



(51) International Patent Classification:

Not classified

(21) International Application Number:

PCT/US2017/062100

(22) International Filing Date:

16 November 2017 (16.11.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/423,196 16 November 2016 (16.11.2016) US
 15/469,452 24 March 2017 (24.03.2017) US
 62/481,601 04 April 2017 (04.04.2017) US

(71) Applicant: **THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY** [US/US];
 Office Of General Counsel, Building 170, 3rd Floor, Main Quad, P.o.box 20386, Stanford, CA 94305-2038 (US).

(72) Inventors: **NAUGHTON, Brian, Thomas**; 217 Ada Avenue, #51, Mountain View, CA 94043 (US). **HARVEY, Colin**; 572 Franklin Street, Mountain View, CA 94041

(US). **SCHLECHT, Ulrich**; 226 Morse Avenue, Sunnyvale, CA 94086 (US). **HILLENMEYER, Maureen, Elizabeth**; 365 N California Avenue, Palo Alto, CA 94301 (US). **HORECKA, Joe**; 38448 Darnel Court, Fremont, CA 95536 (US).

(74) Agent: **THOMAS, Charles, A.**; KPPB LLP, 2190 S. Towne Centre Place, Suite 300, Anaheim, CA 92806 (US).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,

(54) Title: SYSTEMS AND METHODS FOR IDENTIFYING AND EXPRESSING GENE CLUSTERS

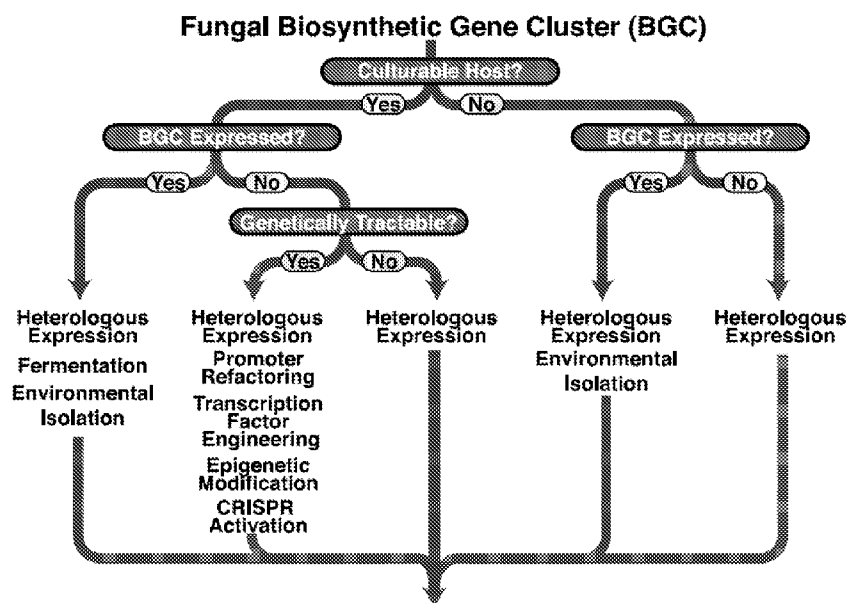


FIG. 1A

(57) Abstract: Methods for identifying biosynthetic gene clusters that include genes for producing compounds that interact with specific target proteins are disclosed. Some methods relate to bioinformatics methods for identifying and/or prioritizing biosynthetic gene clusters. Related systems, components, and tools for the identification and expression of such gene clusters are also disclosed.



UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*
- *with sequence listing part of description (Rule 5.2(a))*

SYSTEMS AND METHODS FOR IDENTIFYING AND EXPRESSING GENE CLUSTERS**CROSS-REFERENCE**

[1] This application claims the benefit of U.S. Provisional Application No. 60/243,196, filed November 16, 2016, U.S. Provisional Application No. 62/481,601, filed April 4 2017, and U.S. Serial No. 15/469,452, filed March 24, 2017, which applications are incorporated herein by reference.

REFERENCE TO A SEQUENCE LISTING SUBMITTED ELECTRONICALLY VIA EFS-WEB

[2] The instant application contains a Sequence Listing which has been filed electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on November 9, 2017, is named 52592-702_601_SL.txt and is 1,208,543 bytes in size.

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[3] This invention was made with Government support under U01 GM110706 awarded by the National Institutes of Health. The government has certain rights to the invention.

TECHNICAL FIELD

[4] The present disclosure generally relates to the introduction of gene clusters into host organisms for the manufacture of small molecules. In some cases, the small molecules are analogs to products produced by the organism in which the gene cluster is identified and that modulate a specific protein. More specifically, the disclosure also relates to the identification of gene clusters likely to produce products that target a specific target protein and methods of expressing gene clusters in host cells. Additionally sequences of various gene clusters are provided, together with structures of compounds produced from these gene clusters.

BACKGROUND

[5] "Secondary metabolites," as used herein, are small molecules that can be produced by the expression of one or more gene clusters. Often, secondary metabolites are not critical for the survival of the organism in which the gene cluster is natively found. Secondary metabolites can be clinically valuable small molecules. Examples include: antibacterial products such as penicillin, and daptomycin; antifungal products such as amphotericin; cholesterol-lowering products such as

lovastatin; anticancer products such as taxol, and eribulin; and immune-modulating products including rapamycin, and cyclosporine. In spite of the great success of secondary metabolites in the history of drug discovery, challenges of secondary metabolites in drug discovery and development can include (i) extremely low yields, (ii) limited supply, (iii) complex structures posing difficulty for structural modifications, and (iv) complex structures precluding practical synthesis. These difficulties have prompted the pharmaceutical industry to embrace new technologies in past decades, particularly combinatorial chemistry, as an alternative to natural product discovery. As a result, the percentage of new secondary metabolites being tested for use in medical treatment of humans has declined steadily since the 1940s due to a greater reliance on synthetic libraries that can be utilized in high throughput screening. Despite the pharmaceutical industry's preference for synthetic libraries, secondary metabolites possess enormous structural and chemical diversity that is unsurpassed by synthetic libraries. Most importantly, secondary metabolites are often evolutionarily optimized as drug-like molecules to target specific proteins and/or pathways.

[6] Now, thousands of bacterial and fungal genomes have been sequenced. These organisms are known to be rich sources of secondary metabolites. These secondary metabolites are enzymatically biosynthesized by the products of one or more genes, often grouped into gene clusters. New genome sequences have revealed that traditional approaches have tapped only a fraction of the biosynthetic potential of these organisms as, on average, fewer than 10% of the biosynthetic gene clusters (BGCs) in a microbial genome are expressed in any single culture condition. Further, millions of fungal species are believed to exist in nature, but have not been cultured in the laboratory. Accordingly, the supply bottleneck for secondary metabolites can be reduced by introducing the genes utilized in the synthesis of the secondary metabolite into a microorganism that can overproduce a desired secondary metabolite or analog thereof. In this way, the vast, untapped, ecological biodiversity of microbes holds renewed promise for the discovery of novel secondary metabolites useful in one or more contexts, such as the treatment of disease.

SUMMARY OF THE INVENTION

[7] In some embodiments, this disclosure refers to a method for screening a plurality of compounds, the method comprising: identifying a gene cluster that includes or is within 20 kilobases of a region that encodes for a protein that is identical with or homologous to a first target protein,

wherein the gene cluster comprises a region that encodes for a protein selected from the group consisting of (1) polyketide synthases and (2) non-ribosomal peptide synthetases, (3) terpene synthetases, (4) UbiA-type terpene cyclases, and (5) dimethylallyl transferases; introducing a plurality of genes from the gene cluster into a vector; introducing the vector into a host cell; expressing the proteins encoded by the plurality of genes in the host cell; and determining whether a compound that is formed or modified by the expressed proteins modulates the first target protein. In some cases, the host cell is a yeast cell. In some cases, the yeast cell is a yeast cell that has been modified to increased sporulation frequency and increased mitochondrial stability. In some cases, each gene of the plurality of genes is under the control of a different promoter. In some cases, the promoters are designed to increase expression when the host cell is in the presence a nonfermentable carbon source. In some cases, the plurality of genes are introduced into the vector via homologous recombination. In some cases, introducing the plurality of genes into the vector via homologous recombination comprises combining a first plurality of nucleotides with a second plurality of nucleotides, wherein: each polynucleotide of the first plurality of DNA polynucleotides encodes for a promoter and terminator, wherein each promoter and terminator is distinct from the promoter and terminator of other polynucleotides of the first plurality of nucleotides; and each polynucleotide of the second plurality of nucleotides includes a coding sequence, a first flanking region on the 5' side of the polynucleotide and a second flanking region on the 3' side of the polynucleotide; and introducing the polynucleotides into a host cell that includes machinery for homologous recombination, wherein the host cell assembles the expression vector via homologous recombination that occurs in the flanking regions of the second plurality of polynucleotides; wherein the expression vector is configured to facilitate simultaneous production of a plurality of proteins encoded by the second plurality of nucleotides. In some cases, the first flanking region and the second flanking region are each between 15 and 75 base pairs in length. In some cases, the first flanking region and the second flanking region are each between 40 and 60 base pairs in length. In some embodiments, the present disclosure provides a method for identifying a gene cluster capable of producing a small molecule for modulating a first target protein, the method comprising: selecting, from a database comprising a list of biosynthetic gene clusters, one or more gene clusters that include or are positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein. In some cases, the one or more gene clusters are selected

from the group consisting of (1) clusters that comprise one or more polyketide synthases and (2) clusters that comprise one or more non-ribosomal peptide synthetases, (3) clusters that comprise one or more terpene synthases, (4) clusters that comprises one or more UbiA-type terpene cyclases, and (5) clusters that comprise one or more dimethylallyl transferases. In some cases, the protein that is encoded by the region that is included in or positioned proximal to the biosynthetic gene cluster is identical to or has greater than 30% homology to the first target protein. In some cases, the region that encodes the protein that is identical with or homologous to the first target protein is within 20,000 base pairs of a region of a portion of the gene cluster that encodes a polyketide synthase, a non-ribosomal peptide synthetase, a terpene synthetase, a UbiA-type terpene cyclase, or a dimethylallyl transferase. In some cases, the first target protein is BRSK1. In some cases, selecting the one or more gene clusters comprises operating a computer, wherein operation of the computer comprises running an algorithm that takes into account both an input sequence for the first target protein and sequence information from a database that includes sequence information from a plurality of species such that the computer returns information corresponding to one or more gene clusters. In some cases, the algorithm takes into account the phylogenetic relationship between gene clusters in the database. In some cases, the one or more gene clusters include a coding sequence for a protein that is an extracellular protein, a membrane-tethered protein, a protein involved in a transport or secretion pathway, a protein homologous to a protein involved in a transport or secretion pathway, a protein with a peptide targeting signal, a protein with a terminal sequence with homology to a targeting signal, an enzyme that degrades small molecules, or a protein with homology to an enzyme that degrades small molecules. In some embodiments the present disclosure provides a method for producing a compound that modulates the first target protein, the method comprising: identifying a gene cluster (e.g., via methods disclosed herein), expressing the gene cluster or a plurality of genes from the gene cluster in a host cell, and isolating a compound produced by the gene cluster. In some cases, the method further comprises screening the isolated compound for modulation of an activity of the first target protein. In some cases, the cluster-encoded protein that is homologous to the first target protein is resistant to modulation by the isolated compound when compared to modulation of the first target protein. In some cases, the compound is not toxic to the species from which the cluster originates due to one or more of (1) sequence differences between

target protein and the cluster-encoded protein, (2) spatial separation of the compound from the cluster-encoded protein and (3) high expression levels for the cluster-encoded protein.

