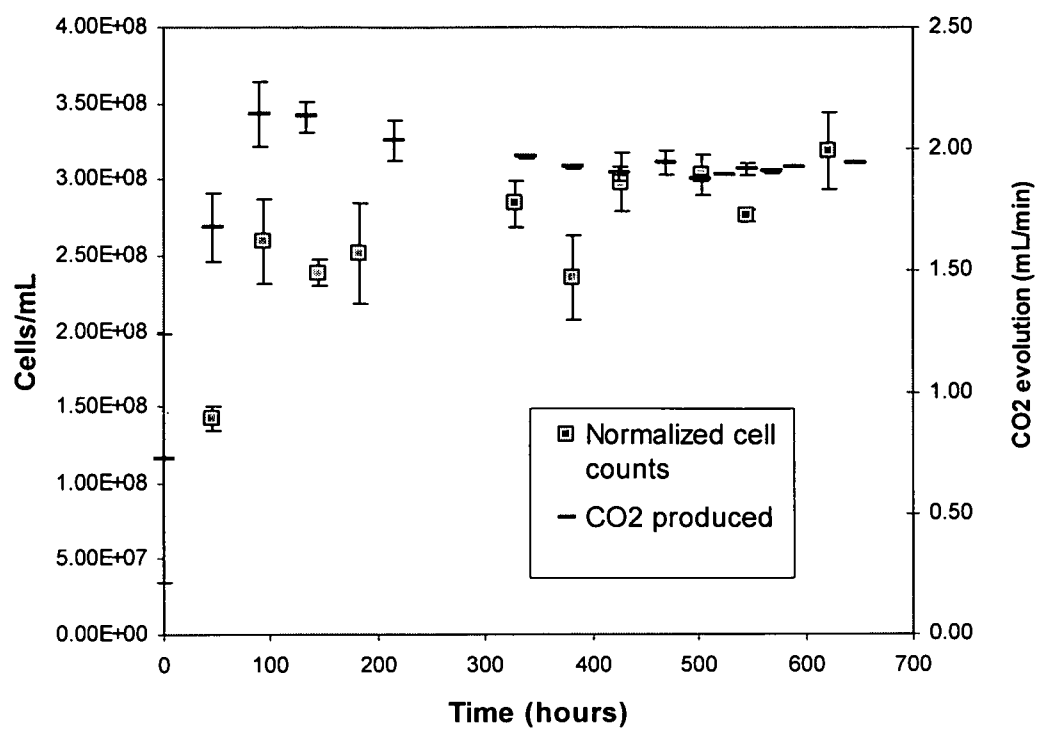
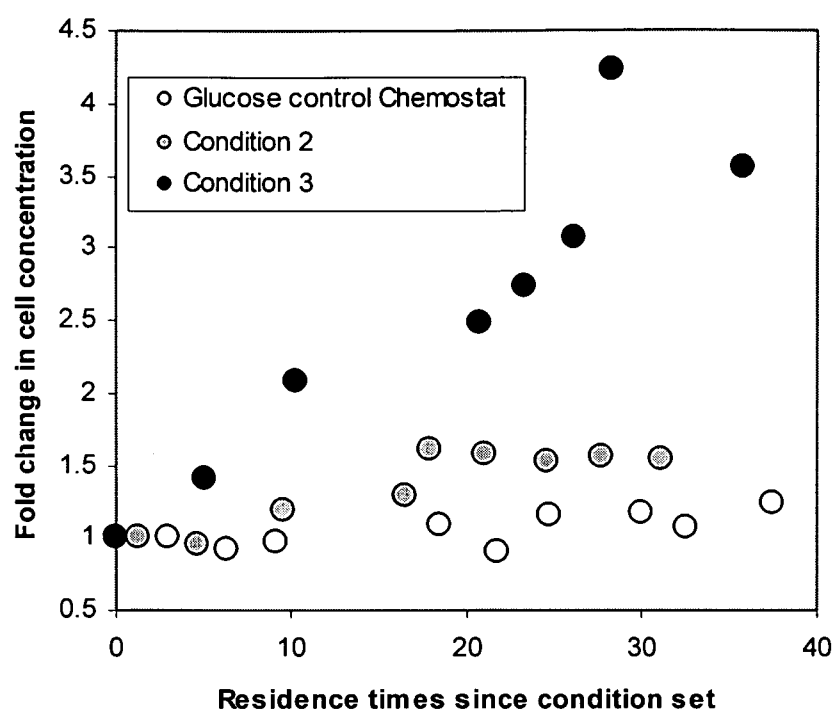