[8] In some embodiments the present disclosure provides a method for making a DNA vector, the method comprising: identifying a gene cluster comprising a plurality of genes that are capable of producing a small molecule for modulating a first target protein; introducing two or more genes of the plurality of genes into a vector, wherein the vector is configured to facilitate expression of the two or more genes in a host organism; wherein (1) the gene cluster encodes one or more proteins selected from the group consisting of a polyketide synthase, a non-ribosomal peptide synthetase, a terpene synthetase, a UbiA-type terpene cyclase, and a dimethylallyl transferase and (2) the gene cluster includes or is positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein. In some cases, the DNA vector is a circular plasmid. In some cases, the DNA vector comprises a plurality of promoters, wherein each promoter of the plurality of promoters is configured to, when the vector is introduced into a *Saccharomyces cerevisiae* cell, promote a lower level of heterologous expression when the cell exhibits predominantly anaerobic energy metabolism than when the cell exhibits aerobic energy metabolism. In some cases, each promoter of the plurality of promoters differs in sequence from one another. In some cases, each promoter of the plurality of promoters has a sequence selected from the group consisting of: SEQ ID NOs: 1–66. In some cases, each promoter of the plurality of promoters has a sequence selected from the group consisting of: SEQ ID NOs: 20–35, and SEQ ID NOs: 41–50. In some cases, when the *Saccharomyces cerevisiae* cell is exhibiting anaerobic energy metabolism, the cell is catabolizing a fermentable carbon source selected from glucose or dextrose; and when the *Saccharomyces cerevisiae* cell is exhibiting aerobic energy metabolism, the cell is catabolizing a non-fermentable carbon source selected from ethanol or glycerol.

[9] In some embodiments the present disclosure provides a method for the heterologous expression of a plurality of genes in a yeast strain, the method comprising: obtaining a yeast strain that includes a vector for expressing a plurality of genes from a single gene cluster of a non-yeast organism; and inducing expression of the plurality of genes; wherein the gene cluster in the non-yeast organism includes or is positioned proximal to a region that encodes a protein that is at least 30% homologous to a target protein. In some cases, the method further comprises introducing the plurality of genes from the single gene cluster into the vector. In some cases, the method further

comprises introducing the vector into the yeast strain. In some cases, expression of the plurality of genes results in the formation of small molecule, wherein the small molecule modulates the activity of the target protein. In some cases, the gene cluster is a gene cluster of a non-yeast fungus. In some cases, the yeast strain is from *Saccharomyces cerevisiae*. In some cases, the target protein is a human protein.

[10] In some embodiments the present disclosure provides a system for identifying a gene cluster for introduction of a plurality of genes from the gene cluster into a host organism, the system comprising: a processor; a non-transitory computer-readable medium comprising instructions that, when executed by the processor, cause the processor to perform operations, the operations comprising: loading the identity or sequence of a first target protein into memory; loading the identity or sequence of a plurality of biosynthetic gene clusters into memory; identifying, from the plurality of biosynthetic gene clusters, one or more gene clusters that encode or are positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein; and scoring the one or more gene clusters based on the likelihood of each gene cluster being capable of use to produce a small molecule that modulates the first target protein. In some cases, scoring the one or more gene clusters comprises comparing the sequence of the first target protein (or a DNA sequence encoding the first target protein) to a sequence of a protein encoded in or proximal to the gene cluster (or to a DNA sequence encoding the protein that is in or proximal to the gene cluster).

[11] In some embodiments the present disclosure provides a system for identifying one or more biosynthetic gene clusters for introduction into a host organism to produce one or more compounds that modulate a specific target protein, the system comprising: a processor; a memory containing a gene cluster identification application; wherein the gene cluster identification application directs the processor to: load data describing at least one target protein into the memory; load data describing a plurality of biosynthetic gene clusters into the memory; score each of the plurality of biosynthetic gene clusters based upon: performing a homolog search for each biosynthetic gene cluster to determine a presence of at least one homolog of a target protein within or adjacent the biosynthetic gene cluster; confidence of homology of the at least one target protein to at least one gene in a biosynthetic gene cluster; a fraction of a homologous gene that meets an identity threshold; a total number of genes homologous to the at least one target protein present in the entire genome of an

organism; homology of the at least one homolog of at least one target protein within or adjacent the biosynthetic gene cluster to genes in the target protein's genome; phylogenetic relationship of the at least one target protein to a gene in a cluster; expected number of homologs of the at least one target protein in or adjacent to a biosynthetic cluster; and a likelihood that at least one target protein is essential for cellular process in the natural environment; and output a report identifying one or more biosynthetic gene clusters that are most likely to produce a compound that modulates the at least one target protein.

[12] In some embodiments the present disclosure provides a method for selecting a biosynthetic gene cluster that produces a secondary metabolite, the method comprising: obtaining a list of gene clusters; performing a phylogenetic analysis of the genes within the clusters compared to known genes from known biosynthetic gene clusters; and selecting the biosynthetic gene cluster based on its phylogenetic relationship with the known genes. In some cases, the biosynthetic gene cluster with the most distant phylogenetic relationship from the known genes is selected.

[13] In some embodiments the present disclosure provides a method for identifying a gene cluster that produces a compound that binds a protein of interest, the method comprising: obtaining sequence information for a plurality of contiguous sequences, wherein each contiguous sequence includes a biosynthetic gene cluster and flanking genomic sequences; analyzing the contiguous sequences for the presence of a gene that encodes a protein with homology to the protein of interest, and selecting a biosynthetic gene cluster which includes, or is proximal to, a gene that encodes a protein that is homologous to the protein of interest. In some cases, the contiguous nucleotide sequence is less than 40,000 base pairs in length.

[14] In some embodiments, the present disclosure provides a modified yeast cell having a BY background, wherein relative to unmodified BY4741 and BY4742, the modified yeast cell has both (1) increased sporulation frequency and (2) increased mitochondrial stability. In some cases, the modified yeast cell grows faster on non-fermentable carbon sources than unmodified BY4741 and BY4742. In some cases, the yeast cell comprises one or more of the following genotypes: MKT1(30G), RME1(INS-308A), and TAO3(1493Q). In some cases, the yeast cell comprises one or more of the following genotypes: CAT5(91M), MIP1(661T), SAL1+, and HAP1+. In some embodiments the present disclosure provides a method of forming an expression vector, the method comprising: combining a first plurality of DNA polynucleotides with a second plurality of

polynucleotides, wherein: each polynucleotide of the first plurality of DNA polynucleotides encodes for a promoter and terminator, wherein each promoter and terminator is distinct from the promoter and terminator of other polynucleotides of the first plurality of nucleotides; and each polynucleotide of the second plurality of nucleotides includes a coding sequence, a first flanking region on the 5' side of the polynucleotide and a second flanking region on the 3' side of the polynucleotide, wherein each flanking region is between 15 and 75 base pairs in length; and introducing the polynucleotides into a host cell that includes machinery for homologous recombination, wherein the host cell assembles the expression vector via homologous recombination that occurs in the flanking regions of the second plurality of polynucleotides; wherein the expression vector is configured to facilitate simultaneously production of a plurality of proteins encoded by the second plurality of nucleotides. In some cases, the host cell is a yeast cell. In some cases, each flanking region is between 40 and 60 base pairs in length. In some cases, at least one polynucleotide of the first plurality of nucleotides encodes a selection marker.

[15] In some embodiments, the present disclosure provides a system for generating a synthetic gene cluster via homologous recombination, the system comprising 1 through N unique promoter sequences, 1 through N unique terminator sequences, and 1 through N unique coding sequences, wherein: coding sequence 1 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 1 and a second end portion is identical to the first 30–70 base pairs of terminator 1; coding sequence 2 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 2 and a second end portion is identical to the first 30–70 base pairs of terminator 2; and coding sequence N is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter N and a second end portion is identical to the first 30–70 base pairs of terminator N. In some cases, terminator 1 and promoter 2 are portions of the same double-stranded oligonucleotide.

[16] In some embodiments, the present disclosure provides a method for assembling a synthetic gene cluster, the method comprising: obtaining 1 through N unique promoters, 1 through N unique terminators, and 1 through N unique coding sequences, wherein: coding sequence 1 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 1 and a second end portion is identical to the first 30–70 base pairs of

terminator 1; coding sequence 2 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 2 and a second end portion is identical to the first 30–70 base pairs of terminator 2; and coding sequence N is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter N and a second end portion is identical to the first 30–70 base pairs of terminator N; transforming the 1 through N promoters, terminators and coding sequences into a yeast cell; isolating a plasmid containing the 1 through N promoters, terminators and coding sequences from the yeast cell. In some cases, the method further comprises a coding sequence for a selection marker. In some cases, the selection marker is an auxotrophic marker. In some cases, the yeast cell has a deficiency in a DNA ligase gene.

[17] In some embodiments the present disclosure provides a yeast strain which allows for both (1) homologous DNA assembly and (2) production of heterologous genes in the same strain. In some cases, the yeast strain is a DHY strain. In some cases, the strain allows DNA assembly via homologous recombination with an efficiency of at least 80% as compared to DNA assembly in BY. In some cases, production of heterologous compounds in the strain is accomplished with an efficiency of at least 80% as compared to heterologous compound production in BJ5464. In some cases, the strain allows production of heterologous proteins with an efficiency of at least 80% as compared to heterologous protein production in BJ5464.

[18] In some embodiments the present disclosure provides a method for isolating a plasmid from a yeast cell, the method comprising: isolating total DNA from a yeast cell that includes a plasmid; incubating the DNA with an exonuclease such that the exonuclease degrades substantially all of the linear DNA in the isolated total DNA from the yeast cell; optionally inactivating the exonuclease; and recovering the plasmid DNA. In some cases, the isolated plasmid DNA is of sufficient purity for use in a sequencing reaction. In some cases, the plasmid DNA is further prepared for a sequencing reaction.

[19] In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 6 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 7 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 8 and a pharmaceutically acceptable excipient.

In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 9 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 10 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 11 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 12 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 13 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 14 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 15 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides pharmaceutical composition comprising Compound 16 and a pharmaceutically acceptable excipient. In some embodiments, the present disclosure provides method of producing Compound 3, the method comprising: providing a vector or vectors comprising the coding sequences of **SEQ ID NOs: 200–206**; transforming a host cell with the vector or vectors; incubating the host cell in culture media under conditions suitable for the expression of the coding sequences; and isolating the compound produced by the host cell.

INCORPORATION BY REFERENCE

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

BRIEF DESCRIPTION OF THE DRAWINGS

[21] The written disclosure herein describes illustrative embodiments that are non-limiting and non-exhaustive. Reference is made to certain of such illustrative embodiments that are depicted in the figures, in which:

[22] **Figure 1A** illustrates strategies that may be used to obtain a secondary metabolite from a fungal strain given different properties of the fungal strain.

- [23] **Figure 1B** illustrates examples of characterized gene clusters which produce known chemicals and a novel gene cluster with a product.
- [24] **Figure 2** illustrates a phylogenetic analysis of enzymes that produce secondary metabolites.
- [25] **Figure 3** illustrates self-resistance mechanisms for potentially toxic secondary metabolites.
- [26] **Figure 4** illustrates a gene cluster that produces lovastatin.
- [27] **Figure 5A** illustrates an exemplary process to extract and utilize compound products in accordance with an embodiment of the invention.
- [28] **Figure 5B** illustrates an exemplary process to produce and extract heterologous, biosynthetic compound products in accordance with an embodiment of the invention.
- [29] **Figure 5C** illustrates an example production pipeline for producing secondary metabolites from gene clusters.
- [30] **Figure 5D** illustrates an example work flow for producing secondary metabolites from gene clusters.
- [31] **Figure 6A** illustrates a yeast phase chart displaying yeast cell concentration in relation to time to provide reference for various embodiments of the disclosure.
- [32] **Figure 6B** illustrates a yeast phase chart displaying glucose or dextrose concentration in relation to time to provide reference for various embodiments of the disclosure.
- [33] **Figure 6C** illustrates a yeast phase chart displaying ethanol or glycerol concentration in relation to time to provide reference for various embodiments of the disclosure.
- [34] **Figure 7A** illustrates a DNA vector having a production-phase promoter in accordance with an embodiment of the disclosure.
- [35] **Figure 7B** illustrates a DNA vector having multiple production-phase promoters in accordance with an embodiment of the disclosure.
- [36] **Figure 8A** illustrates a DNA expression vector having a production-phase promoter within an expression cassette in accordance with an embodiment of the disclosure.
- [37] **Figure 8B** illustrates a DNA expression vector having multiple production-phase promoters, each within an expression cassette in accordance with an embodiment of the disclosure.
- [38] **Figure 9** illustrates a method to construct and utilize production-phase promoter DNA vectors in accordance with various embodiments of the disclosure.

- [39] **Figure 10A** illustrates an overview of an approach involving yeast homologous recombination assembly.
- [40] **Figure 10B** illustrates the homologous recombination in yeast of the parts from FIG. 10A.
- [41] **Figure 10C** illustrates the plasmid which results from the parts of **FIG. 10A**, homologously recombined as in **FIG. 10B**.
- [42] **Figure 10D** illustrates improved sequencing results obtained via disclosed methods.
- [43] **Figure 10E** illustrates assembly of plasmid DNA from up to 14 individual fragments.
- [44] **Figure 11A** illustrates improved assembly efficiency in a background lacking the DNL4 ligase.
- [45] **Figure 11B** illustrates equivalent sequencing efficiencies using DNA prepared from both colonies (red) and liquid cultures (blue) for four test assemblies.
- [46] **Figure 11C** illustrates sequencing efficiencies observed using both standard and modified NexteraXT library preparation methods.
- [47] **Figure 11D** illustrates a workflow for sequencing plasmids from yeast via a step of transforming the plasmids into *E. coli*.
- [48] **Figure 11E** illustrates a workflow for sequencing plasmids from yeast in accordance with methods described herein.
- [49] **Figure 12** illustrates repaired SNPs in yeast strain DHY674 relative to BY4741.
- [50] **Figure 13** illustrates DHY213, BJ5464, and X303 (a W303 derivative) grown on glucose (fermentation) and ethanol/glycerol (respiration) media.
- [51] **Figure 14A** illustrates relative growth rates of strains described here in YPD culture. The dotted line denotes the diauxic shift (the point at which the culture exhausts all glucose and transitions from fermentation to respiration). The DHY derived strain JHY692 shows significantly improved growth in the respiration phase of the culture.
- [52] **Figure 14B** illustrates expression of eGFP driven by the PADH2 promoter in the strains from **Figure 14A**.
- [53] **Figure 14C** illustrates expression of eGFP driven by the PPCK1 promoter in the strains from **Figure 14A**.
- [54] **Figure 15A** illustrates an example gene cluster.
- [55] **Figure 15B** illustrates a polyketide produced by a 5-gene gene cluster.

[56] **Figure 16** is a heat map graphic generated in accordance with various embodiments of the disclosure with data of expression of enhanced-green fluorescent protein driven by various *S. cerevisiae* promoters.

[57] **Figure 17** is a data graph of enhanced-green fluorescent protein expression driven by various *S. cerevisiae* promoters.

[58] **Figure 18** illustrates fluorescence intensity of 10^5 cells expressing enhanced-green fluorescent protein driven by various promoters.

[59] **Figure 19** illustrates a phylogenetic tree of *Saccharomyces sensu stricto* subgenus.

[60] **Figure 20** illustrates a multiple sequence alignment of various *Saccharomyces sensu stricto* species' upstream activating sequences in ADH2 promoters.

[61] **Figure 21** illustrates homology between various *Saccharomyces sensu stricto* species' ADH2 promoters.

[62] **Figure 22** is a heat map graphic generated in accordance with various embodiments of the disclosure with data of expression of enhanced-green fluorescent protein driven by various *S. sensu stricto* ADH2 promoters.

[63] **Figure 23** is a data graph of enhanced-green fluorescent protein expression driven by various *S. sensu stricto* ADH2 promoters.

[64] **Figure 24** illustrates four multi-gene expression vector constructs and a data graph of the resultant compound production, in accordance with an embodiment of the invention.

[65] **Figure 25** illustrates a biosynthetic process that produces the compound emindole SB via a fungal four-gene cluster.

[66] **Figure 26** is a data graph of the production results of two product compounds generated.

[67] **Figure 27A** illustrates two plasmid vector constructs in accordance with an embodiment of the disclosure.

[68] **Figure 27B** illustrates a further vector construct in a yeast cell in accordance with an embodiment of the disclosure.

[69] **Figure 28A** illustrates a phylogenetic analysis of further gene clusters. Abbreviations used may include Adenylation domain (A), a,b-hydrolase (a,b-h), ATP-binding cassette transporter (ABC), Acyl carrier protein (ACP), Alcohol dehydrogenase (ADH), Aldo-keto reductase (AK-red), Aminooxidase (AmOx), Aminotransferase (AmT), Arylsulfotransferase (ArST), Acyltransferase

domain (AT), C-methyltransferase (cMT), Terpene cyclase (Cyc), Dehydratase (DH), Domain of unknown function 4246 (DUF4246), Flavin adenine dinucleotide (FAD) binding protein (FAD), Iron dependent alcohol dehydrogenase (Fe-ADH), Flavin-dependent monooxygenase (FMO), Geranyl-geranyl pyrophosphate synthase (GGPPS), Glycosidase (glycos.), Glucose-methanol-choline oxidoreductase (GMC), Halogenase (Halo), Highly reducing polyketide synthase (HR-PKS), Hypothetical protein (Hyp), Indole-3-acetic acid-amido synthetase (IAS), Ketosynthase domain (KS), metallo-B-lactamase (mBla), Mitochondrial phosphate carrier protein (MCP), Major facilitator superfamily transporter (MFS), Metallohydrolase (MH), Methyl transferase (MT), Nicotine adenine dinucleotide dependent dehydrogenase (NAD-DH), Nicotine adenine dinucleotide phosphate (NADP) dependent reductase (NADP-R), N-methyltransferase (nMT), Non reducing polyketide synthase (NR-PKS), O-succinylhomoserine sulfhydrylase (O-suc-SH), O-methyltransferase (oMT), Oxidoreductase (OxR), Cytochrome p450 (p450), Dipeptidyl peptidase (Pep), Prenyl transferase (PrT), Product template domain (PT), Riboflavin biosynthesis protein RibD (RibD), RNA helicase (RNAh), starter unit:ACP transacylase domain (SAT), Short-chain dehydrogenase (SDH), Short-chain dehydrogenase/reductase (SDR), Serine hydrolase (SH), Sugar transport protein (ST), Thiolation domain (T), Terminal domain (TD), Thioesterase domain (TE), Transcription factor (TF), and UbiA-type terpene cyclase (UTC).

[70] **Figure 28B** illustrates various gene clusters and biosynthetic compound products in accordance with various embodiments of the invention.

[71] **Figure 28C** illustrates various gene clusters and biosynthetic compound products in accordance with various embodiments of the invention.

[72] **Figure 28D** illustrates various gene clusters and biosynthetic compound products in accordance with various embodiments of the invention.

[73] **Figure 28E** illustrates various gene clusters and biosynthetic compound products in accordance with various embodiments of the invention.

[74] **Figure 28F** illustrates various gene clusters and biosynthetic compound products in accordance with various embodiments of the invention.

[75] **Figure 29A** illustrates a phylogenetic analysis of gene clusters.

[76] **Figure 29B** illustrates correction of a gene cluster.

[77] **Figure 30A** illustrates schematics of PKS enzyme containing BGCs examined herein.

- [78] **Figure 30B** illustrates schematics of UTC containing BGCs examined herein.
- [79] **Figure 31** illustrates a volcano plot of all spectral features identified in the automated analysis of strains expressing PKS containing BGCs. All features determined to be specific to the BGC expressing strain were identified by comparison to a negative vector control.
- [80] **Figure 32** illustrates features produced by BGC PKS1.
- [81] **Figure 33** illustrates features produced by BGC PKS2.
- [82] **Figure 34** illustrates features produced by BGC PKS4.
- [83] **Figure 35** illustrates features produced by BGC PKS6.
- [84] **Figure 36** illustrates features produced by BGC PKS8.
- [85] **Figure 37** illustrates features produced by BGC PKS10.
- [86] **Figure 38** illustrates features produced by BGC PKS13.
- [87] **Figure 39** illustrates features produced by BGC PKS14.
- [88] **Figure 40** illustrates features produced by BGC PKS15.
- [89] **Figure 41** illustrates features produced by BGC PKS16.
- [90] **Figure 42** illustrates features produced by BGC PKS17.
- [91] **Figure 43** illustrates features produced by BGC PKS18.
- [92] **Figure 44** illustrates features produced by BGC PKS20.
- [93] **Figure 45** illustrates features produced by BGC PKS22.
- [94] **Figure 46** illustrates features produced by BGC PKS23.
- [95] **Figure 47** illustrates features produced by BGC PKS24.
- [96] **Figure 48** illustrates features produced by BGC PKS28.
- [97] **Figure 49** illustrates NMR data and structure of Compound 6.
- [98] **Figure 50** illustrates NMR data and structure of Compound 7.
- [99] **Figure 51** illustrates NMR data and structure of Compound 8.
- [100] **Figure 52** illustrates NMR data and structure of Compound 9.
- [101] **Figure 53** illustrates NMR data and structure of Compound 10.
- [102] **Figure 54** illustrates NMR data and structure of Compound 11.
- [103] **Figure 55** illustrates NMR data and structure of Compound 12.
- [104] **Figure 56** illustrates NMR data and structure of Compound 13.
- [105] **Figure 57** illustrates NMR data and structure of Compound 14.

[106] **Figure 58** illustrates NMR data and structure of Compound 15.

[107] **Figure 59** illustrates NMR data and structure of Compound 16.

DETAILED DESCRIPTION

[108] Systems and methods in accordance with various embodiments of the disclosure identify, refactor, and express biosynthetic gene clusters in host organisms to produce secondary metabolites. Systems and methods in accordance with various embodiments of the disclosure utilize host organisms to produce secondary metabolites. Using host organisms allows for production of secondary metabolites from biosynthetic gene clusters regardless of whether the native host cell can be cultured, or whether the cluster is expressed in the native host, see **Figure 1A**. In some cases, the secondary metabolites may bind to one or more specific proteins. In some embodiments, the host organism ordinarily does not produce the secondary metabolite. The host organism obtains the ability to produce the secondary metabolite due to the introduction of a biosynthetic gene cluster identified using a cluster identification process performed in accordance with an embodiment of the disclosure. In one embodiment, a cluster identification process identifies a biosynthetic gene cluster that possesses characteristics suggesting that it is responsible for producing a secondary metabolite with novel chemistry. In another embodiment, a cluster identification process (discussed further below) identifies a biosynthetic gene cluster that possesses characteristics suggesting that it is responsible for producing a secondary metabolite that binds to a specific protein of interest. In some embodiments this disclosure provides the inclusion of a biosynthetic gene cluster identified using a cluster identification process, in accordance with various embodiments of the disclosure, within a host organism enabling the host organism to express a secondary metabolite. In some cases, the secondary metabolite produced in the host cell may be identical to a secondary metabolite that is naturally produced by the organism in which the biosynthetic gene cluster was originally identified. In some cases, the secondary metabolite produced by the host cell is an analog of a secondary metabolite produced by the organism from which the cluster was identified. In other cases, the secondary metabolite produced in the host cell may be structurally distinct from the secondary metabolite produced by the cluster in the originating species. Differences in the product produced may arise from differences in expression and localization of the coding sequences from the cluster.

Additionally coding sequences, or expression products thereof, from the cluster may interact with other coding sequences, or expression products thereof, not contained in the cluster. Host cell produced secondary metabolites which are distinct from those produced in the originating species of the cluster may be termed non-naturally occurring secondary metabolites or non-natural secondary metabolites. The secondary metabolite produced by the host organism can be isolated and then can be used, for example, in a treatment. In some cases, the secondary metabolites may be used in a treatment of a disease or disorder which involves aberrant activity of a specific protein. This disclosure also describes a panel of sesquiterpenoid and polyketide products.