Figure 8

**Figure 9**

**Figure 10**

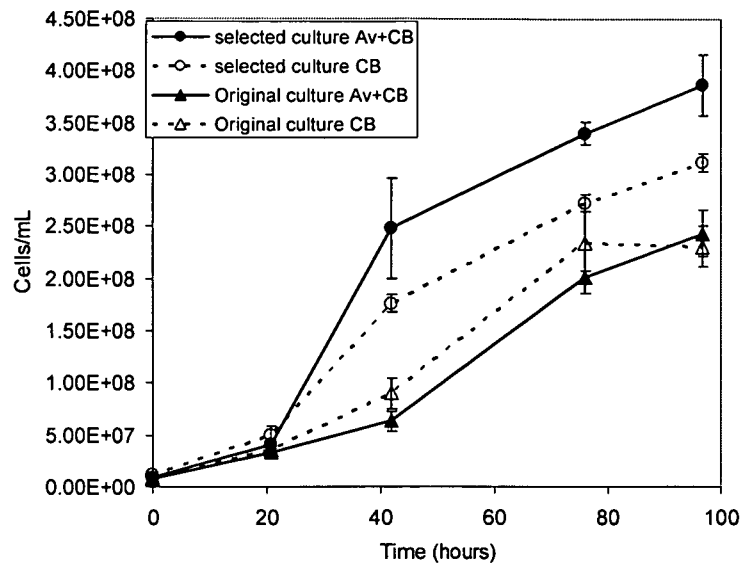


Figure 11A

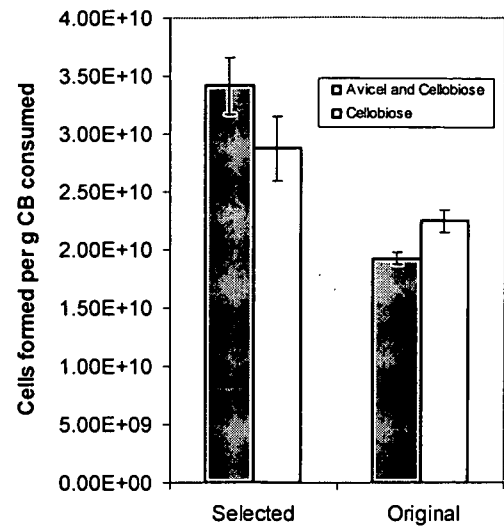