Cluster identification

[109] Cluster identification processes, in accordance with many embodiments of the disclosure, utilize specific properties of a biosynthetic gene cluster to identify gene clusters of interest from sequence data. Sequence data used with the methods of this disclosure may comprise genomic sequence data, transcriptome sequence data, or other sequence data. In some cases, sequence data may be generated by sequencing a DNA sample obtained from an environmental sample. In other cases, sequence data may be obtained from publically available genome sequence libraries, or may be purchased. Genome sequences may be derived from any organism. In some cases, genome sequences may be derived from a fungal, bacterial, archaeal or plant species.

[110] In some cases, the genome sequences may be derived from fungi. In some cases, the genome sequences may be derived from fungi which are poorly characterized or difficult to culture. In some cases, the genome sequences may be derived from fungi which are well characterized, or partially characterized. Examples of types of fungi from which sequences may be derived include fungi from one of the following Phyla: Basidiomycota, Ascomycota, Neocallimastigomycota, Blastocladiomycota, Glomeromycota, Chytridiomycota and Microsporidia. Examples of fungal species which may be the source of sequence data include, but are not limited to: *Aspergillus tubingensis*, *Hypomyces subiculosus*, *Coniothyrium sporulosum*, *Acremonium Sp. KY4917*, *Aspergillus niger*, *Thielavia terrestris*, *Trichoderma virens*, *Pseudogymnoascus pannorum*, *Scedosporium apiospermum*, *Metarhizium anisopliae*, *Cochliobolus heterostrophus*, *Verruconis gallopava*, *Moniliophthora roreri*, *Punctularia strigosozonata*, *Hydnomerulius pinastri*, *Arthroderma gypseum*, *Setosphaeria turcica*, *Pyrenophora teres*, *Cladophialophora yegresit*,

Talaromyces cellulolyticus, *Endocarpon pusillum*, *Hypholoma sublateritium*, *Ceriporiopsis subvermispora*, *Botryotonia cinerea*, *Formitiporia mediterranea*, *Heterobasidion annosum*, *Ge/atoporia subvermispora*, *Dichomitus squalens*, *Pleurotus ostreatus*, *Schizophyllum commune*, *Stereum hirsutum*, *Sternum hirsutum*, and *Dacryopinax primogenitus*.

[111] Provided in **Figure 1B** is a flow chart showing a process embodiment that can be implemented using computer systems for identifying and ranking BSGs that are likely to produce secondary metabolites capable for anthropogenic use (e.g., medicinal). As shown, Process 1000 can begin by obtaining genetic sequence data from a biological source or sequence database (1001). In some cases, the sequence data may be derived from single cell sequencing of a fungal cell of unknown species. For example an environmental sample may contain many different cells which may be separated and sequenced. In some cases, the sequence data is obtained from a publically available genetic data library, or may be purchased. Often the genetic sequence data is a genomic sequence of an organism, however, any genetic data sequence, including partial genome sequence data, may be used.

[112] Process 1000 also identifies biosynthetic gene clusters (1003). Clusters may be identified using bioinformatics methods to scan genome sequences. Characteristics of a gene cluster may include a grouping of two or more genes within about 5 kb, 10 kb, 15 kb, 20 kb, 25 kb, 30 kb, 35 kb, 40kb or 45 kb of each other. Genes may be bioinformatically identified by the presence of known promoter sequences, transcription initiation sequences, or homology to known genes or gene features. The term homology as used herein refers to sequences with high sequence identity, for example a sequence identity of at least about 50%, 60%, 70%, 80%, 90%, 95%, 97%, 98%, 99% or more than 99%. Sequence identity may be determined using alignment tools such as Basic Local Alignment Search Tool (BLAST), available via the National Center for Biotechnology Information (NCBI), to determine areas of conserved sequence. Biosynthetic gene clusters may be identified using a bioinformatics tool, such as ClustScan, SMURF, CLUSEAN, and/or antiSMASH. Biosynthetic gene clusters typically contain one or more core biosynthetic genes and one or more tailoring genes. Clusters which contain the same, or highly similar, core enzymes with different tailoring genes may produce quite different compounds, as seen in **Figure 1C**. Expressing a subset of genes from a cluster may also result in production of a different compound compared with the compound produced by expression of all the genes in the cluster.

[113] Process 1000 also scores and/or ranks gene clusters utilizing various factors (1005). In some cases, scores and rankings are based on the type of secondary metabolite to be produced. In other cases, scores and rankings are based on a protein or domain existing within the cluster. In even more cases, the level of homology of a particular protein within the cluster to a protein of interest is considered. It should be understood that many different factors can be used, as determined by the application and use of the biosynthetic gene cluster data.

[114] An embodiment of a process for identifying biosynthetic gene clusters using computer systems is provided in **Figure 1D**. Process 2000 can begin with obtaining genetic sequence data of an organism having gene clusters (2001). In many cases, the genetic sequence data is a genomic sequence of an organism. In other cases, the genetic data sequence is a partial genome sequence data. In addition to genetic sequence data, Process 2000 also obtains target sequence data keying to biosynthetic gene clusters of interest (2003). The target sequence is any sequence the user wishes to define and identify clusters. In some cases, the target sequence is a particular protein domain of interest. In other cases, the target sequence is a particular protein, protein homolog or protein class.

[115] In some embodiments, clusters are scanned for the presence of genes encoding proteins known to be involved in biosynthetic pathways. Key proteins involved in biosynthetic pathways include terpene synthases, polyketide synthases (PKSs, both highly-reducing and non-reducing), non-ribosomal peptide synthetases, UbiA-type terpene cyclases (UTCs), polyketide synthase non-ribosomal peptide synthetase hybrids, and dimethyl allyl transferases (see **Figure 2**).

[116] Process 2000 also identifies the target sequence within genetic sequence data using homologous alignment scores (2005). Using an appropriate application, the sequence of the target is used to align to the genetic sequence data, looking for a threshold of homology. In some cases, a positive homologous event occurs when the target sequences aligns with the a portion of the genetic sequence having homology of at least about 50%, 60%, 70%, 80%, 90%, 95%, 97%, 98%, 99% or more than 99%. Sequence homology may be determined using any alignment tool, such as, for example BLAST.

[117] Using the homologous alignment scores, candidate biosynthetic gene clusters may be identified in the region surrounding the homologous event (2007). In many cases, the proximal upstream and downstream genes are examined to determine and define a gene cluster. In some of these cases, 5, 6, 7, 8, 9, 10 or more proximal genes in either direction are examined. In addition or

alternatively, a gene cluster may be defined by genetic distance of the homologous event. Gene clusters may also be defined using a bioinformatics tool, such as ClustScan, SMURF, CLUSEAN, and/or antiSMASH. Once identified and defined, clusters may be stored as data and/or reported via an output interface (2009).

[118] An embodiment for ranking biosynthetic gene clusters using computer systems is provided in **Figure 1E**. As shown, Process 3000 can begin with obtaining genet sequence data of multiple biosynthetic gene clusters. In this process, the clusters obtained each have a homolog protein of interest. The homolog of interest depends on the user's desired result. In many cases, the homolog of interest has a human ortholog that is known to be involved in a human condition, disorder, or disease. In some of these cases, the human ortholog is known to have mutations, either congenital or somatic, that lead to a condition, disorder, or disease. In some other of these cases, the human ortholog is involved in biological pathways that are involved in a condition, disorder, or disease. In other cases, the homolog of interest has an ortholog in an infectious species, including, but not limited to bacterial, fungal, protozoan species. In many of these cases, the ortholog in the infectious species is essential to the vitality of the organisms of the species. In some other of these cases, the ortholog in the infectious species is involved in producing a toxin. Furthermore, in many cases, the user has a desired result to prioritize clusters that will produce a secondary metabolite that may target the of human or infectious species ortholog.

[119] Identifying a gene cluster that may produce a secondary metabolite that binds a target protein may involve multiple different steps. In some cases, such a gene cluster may be identified by the presence of one or more genes encoding a homolog of the target protein, within or adjacent to the cluster, as determined by a homology search (for example, using the tblastn algorithm, with a maximum score granted when one homolog is found). In some cases, a gene cluster may be identified by the confidence in homology of the target to a gene or genes in a cluster. For example, according to the tblastn algorithm, gene clusters containing a gene with an e value less than about 10^{-10} , 10^{-20} , 10^{-30} , 10^{-35} or 10^{-40} may be selected. Gene clusters may also be selected using a protein blast method such as blastp to compare a predicted protein sequence against either known protein sequence or other predicted protein sequence. For blast searches using a known or predicted protein sequence gene clusters containing a gene with an e value less than about 10^{-10} , 10^{-20} , 10^{-30} , 10^{-35} or 10^{-40} may be selected. Selected gene clusters may be prioritized with increasing priority scores for

lower e-values. In some cases, a gene cluster may be identified by the fraction of the homologous gene that meets a certain threshold of identity (for example, with an increasing score for more identity, and a lower bound threshold of at least about 99%, 95%, 80%, 85%, 75%, 70%, 65%, 60%, 50%, 45%, 40%, 35%, 30%, 25%, 20% or 15% coverage). In some cases, a gene cluster may be selected if it contains a gene which produces a protein with at least about 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97%, 98%, 99% or 100% homology to the target protein. In some cases, a protein that is encoded by the region that is included in, or positioned proximal to, a BGC may be identical to, or have at least about 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99% or about 100% homology to a target protein. In some cases, a gene cluster may be identified by the total number of genes homologous to the target protein present in the entire genome of the organism (for example, with a maximum score granted to cases with 2–4 homologs per genome). In some cases, a gene cluster may be identified by the homology of the gene in, or adjacent to, the cluster to the target protein (for example, using the blastx algorithm, with a maximum score granted when the gene in the biosynthetic gene cluster's closest homolog in the target protein's genome is the target protein itself). In some cases, a gene cluster may be identified by the phylogenetic relationship of the target protein to the gene in the cluster (for example, with an increasing score for homologs in the gene cluster that clade with the target protein, with confidence assigned by a bootstrap test or Bayesian inference of phylogeny, and a lower bound threshold defined as homologs in a phylogenetic context that appear in a clade with bootstrap value of 0.7 or Bayesian posterior probability of 0.8). In some cases, a gene cluster may be identified by the expected number of homologs of the target in or adjacent to the biosynthetic cluster (for example, with a greater score the lower the probability of a homolog of the target being present in or adjacent to a biosynthetic cluster of a certain size, given the number of total homologs in the genome, as determined by a permutation test). In some cases, a gene cluster may be identified by the likelihood that the target protein is essential for viability, growth, or other cellular processes in the native environment (for example, through evidence that deletion of homologs in related organisms (such as *S. cerevisiae*) render the organism inviable). In some cases, a gene cluster may be identified by one or more of the above methods, or by any two or more of the above methods.

[120] Utilizing identification steps, Process 3000 also scores each biosynthetic gene cluster based on factors indicative of secondary metabolite synthesis related to the homolog of interest (3003).

Accordingly, a score is constructed for each biosynthetic cluster based on one or more of the following:

- (a) the presence of one or more homologs of target orthologs within or adjacent to the cluster, as determined by a homology search (e.g., the BLAST algorithm);
- (b) the degree of homology of one or more target orthologs to genes in a cluster (e.g., as defined by the e-value);
- (c) the fraction of the homologous gene that meets a certain threshold (e.g., the number of homologous protein domains);
- (d) the total number of genes homologous to the target ortholog present in the entire genome of the organism;
- (e) the degree of homology of a gene in or adjacent to the cluster to the target ortholog
- (f) the number of homologs in the gene family (e.g., the number of homologs of a human target gene in the human genome); and/or
- (g) the expected number of homologs of the target ortholog in, or neighboring, a biosynthetic cluster (e.g., the probability of a homolog of the target being present in or adjacent to a biosynthetic cluster of a certain size, given the number of total homologs in the genome, and as determined from a permutation test)
- (h) the synteny of the gene cluster with related species (e.g., conservation of gene cluster)
- (i) the function class of the target ortholog
- (j) the presence of specific promoters adjacent to the homolog(s) within the gene cluster (e.g., identification of bidirection promoter upstream the homolog and biosynthetic gene)
- (k) the presence of specific regulatory elements in the biosynthetic cluster (e.g., the number of transcription factor binding sites shared among target orthologs and homologs/biosynthetic genes in the cluster)
- (l) the presence of homologs outside the cluster that are co-regulated with some or all the genes within the biosynthetic cluster
- (m) the presence of protein- and DNA-sequence derived features within the clusters that have successfully been shown to produce secondary metabolites. It should be understood that a particular user may desire to use one, some, or all the factors listed here, and/or other factors not listed. The factors utilized depend on the user's application and desired result.

[121] Process 3000 also has the option to calibrate the score to by referencing a set of “true positives” (i.e., cases where there is one or more known targets in or adjacent to a biosynthetic cluster that produces a small molecule targeting the target, such as the lovastatin BGC) (3005). The output of this process may be a ranked and/or scored list of biosynthetic clusters, which may be used to identify the clusters that will produce therapeutic small molecules targeting the products of a disease-related gene (3007). The ranked and/or scored list of clusters can be stored as data or reported via an output interface (3009).

[122] Turning now to **Figure 1F**, computer systems (4001) may be implemented on a single or multiple computing devices in accordance with some embodiments of the invention. Computer systems (4001) may be personal computers, laptop computers, and/or any other computing devices with sufficient processing power for the processes described herein. The computer systems (4001) include a processor (403), which may refer to one or more devices within the computing devices that can be configured to perform computations via machine readable instructions stored within a memory (4007) of the computer systems (4001). The processor may include one or more microprocessors (CPUs), one or more graphics processing units (GPUs), and/or one or more digital signal processors (DSPs). According to other embodiments of the invention, the computer system may be implemented on multiple computers.

[123] In a number of embodiments of the invention, the memory (4007) may contain a gene cluster identification and/ scoring application (4009) that performs all or a portion of various methods according to different embodiments of the invention described throughout the present application. As an example, processor (4003) may perform a gene cluster identification and/or scoring method similar to any of the processes described above with reference to Figures 1D and 1E, during which memory (4007) may be used to store various intermediate processing data such as the genetic sequence alignment data (e.g., BLASTn) (4009a), identification of key target sequences with gene clusters (4009b), identification of homolog(s) to a protein of interest (4009c), characterization of homologs (4009d), scores and/or ranks of gene clusters (4009e), and calibration of gene cluster scores (4009f).

[124] In some embodiments of the invention, computer systems (4001) may include an input/output interface (4005) that can be utilized to communicate with a variety of devices, including but not limited to other computing systems, a projector, and/or other display devices. As can be

readily appreciated, a variety of software architectures can be utilized to implement a computer system as appropriate to the requirements of specific applications in accordance with various embodiments of the invention.

[125] Although computer systems and processes for chimeric sequence unveiling and performing actions based thereon are described above with respect to Figure 1F, any of a variety of devices and processes for data associated with cluster identification and/or scoring as appropriate to the requirements of a specific application can be utilized in accordance with many embodiments of the invention.

[126] In some embodiments, gene sequences within a novel cluster, or a set of novel clusters, may be compared to gene sequences from known and characterized biosynthetic gene clusters. In some cases, phylogenetic comparisons may be carried out between gene sequences in a novel cluster and gene sequences from characterized gene clusters. As shown in **Figure 2**, of the many biosynthetic enzymes which have been identified from sequence data, only a small fraction have been characterized, suggesting potential for many novel chemistries. Phylogenetic analysis may be performed on the core biosynthetic gene or genes, or on one or more tailoring genes of the novel cluster. Preferred clusters may be clusters containing one or more gene sequences which do not share a close phylogenetic relationship with a sequence from a characterized gene cluster. In some cases, gene clusters may be ordered according to their phylogenetic relationship to characterized gene clusters, clusters with the most distant relationships may be preferred for further analysis.

[127] In several embodiments, a specific secondary metabolite is used as a weapon against another organism (i.e., is toxic to or inhibits the growth of specific type of organism). In many instances, the toxic secondary metabolite may also be toxic to the organism that produces the secondary metabolite. Accordingly, the producing organism may defend against self-harm in a number of ways including (but not limited to) pumping the secondary metabolite out of the cell, enzymatically negating the secondary metabolite, or producing an additional version of the protein targeted by the secondary metabolite that is less sensitive or insensitive to the secondary metabolite, (see **Figure 3**). In instances in which an organism produces an additional version of the target protein, the “protective” version of the gene that encodes the additional version of the target protein that is less sensitive to the secondary metabolite is often colocalized with the biosynthetic gene cluster, for example the HMGR gene in the lovastatin cluster shown in **Figure 4**. Although different to the gene

that produces the target protein, the protective version of the gene should produce a protein that maintains detectable homology to the target protein. In several embodiments, the cluster identification process takes advantage of this homology to identify those biosynthetic gene clusters that contain or are adjacent to a gene that encodes a protective homolog of the target protein. In a number of embodiments of the disclosure, the genetic sequences of multiple organisms are analyzed to detect biosynthetic gene clusters possessing this characteristic.

[128] A target protein may be any protein of interest which has a homolog in the genome sequence/s from which the gene clusters were obtained. In some cases, the target protein is an enzyme. In some cases, the target protein is a signaling protein. In some cases, the target protein is one which is required by the species of origin. For example, the target protein may be one which contributes to viability of growth of the cell, and deletion or inactivation of the target protein from the cell may have deleterious effects on the viability or growth of the cell. In some cases, the target protein may have a vertebrate or mammalian homolog. In some cases, the target protein has a human homolog. The human homolog may be a protein which is dysregulated in a disease. An example of a gene cluster which was identified using methods disclosed herein, and which comprises a homolog of the BRSK1 gene is shown in Figure 15A.

[129] In some embodiments, a biosynthetic cluster of interest which produces a secondary metabolite that interacts with a target protein may also produce a protein for inactivating the secondary metabolite or for secreting the secondary metabolite from the cell in which it is produced. A protein which inactivates the secondary metabolite may be omitted when designing an expression construct to express this cluster in a host cell. A protein involved in secreting the secondary may be included or omitted when designing an expression construct to express this cluster in a host cell. To identify such clusters several approaches may be used. For example a set of biosynthetic gene clusters may be identified from genome data and the identified clusters may be analyzed for the presence of enzymes with activities that may be useful for degradation of a secondary metabolite. In another example, a set of biosynthetic gene clusters may be analyzed for the presence of proteins involved in transport or secretion pathways. In another example a homology search may be identified across one or more genome sequences to find genes encoding proteins homologous to an enzyme known to degrade a toxic compound. Once such genes have been identified they may be

analyzed for proximity to a biosynthetic gene cluster. Proximity to a gene cluster may be defined as within about 50 kb, 40 kb, 30 kb, 20 kb, 10 kb, 5 kb, 1kb, or less than 1 kb.

[130] In some embodiments, a gene cluster which produces a secondary metabolite that may be toxic to the host cell may also contain signals to direct the production of the secondary metabolite to a specific cellular location. In some cases, the enzymes may be membrane tethered to the intracellular or extracellular side of a cell membrane, or may be secreted through a membrane to the intracellular or extracellular side (including the insides of organelles and vacuoles). For example, the enzymes of the cluster may contain membrane-targeting signals to target the enzymes to either the extracellular membrane or to an intracellular organelle or vacuole. In some cases, the enzymes may be targeted to the extracellular membrane in an orientation which results in the active region of the enzyme being inside of an organelle or vacuole. In some cases, the enzymes may be targeted to the extracellular membrane in an orientation which results in the active region of the enzyme being outside of the cell. Such clusters may be identified by analyzing the predicted proteins of the cluster for the presence of peptide targeting signals or of terminal sequences with homology to targeting signals.

[131] In some cases, a genome or genomes may be searched for gene clusters and the set of identified gene clusters may be searched for genes which are homologous to a target gene, or which produce proteins homologous to a target protein. In other cases, a genome or genomes may be searched for a gene or genes homologous to a target gene, and the identified genes may be analyzed for association with a gene cluster. In other cases, a genome or genomes may be searched for gene clusters and the set of identified gene clusters may be phylogenetically analyzed to determine relationships between the identified gene clusters and known, characterized, gene clusters. In yet other cases, a novel genome or genomes may be searched for sequences distantly homologous to a known biosynthetic enzyme, and the identified genes may be analyzed for association with a gene cluster.

[132] Specific secondary metabolites synthesized using methods in accordance with a number of embodiments of the disclosure and the proteins utilized to identify the biosynthetic gene clusters used to synthesize the secondary metabolites and targeted by the secondary metabolites are described herein. An example of a method which may be used to produce a secondary metabolite is shown in **Figure 5**. As shown in an exemplified embodiment process in **Figure 5A**, Process 100

generates and extracts a heterologous compound for use. Exemplary Process 100 can begin by searching genetic data of various organismal species for pathways or BGCs that produce a compound product (101). The organismal species used in this step can be any species. For example, fungal and bacterial species often contain multiple BGC pathways that are encoded in its DNA. Likewise, the genetic data to be searched can be any genetic data available or determinable by the user. Accordingly, on one end of the spectrum, the genetic data may be a fully annotated, publicly available genome of a well-studied species (e.g., *Penicillium notatum*). On the other end of the spectrum, the genetic data may be a publicly unavailable, partial genomic sequence of a newly discovered species incapable of anthropogenic cultivation, wherein the partial sequence is found to have a gene cluster that may produce a compound.

[133] Exemplary embodiment Process 100 may continue by using the genetic data to reconstruct the compound product pathway in an acceptable genetic expression system (103). Often, to reconstruct the compound pathway, the genetic data is used to create nucleic acid molecules (e.g., DNA) comprising the coding sequences of the pathway genes sufficient to produce the product in the acceptable genetic expression system. The nucleic sequences are to be transferred into the expression system. Expression systems are any organism capable of producing the heterologous compound by heterologous expression of the pathway genes. Typical expression systems include (but are not limited to) *E. coli* and *S. cerevisiae*.

[134] Once the compound product pathway is reconstructed and transferred within an expression system, the expression system produces the compound (105) in exemplary Process 100. Typically, production of the compound results from coordinated expression of the pathway genes in the expression system. The coordinated expression of the pathway genes results in the production of the enzymes primarily responsible for constructing the heterologous compound product.

[135] **Figure 5B** depicts another exemplary embodiment process. Exemplary Process 200 produces, extracts, and characterizes a biosynthetic compound derived from heterologous expression of a gene cluster. The process may begin with the identification and selection of a gene cluster with an identifiable trait that is indicative of compound production (201). Several indicative processes to select gene clusters exist, including several computer-implemented programs. One such program is antiSMASH2.0, which is platform for mining BGC clusters for production of secondary metabolites that searches for core structures to identify putative BGCs (K. Blin, et al. Nucleic Acids

Res. 41:W204-12, 2013, the disclosure of which is incorporated herein by reference in its entirety). Another such method is described herein which utilizes homolog sequences within a proximal region in the genome to identify putative BGCs. It should be noted that many other methodologies could be used to select BGCs.

[136] Once a gene cluster has been selected, Process 200 continues by appropriating nucleic acid molecules with the coding sequences of the various genes within the cluster (203 in **Figure 5B**). Typically, the nucleic molecules are DNA, but other nucleic molecules (e.g., RNA) can be used for certain applications. When extracting gene sequence data from the host organism, it is usual to remove the non-translated portions (e.g., UTRs, introns) from the gene, leaving only the coding sequence, but the non-translated portions may also be used, especially if they provide a beneficial characteristic. Appropriation of the nucleic molecules can be performed by many different methods including (but not limited to) direct extraction from the host, chemical synthesis, and/or cDNA generation methods (e.g., reverse transcription of host RNA). Regardless of the method used, the resulting nucleic acid molecules may be available to build into expression vectors for heterologous expression.

[137] Exemplary Process 200 utilizes the appropriated nucleic acid molecules to assemble expression vectors for expression in an appropriate organismal expression system (e.g., *E. coli*, *S. cerevisiae*). Expression vectors are nucleic acid molecules that have the necessary components to express a heterologous gene in the expression system. Common expression vectors are plasmid DNA and viral vectors, but also include kits of DNA molecules that can be joined together to form a longer DNA molecule by a recombination methodology (e.g., yeast homologous recombination (YHR)).