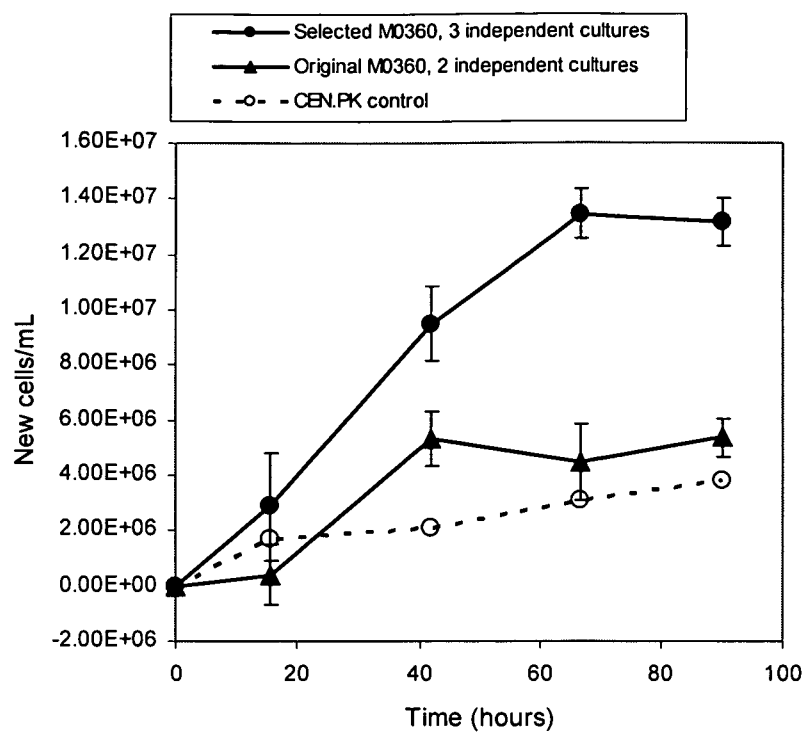
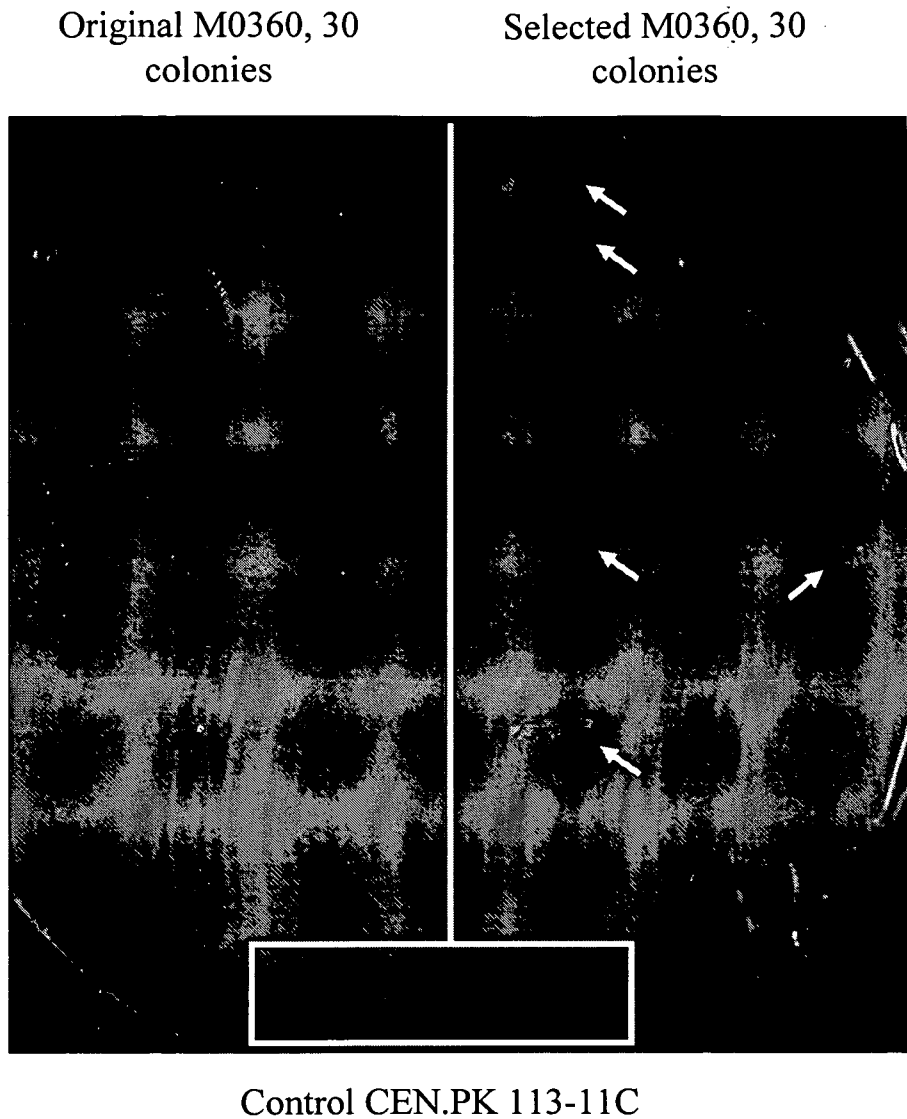


Figure 11B

11/12

**Figure 12**

**Figure 13**