[138] To express a heterologous gene from an expression vector, an expression cassette is may be used, which comprises the sequences of an appropriate promoter and an appropriate terminator along with the heterologous gene sequence. The promoter is typically located upstream of the heterologous gene and can regulate the gene's expression. Many different types of promoters can be used. The selection of the appropriate promoter depends on the application and expression profile desired. For example, in the *S. cerevisiae* expression system, production-phase promoters may express heterologous genes only in the production-phase of the yeast culture's life cycle, which may have desirable properties. For more description of production-phase promoters, please refer to the related

U.S. Patent Application No. 15/469,452 ("Inducible Production-Phase Promoters For Coordinated Heterologous Expression in Yeast"), the disclosure of which is incorporated herein by reference in its entirety. However, it should be understood, that constitutive promoters and other response-driven promoters could be used within the system.

[139] The sequences of promoters to be used in an expression vector can be derived from various sources. *E. coli* expression systems often use the T7 promoter derived from the T7 bacteriophage because the promoter reliably produces high expression in *E. coli*. Endogenous promoter sequences (e.g., the lac operon in *E. coli*) are expected to perform well within the organismal expression system.

[140] Expression vectors often have other sequences that benefit duplication, selection and stability of the vector within the organismal expression system, in addition to the expression cassette. In several instances, some of these sequences are necessary for maintenance in the expression system. For example, plasmid vectors within an *E. coli* or *S. cerevisiae* host require a host origin of replication and a selectable marker. The origin of replication signals the host expression system to replicate the plasmid vector in order to produce more copies of the plasmid as the host cells duplicate and divide. The selectable marker ensures that only the host cells that contain the vector continue to survive and propagate. Accordingly, these sequences may be necessary for viable heterologous expression.

[141] Once the expression vector is assembled, the heterologous BGC genes are expressed using the organismal expression system (207). Accordingly, the expression vector is to exist within the organismal host such that the host will express the heterologous BGC genes to produce the encoded enzymes. The enzymes produce a biosynthetic compound. This compound is be extracted from the expression system (209).

[142] Extracted heterologous biosynthetic compounds can be characterized to determine their various structures and conformations. Some resultant products may have a solitary structure and conformation while other products will have several different structures with multiple conformations. The various structures and conformations can be determined using mass spectrometry, chromatography, and/or other methods.

[143] There are numerous classes of biosynthetic compounds. For example, polyketides and terpenes are a class of compounds derived from various organismal species. Many novel

biosynthetic compounds are likely to have beneficial properties, as a multitude of biosynthetic compounds have been found to be useful in several industries.

[144] Illustrated in **Figure 5C** is an exemplary pipeline to produce heterologous, biosynthetic compounds. Exemplary Pipeline 300 takes advantage of a yeast expression system to reproduce a fungal BGC in order to produce a heterologous compound product.

[145] Pipeline 300 begins with selection of a biosynthetic gene cluster (301). Depicted, as an example, is a phylogenetic tree of numerous fungal BGCs. Using the phylogenetic data, a BGC having a desired trait is selected. The coding sequences of the various BGC genes are then used to chemically synthesize DNA molecules (303). The synthesized BGC DNA molecules are then to be assembled into a heterologous expression construct (305). In this example, the DNA molecules are assembled by yeast homologous recombination. Accordingly, the synthesized DNA molecules have overlapping homologous sequences that the yeast will use to recombine the various DNA molecules into a plasmid DNA vector.

[146] Pipeline 300 then utilizes the assembled expression vectors to maintain and express the BGC in the yeast (307). The expression of the various heterologous genes results in expression of a number of heterologous enzymes that then can produce the heterologous biosynthetic compounds. Once a sufficient titer of compound is produced, it can be characterized to determine its structures and properties (309).

[147] Briefly, the method comprises gene cluster selection as discussed herein, synthesis of coding sequence, promoters and terminators, assembly of the cluster coding sequence, expression in a fungal host, and isolation and characterization of compounds produced. An example of a gene cluster identified using the methods herein, and the compound created by expression of the identified genes in yeast, is shown in **Figure 15B**.

Cluster engineering

[148] Once a gene cluster is identified according to the methods described herein, the cluster may be prepared for expression in a heterologous host cell. Editing of a putative gene cluster may involve steps such as: removal of introns, replacement of promoters, replacement of terminators, gene shuffling, and codon optimization. For example, if a gene cluster is to be expressed in *S.*

cerevisiae then coding sequences from the gene cluster may be codon optimized for *S. cerevisiae*, and operably linked to *S. cerevisiae* promoters and terminators.

[149] The gene cluster editing may rely on automatic annotation of expressed sequences, introns, and exons, or on manual inspection of the cluster. The gene cluster editing may be done *in silico* using the sequence data, or may be done *in vitro* or *in vivo* using the DNA sequence in a suitable vector such as a cloning vector. In some cases an initial edited gene cluster may not produce a product and reanalysis of the predicted coding sequences, and introns of the same, may reveal errors in the predicted transcription start sites, transcription termination sites and/or splice sites.

[150] In an embodiment, this disclosure provides sequences (**SEQ ID NOs: 67-483**) of cryptic BGCs which encode various products. These BGCs may also be reengineered to provide the coding sequences without the endogenous regulatory sequences. The coding sequences from these clusters may be isolated and cloned into one or more expression vectors for expression in a model host system. The expression vectors may be plasmids, viruses, linear DNA, bacterial artificial chromosomes or yeast artificial chromosomes. The expression vectors may be designed to integrate into the host genome, or to not integrate. In some cases, the expression vector may be a high copy number plasmid.

Promoters

[151] Expression of a refactored gene cluster in a host organism may require coordinated expression of several different coding sequences. In some cases, the expression of multiple different coding sequences in a host organism may require the use of multiple different promoters suitable for that organism. This disclosure provides methods for discovering multiple promoters with similar activities and expression patterns but with differing DNA sequence. Such methods may involve use of closely related species, such as different species of *Saccharomyces*. *Saccharomyces* (S.) is a genus of fungi composed of different yeast species. The genus can be divided into two further subgenera: *S. sensu stricto* and *S. sensu lato*. The former have relatively similar characteristics, including the ability to interbreed, exhibiting uniform karyotype of sixteen chromosomes, and their use in the fermentation industry. The later are more diverse and heterogeneous. Of particular importance is the *S. cerevisiae* species within the *S. sensu stricto* subgenus, which is a popular model organism used for genetic research.

[152] The yeast *S. cerevisiae* is a powerful host for the heterologous expression of biosynthetic systems, including production of biofuels, commodity chemicals, and small molecule drugs. The yeast's genetic tractability, ease of culture at both small and large scale, and a suite of well-characterized genetic tools make it a desirable system for heterologous expression. Occasionally, production systems require coordinated expression of two or more heterologous genes. Coordinated expression systems in bacteria (e.g., *E. coli*) has long exploited the operon structure of bacterial gene clusters (e.g., lac operon), allowing a single promoter to control the expression of multiple genes.

[153] The construction of synthetic operons therefore allows a single inducible promoter to control the timing and strength of expression of an entire synthetic system. In yeast, many heterologous-expression systems do not rely on the operon system, but instead rely on a one-promoter, one-gene paradigm. Accordingly, multi-gene heterologous expression is generally performed using multiple expression cassettes with a well-characterized promoter and terminator, each on a single expression vector (e.g., plasmid DNA) (See D. Mumberg, R. Muller, and M. Funk *Gene* 156:119-22, 1995). With traditional restriction-ligation cloning, it is also possible to recycle a promoter on a single plasmid by the serial cloning of multiple genes (M. C. Tang, et al., *J Am Chem Soc* 137:13724-27, 1995).

[154] Turning now to the drawings and data, disclosed embodiments are generally directed to systems and constructs of heterologous expression during the production phase of yeast. In many of these embodiments, the expression system involves coordinated expression of multiple heterologous genes. More embodiments are directed to production-phase promoter systems having promoters that are inducible upon an event in the yeast's growth or by the nutrients and supplements provided to the yeast. Specifically, a number of embodiments are directed to the promoters that are capable of being repressed in the presence of glucose and/or dextrose. In more embodiments, the promoters are capable of being induced in the presence of glycerol and/or ethanol. In additional embodiments, at least one production-phase promoter exists within an exogenous DNA vector, such as (but not limited to), for example, a shuttle vector, cloning vector, and/or expression vector. Embodiments are also directed to the use of expression vectors for the expression of heterologous genes in a yeast expression system.

[155] Controlled gene expression is desirable in heterologous expression systems. For example, it would be desirable to express heterologous genes for production during a longer stable phase.

Accordingly, decoupling the anaerobic growth and aerobic production phases of a culture allows the yeast to grow to high density prior to introducing the metabolic stress of expressing unnaturally high amounts of heterologous protein. In accordance with many embodiments, the anaerobic growth phase is defined by the yeast culture's energy metabolism in which the yeast cells predominantly catabolize fermentable carbon sources (e.g., glucose and/or dextrose), and a high growth rate (i.e., short doubling-time). In contrast, and in accordance with several embodiments, the aerobic production phase is defined by the yeast culture's energy metabolism in which the yeast cells predominantly catabolize nonfermentable carbon sources (e.g., ethanol and/or glycerol), and a steady growth rate (i.e., long doubling-time). Accordingly, each yeast cell's energy metabolism can be predominantly in aerobic or nonaerobic phase, and dependent on the local concentration of the carbon source.

[156] **Figure 6A** depicts the phases of a yeast culture when provided a fermentable sugar, such as glucose or dextrose sugar, at a concentration of around 2–4% as its main carbon source. Initially, a yeast culture will predominantly catabolize the fermentable sugar, which correlates with an exponential growth with very high doubling rates. The growth phase typically lasts approximately 4–10 hours. During this phase, the catabolism of the fermentable sources results in the production of ethanol and glycerol.

[157] Once glucose becomes scarce, the growth of a yeast culture passes a diauxic shift and begins to predominantly catabolize nonfermentable carbon sources (e.g., ethanol and/or glycerol) (**Figure 6B**). The predominant catabolism of nonfermentable carbon source correlates with a longer and more stable production phase that can last for several days, or even weeks in an industrial-like setting (**Figure 6A**). During the production phase, yeast cultures reach and maintain a high concentration, but have a much lower doubling time (**Figure 6A**). Due to the decrease in doubling rate, yeast cultures no longer expend a great amount of energy and resources on rapid growth and thus can reallocate that energy and those resources to other biological activities, including heterologous expression. Accordingly, it is hypothesized that limiting the transcription of heterologous genes to the production phase would allow a yeast culture to reach a high, healthy confluency that would in turn allow better heterologous protein expression and biosynthetic production.

[158] In yeast, transcriptional regulation can be achieved in several ways, including inducement by chemical substrates (e.g., copper or methionine), the tetON/OFF system, and promoters engineered to bind unnatural hybrid transcription factors. Perhaps the most commonly employed inducible promoters are the promoters controlled by the endogenous GAL4 transcription factor. GAL4 promoters are strongly repressed in glucose, and upon switching to galactose as a carbon source, strong induction of transcription is observed (M. Johnston and R. W. Davis, *Mol. Cell Biol.* 4:1440-48, 1984). While this system leads to high-level transcription, only four galactose-responsive promoters are known, and galactose is both a more expensive and a less efficient carbon source as compared to glucose (S. Ostergaard, et al., *Biotechnol. Bioeng.* 68:252-59, 2000). Other carbon-source dependent promoters have also been used for heterologous gene expression. The *S. cerevisiae* ADH2 gene exhibits significant derepression upon depletion of glucose as well as strong induction by either glycerol or ethanol (K. M. Lee & N. A. DeSilva *Yeast.* 22:431-40, 2005). Once induced, genes driven by the ADH2 promoter (pADH2) display expression levels equivalent to those driven by highly expressed constitutive counterparts. This induction profile was found to work in heterologous expression studies, as the system auto-induces upon glucose depletion in the late stages of fermentative growth after cells have undergone diauxic shift. The ADH2 promoter has been used extensively for yeast heterologous expression studies, resulting in high-level expression of several heterologous biosynthetic proteins (For example, see C. D. Reeves, et al., *Appl. Environ. Microbiol.* 74:5121-29, 2008).

[159] As shown in **Figure 6C**, the concentration of ethanol and glycerol increases as glucose and dextrose sugar decreases, due to anaerobic glycolysis (i.e., breaking down the fermentable sugar) and subsequent fermentation (i.e., converting the broken-down glucose into alcohol) and glycerol biosynthesis (i.e., converting the broken-down glucose into glycerol). Upon fermentable sugar depletion, yeast cultures undergo a diauxic shift and begin to use ethanol and glycerol as a carbon source instead of glucose. A diauxic shift, as understood in the art, is defined as a point in time when an organism switches from primarily consumption of one source for energy, to primarily another source. This shift typically elicits significant changes to a yeast culture's gene-expression pattern. Accordingly, it is hypothesized that higher concentrations of ethanol, (e.g., ~2–4%) and or glycerol (e.g., ~2%) could be used to stimulate promoters that either directly or indirectly respond to these concentrations (See Figures 6A and 6C).

[160] Various disclosed embodiments are based on the discovery of inducible promoters that can be used for the coordinated expression of multiple genes (e.g., gene cluster pathway) in *Saccharomyces* yeast. Described below are sets of inducible promoters from *S. cerevisiae* and related species that are inactive during anaerobic growth, activating transcription only after a diauxic shift when glucose is near-depleted and the yeast cells are respiring (i.e., the production phase). As portrayed in various embodiments, various production-phase promoters are auto-inducing and allow automatic decoupling of the growth and production phases of a culture and thus initiate heterologous expression without the need for exogenous inducers. It should be noted, however, that many embodiments include production-phase promoters that are also inducible in the presence of nonfermentable carbon-sources (e.g., ethanol and/or glycerol) supplied to the yeast. As such, multiple embodiments employ recombinant production-phase promoters that act much like constitutive promoters when the host yeast cultures are constantly maintained in ethanol and/or glycerol-containing media.

[161] Once activated, the strength of various production-phase promoters can vary as much as 50-fold. The strongest production-phase promoters stimulate heterologous expression greater than that observed from strong constitutive promoters. The production-phase promoters could be employed in many different applications in which high expression of multiple genes is beneficial. Accordingly, the promoters can be used, for example, in multiple subunit protein production or for the production of biosynthetic compounds that are produced by multiple proteins within a pathway. Discussed in an exemplary embodiment below, some embodiments are used to express multiple proteins involved in production of indole diterpene compound product. When compared to constitutive promoters, the production-phase promoters produced greater than a 2-fold increase in titer of the exemplary diterpene compounds. In other exemplary embodiments, it was found that the production-phase promoter system outperformed constitutive promoters by over 80-fold. Thus, these promoters can enable heterologous expression of biosynthetic systems in yeast.

[162] The practice of several embodiments will employ, unless otherwise indicated, conventional methods of chemistry, biochemistry, and molecular biology and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., A.L. Lehninger, *Biochemistry* (Worth Publishers, Inc., 30 current addition); Sambrook, et al., *Molecular*

Cloning: A Laboratory Manual (3rd Edition, 2001); Methods In Enzymology (S. Colowick and N. Kaplan eds., Academic Press, Inc.).

Inducible Production-Phase Promoters for Heterologous Expression in Yeast

[163] In accordance with several embodiments, inducible production-phase promoters can be constructed into exogenous expression vectors for production of at least one protein in *Saccharomyces* yeast. In many embodiments, the constructed expression vectors have multiple inducible production-phase promoters in order to express multiple heterologous genes. Several embodiments are directed to production-phase promoters and DNA vectors incorporating these promoters. Promoters, in general, are defined as a noncoding portion of DNA sequence situated proximately upstream of a gene to regulate and promote its expression. Typically, in *S. cerevisiae* and similar species, the promoter of a gene can be found within 500-bp upstream of a gene's translation start codon. In some cases, a promoter may be about 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 1.5 kb, 2 kb or more than 2 kb upstream of a gene's transcription start site.

[164] In accordance with several embodiments, production-phase promoters have two defining characteristics. First, production-phase promoters are capable of repressing heterologous expression of a gene in *S. cerevisiae* and similar species when the yeast is exhibiting anaerobic energy metabolism. As described previously, yeast exhibit anaerobic metabolism in the presence of a nontrivial concentration of fermentable carbon sources such as, for example, glucose or dextrose. In addition, production-phase promoters are also capable of inducing heterologous expression of a gene in *S. cerevisiae* and similar species when the yeast is exhibiting aerobic energy metabolism. As described previously, yeast exhibit aerobic metabolism when fermentable carbon sources are near depleted and the yeast cells switch to a catabolism of nonfermentable carbon sources such as glycerol or ethanol. These characteristics correspond to the phase charts in Figures 6A–6C. Tables 1 and 2 provide several examples of production-phase promoters in accordance with several embodiments.

[165] The production-phase promoters can be characterized based on their level of transgene expression relative to each other and to constitutive promoters. As described in an exemplary embodiment below, it was found that the sequence of endogenous promoters of the *S. cerevisiae* genes ADH2, PCK1, MLS1, and ICL1 exhibited high-level expression and thus can be characterized

as strong production-phase promoters (**Table 1**). Sequences of the endogenous promoters of the *S. cerevisiae* genes YLR307C-A, ORF-YGR067C IDP2, ADY2, CAC1, ECM13, and FAT3 exhibited mid-level expression and thus can be characterized as semi-strong production phase promoters (**Table 1**). In addition, sequences of the endogenous promoters of the *S. cerevisiae* genes PUT1, NQM1, SFC1, JEN1, SIP18, ATO2, YIG1, and FBP1 exhibited low-level expression and thus can be characterized as weak production-phase promoters (**Table 1**).

Table 1. Production-Phase Promoters Expression Phenotype

Gene Name	Systematic Name	Expression Phenotype	Sequence ID Number
ADH2	YMR303C	Strong	1
PCK1	YKR097W	Strong	2
MLS1	YNL117W	Strong	3
ICL1	YER065C	Strong	4
YLR307C-A	YLR307C-A	Semi-Strong	5
YGR067C	YGR067C	Semi-Strong	6
IDP2	YLR174W	Semi-Strong	7
ADY2	YCR010C	Semi-Strong	8
GAC1	YOR178C	Semi-Strong	9
ECM13	YBL043W	Semi-Strong	10
FAT3	YKL187C	Semi-Strong	11
PUT1	YLR142W	Weak	12
NQM1	YGRO43C	Weak	13
SFC1	YJR095W	Weak	14
JEN1	YKL217W	Weak	15
SIP18	YMR175W	Weak	16
ATO2	YNR002C	Weak	17
YIG1	YPL201C	Weak	18
FBP1	YLR377C	Weak	19

[166] The closely related *S. sensu stricto* species have similar genetics and growth characteristics. Accordingly, the phase charts provided in **Figures 6A–6C** apply generally to *S. sensu stricto* species. **Table 2** provides a list of strong production-phase exogenous promoters of similarly related species in accordance with numerous embodiments of the disclosure.

Table 2. Strong Production-Phase Promoters of *S. sensu stricto* species

Species	Gene Name	Sequence ID Number
<i>S. paradoxus</i>	ADH2	36
<i>S. kudriavzevii</i>	ADH2	37

<i>S. bayanus</i>	ADH2	38
<i>S. paradoxus</i>	PCK1	41
<i>S. kudriavzevii</i>	PCK1	42
<i>S. bayanus</i>	PCK1	43
<i>S. paradoxus</i>	MLS1	44
<i>S. kudriavzevii</i>	MLS1	45
<i>S. bayanus</i>	MLS1	46
<i>S. paradoxus</i>	ICL1	47
<i>S. kudriavzevii</i>	ICL1	48
<i>S. bayanus</i>	ICL1	49

[167] It should be noted that substantially similar sequences to the production-promoter sequences are expected to regulate heterologous expression in *S. cerevisiae* and achieve similar results. Accordingly, a substantially similar sequence of a production-phase promoter, in accordance with numerous embodiments, is any sequence with a high functional equivalence such that when regulating heterologous expression in *S. cerevisiae* that it achieves substantially similar results. For example, in an exemplary embodiment below, it was found that the ADH2 promoter of *S. bayanus* is only 61% homologous, yet achieved strong heterologous expression in *S. cerevisiae*, similar to the endogenous ADH2 promoter. In some cases, a substantially similar sequence may be homologous to the promoter sequences identified herein (e.g., have a nucleotide BLAST e value of less than or equal to 10^{-10} , 10^{-20} , 10^{-30} , 10^{-35} or 10^{-40}).

[168] In **Figure 7A**, an exemplary schematic of a section of an exogenous DNA vector (e.g., cloning vector, expression vector, and/or shuttle vector) having a production-phase promoter sequence embedded within. A vector is capable of transferring nucleic acid sequences to target cells (e.g., yeast). Typical DNA vectors include, but are not limited to, plasmid or viral constructs. DNA vectors are also meant to include a kit of various linear DNA fragments that are to be recombined to form a plasmid or other functional construct, as is common in yeast homologous recombination methods (See e.g., Z. Shao, H. Zhao & H. Zhao, 2009, Nucleic Acids Research 37:e16, 2009, the disclosure of which is incorporated herein by reference). Often, embodiments of cloning vectors will incorporate other sequences in addition to the production-phase promoter. As depicted in **Figure 7A**, the exemplary cloning vector has a terminator sequence and cloning/recombination sequence in addition to the production-phase promoter, each of which can assist with expression vector construction. Furthermore, other sequences necessary for growth and amplification can be

incorporated into the promoter vector. Embodiments of these sequences may include, for example, at least one appropriate origin of replication, at least one selectable marker, and/or at least one auxotrophic marker. It should be noted, however, that various embodiments of the disclosure are not required to contain cloning, terminator, or either sequences. For example, embodiments of a typical shuttle vector may only contain the production-phase promoter sequence along with the necessary sequences for amplification in a biological system.

[169] For purposes of this application, an exogenous DNA vector is any DNA vector that was constructed, at least in part, exogenously. Accordingly, DNA vectors that are assembled using the yeast's own cell machinery (e.g., yeast homologous recombination) would still be considered exogenous if any of the DNA molecules transduced within yeast for recombination contain exogenous sequence or were produced by a non-host methodology, such as, for example, chemical synthesis, PCR amplification, or bacterial amplification.

[170] As shown in **Figure 7B**, various embodiments of the disclosure are directed to DNA vectors having multiple production-phase promoters. In these various embodiments, multiple different production-phase promoters are incorporated, preferably each having a unique sequence and derived from a different gene and/or *S. sensu stricto* species. Having unique promoter sequences can prevent complications that can arise during product production in yeast, such as, for example, unwanted DNA recombination at sites similar to the promoter sequences that render the DNA vector constructs undesirable. In many embodiments, the DNA vector has at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more than 20 production-phase promoters. As the size of the DNA vector increases, the utility may decrease, as larger vectors may become unwieldy for the intended organism to handle. For example, plasmids for amplification in *E. coli* are often somewhere between 2,000 and 10,000 base pairs (bp) but can handle up to 20,000 bp or so. Likewise, plasmids for amplification and growth in yeast can vary from approximately 10,000 to 30,000 bp. Viral vectors, on the other hand, often have a limited construct size and thus may require a more precise vector size. Thus, depending on vector and intended use, the number of production-phase promoters within a DNA vector will vary.

[171] Although **Figure 7B** depicts recombination sites, cloning sites, and terminator sequences, it should be noted that these sequences may or may not be included in various embodiments of DNA vectors having multiple production-phase promoters. The incorporation of these sequences or other

various sequences is often dependent on the purpose of the DNA vector. For example, cloning vectors may not include a terminator sequence if that sequence is to be incorporated into an expression construct at another stage of assembly.

[172] **Figure 8A** depicts an exemplary heterologous expression vector having a production-phase promoter for expression in yeast, in accordance with various embodiments of the disclosure. Expression constructs contain an expression cassette that has a promoter, a heterologous gene, and a terminator sequence in order to produce an RNA molecule in an appropriate host. Expression cassette in accordance with numerous embodiments will have a production-phase promoter situated proximately upstream of a heterologous gene of which the promoter is to regulate expression. It should be understood, that the precise location of the production-phase promoter upstream of the heterologous gene may vary, but the promoter generally is within a certain proximity to adequately function.

[173] In many embodiments of the disclosure, a heterologous gene is any gene driven by a production-phase promoter, wherein the heterologous gene is different than the endogenous gene that the promoter regulates within its endogenous genome. Accordingly, a *S. cerevisiae* production-phase promoter could regulate another *S. cerevisiae* gene provided that the gene to be regulated is not the gene endogenously regulated. For example, the *S. cerevisiae* ADH2 promoter should not regulate the *S. cerevisiae* ADH2 gene; however, the *S. cerevisiae* ADH2 promoter can regulate any other *S. cerevisiae* gene or the ADH2 gene from any other species. Often, in accordance with many embodiments, the heterologous gene is from a different species than the species from which the production-promoter sequence was obtained.

[174] Although not depicted, various embodiments of expression cassettes may include other sequences, such as, for example, intron sequences, Kozak-like sequences, and/or protein tag sequences (e.g., 6×-His) that may or may not improve expression, production, and/or purification. In yeast, various embodiments of expression vectors will also minimally have a yeast origin of replication (e.g., 2-micron) and an auxotrophic marker (e.g., URA3) in addition to the expression cassette. Other nonessential sequences may also be included, such as, for example, bacterial origins of replication and/or bacterial selection markers that would render the expression capable of amplification in a bacterial host in addition to a yeast host. Accordingly, various embodiments of

expression vectors would include the essential sequences for heterologous expression in yeast and other various embodiments would include additional nonessential sequences.

[175] In accordance with various embodiments, a DNA vector having a production-phase promoter expression cassette can be transformed into a yeast cell. Or alternatively, and in accordance with numerous embodiments, a DNA vector having a production-phase promoter expression cassette can be assembled within yeast using homologous recombination techniques. Once existing within a yeast cell, the production-phase promoter can regulate the expression of a heterologous gene in accordance with the yeast cell's energy metabolism. As described previously, and in accordance with many embodiments, production-phase promoters repress heterologous expression when the yeast cell is in an anaerobic energy metabolic state. Alternatively, and in accordance with a number of embodiments, production-phase promoters induce heterologous expression when the yeast cell is in an aerobic energy metabolic state

[176] Depicted in **Figure 8B** are alternative exemplary heterologous expression vectors having multiple production-phase promoters for expression of multiple genes in yeast in accordance with numerous embodiments. In some embodiments, the expression vectors will include at least two expression cassettes, each with a unique promoter, gene, and terminator sequence in order to prevent unwanted recombination. The number of expression cassettes will vary based on vector construct design and application. For heterologous expression in *S. cerevisiae*, it has been found that plasmid expression vectors of approximately 30,000 bp are still tolerated. Thus, vectors containing up to seven production-phase promoter expression cassettes can be incorporated into an expression vector and have been found to be able to maintain adequate gene expression and protein production. Larger vectors with more expression cassettes may be tolerated.

[177] Although **Figure 8B** depicts multiple expression cassettes sequentially in the same orientation (5' to 3'), it should be understood that the combination of two or more expression cassettes is not limited to sequential linear organization in the same orientation. Expression cassettes in accordance with many embodiments exist within the expression vector in any orientation and in any sequential order. Furthermore, it should be understood that other sequence elements of an expression vector (e.g., an auxotrophic marker) may be among and/or between the multiple expression cassettes. Optimal vector design is likely to depend on various factors, such as, for

example, optimizing the location of the auxotrophic marker to enable the final expression vector to include each expression cassette to be incorporated.

[178] DNA heterologous expression vectors are a class of DNA vectors, and thus the description of general DNA vectors above also applies to the expression vectors. Accordingly, many embodiments of the expression vectors are formulated into a plasmid vector, a viral vector, a circular vector, or a kit of linear DNA fragments to be recombined into a plasmid by yeast homologous recombination. In several of these embodiments, the end-product vector contains at least one expression cassette having a production-phase promoter. It should be understood, that in addition to the at least one production-phase promoter, some vector embodiments incorporate expression cassettes that include other promoters, such as (but not limited to), constitutive promoters that maintain high expression during the growth and production phases.

[179] The various embodiments of heterologous expression vectors having at least one production-phase promoter can be used in numerous applications. For example, high expression in the production phase can lead to better, prolonged expression, as compared to constitutive promoters. In many applications, the end product is a protein from a single gene or a protein complex of multiple genes to be purified from the culture. For these applications, high, prolonged expression using production-phase promoters can lead to better yields of proteins. Furthermore, when the heterologous protein is toxic to the host yeast cells, the use of production-phase promoters prevents the expression of the toxic protein during growth phase, allowing the yeast to reach a healthy confluency before mass protein production.

[180] The production-phase promoter vectors can also benefit the production of a biosynthetic compound from a gene cluster. Many products derived from various natural species are produced from a cluster of genes with sequential enzymatic activity. For example, the antibiotic emindole SB is produced from a cluster of four genes that is expressed in *Aspergillus tubingensis*. To reproduce this gene cluster in a yeast production model, a production-promoter vector system with four different expression cassettes could work. This system would allow the yeast to reach a healthy confluency before the energy-draining expression of four heterologous proteins begins, leading to better overall yields of the antibiotic product. In fact, experimental results provided in an exemplary embodiment described in Example 1 below demonstrate that a production-phase promoter vector

outperformed a constitutive promoter vector approximately 2-fold to produce the emindole SB product.

[181] **Figure 9** depicts an exemplary process (Process 400) to implement various embodiments of production-phase promoters. To begin, Process 400 identifies and selects at least one gene for heterologous expression in yeast (401). The choice of gene(s) for expression would depend on the desired outcome. For example, to produce a biosynthetic compound, one would likely select to express all, or a subset, of the genes within a biosynthetic gene cluster of a particular organism. Once the gene(s) have been selected, Process 400 then appropriates DNA molecules having the coding sequence of the selected genes (403). As is well known in the art, there are many ways to appropriate DNA molecules, which include chemical synthesis, extraction directly from the biological source, or amplification of a gene by polymerase chain reaction (PCR).

[182] Process 400 then uses the appropriated DNA molecules to assemble these molecules into an expression vector having production-phase promoters (405). There are many ways to assemble DNA expression vectors that are well known in the art, which include popular methodologies such as homologous recombination and restriction digestion with subsequent ligation. After assembly, the resultant expression vectors can be expressed in *Saccharomyces* yeast to obtain the desired outcome (407).

Yeast homologous recombination for plasmid construction and direct plasmid sequencing

[183] Each of the one or more expression vectors may contain one or more promoters suitable for expression of a heterologous gene in a model host system. Each expression vector may contain a single coding sequence or multiple coding sequences. Multiple coding sequences may be functionally linked to a single promoter, for example via an internal ribosome entry site, or may be linked to multiple promoters. The expression vectors may also contain additional elements to regulate or increase the transcriptional activity, for example enhancers, polyA sequences, introns, and posttranscriptional stability elements. The expression vectors may also contain one or more selectable markers.

[184] To improve high-throughput assembly and characterization of orphan biosynthetic systems or other systems, an automated DNA assembly pipeline using yeast homologous recombination, (YHR), as its core technology was developed. An example of the design strategy for assembling

DNA parts for this pipeline is illustrated in **Figure 10A**. In one embodiment a method is provided for assembling synthetic gene clusters with heterologous regulation. DNA polynucleotides coding for a series of distinct promoters and terminators may be obtained in bulk and used for a variety of different synthetic gene clusters. Once a gene cluster of interest is identified the coding sequences are determined and then all coding sequences are synthesized with a flanking sequence (assembly overhang) on each side. The assembly overhangs, encode either for the flanking promoter and terminator if the gene is small enough to be ordered as a single piece, or for the adjacent gene fragments for longer sequences. The length of the flanking sequences may vary, In some cases, the flanking sequences may be about 30 bp, 40 bp, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 30-100 bp, 30-70 bp, 40-60 bp, 40-80 bp, or 45-55 bp in length. Placing assembly overhangs exclusively on the unique coding sequence fragments allows for all regulatory cassettes to be generated in bulk and stockpiled as the same fragments are used in all assemblies. For example, in an assembly involving three or more genes, an auxotrophic marker may be placed between the second terminator and third terminator while no marker is present on the vector. By providing the auxotrophic marker and origin of replication on separate fragments, reaction background was significantly reduced. Additional modest increases in efficiency were observed when the assembly host is lacking a DNA ligase, such as the DNL4 DNA ligase.

[185] In some embodiments, this disclosure provides a system for generating a synthetic gene cluster via homologous recombination. The system comprises 1 through N unique promoter sequences, 1 through N unique terminator sequences, and 1 through N unique coding sequences. Each terminator sequence may be linked to the following promoter sequence, for example terminator 1 is linked to promoter 2, terminator 2 is linked to promoter 3, and so forth till terminator N-1 which is linked to promoter N. In some cases, promoter 1 and terminator N may be attached to a linear plasmid backbone. Coding sequence 1 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical or homologous to the last 30–70 base pairs of promoter 1 and a second end portion is identical or homologous to the first 30–70 base pairs of terminator 1. Coding sequence 2 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical or homologous to the last 30–70 base pairs of promoter 2 and a second end portion is identical or homologous to the first 30–70 base pairs of terminator 2. Coding sequence N is also attached to an additional 30–70 base pair sequence on each end such that a first end portion

is identical or homologous to the last 30–70 base pairs of promoter N and a second end portion is identical or homologous to the first 30–70 base pairs of terminator N. These DNA fragments may be assembled transforming the 1 through N promoters, terminators and coding sequences into a yeast cell where they are combined through yeast homologous recombination, and then isolating a plasmid containing the 1 through N promoters, terminators and coding sequences from the yeast cell.

[186] An example of this system is shown in **FIG. 10A**. In this example N equals four. The system comprises four unique promoters (110, 120, 130 and 140), four unique terminator sequences (210, 220, 230 and 240), and four unique coding sequences (310, 320, 330, and 340). Each of the coding sequences is created with an additional 30–70 base pair sequence that is homologous or identical to the sequence of the preceding promoter and an additional 30–70 base pair sequence that is homologous or identical to the sequence of the subsequent terminator. Thus, coding sequence 310 is flanked by sequence 111 which is identical or homologous to at least a part of sequence 110, and sequence 211 which is identical or homologous to at least a part of sequence 210. Coding sequence 320 is flanked by sequence 121 which is identical or homologous to at least a part of sequence 120, and sequence 221 which is identical or homologous to at least a part of sequence 220. Coding sequence 330 is flanked by sequence 131 which is identical or homologous to at least a part of sequence 130, and sequence 231 which is identical or homologous to at least a part of sequence 230. Coding sequence 340 is flanked by sequence 141 which is identical or homologous to at least a part of sequence 140, and sequence 241 which is identical or homologous to at least a part of sequence 240. In this example promoter sequence 110 and terminator sequence 240 are attached to the ends of a linearized plasmid backbone, and the DNA fragment comprising terminator 210 and promoter 120 further comprises an auxotrophic marker (400). Terminator 210 is linked to promoter 120, terminator 220 is linked to promoter 130, and terminator 230 is linked to promoter 140.

[187] As shown in **FIG. 10B**, once the DNA sequences from **FIG. 10A** are transfected into a yeast cell the homologous sequences are paired up and the fragments are linked together through yeast homologous recombination. The resultant DNA plasmid is illustrated in **FIG. 10C**.

[188] Traditionally, for yeast homologous recombination plasmid assemblies, plasmid DNA is isolated from assembly clones and transformed into *E. coli* in order to obtain sufficiently pure DNA to enable sequencing. The necessity of this step arises from the relatively low plasmid yields from yeast and the large amounts of contaminating genomic DNA in every sample. This disclosure

provides a method by which plasmid DNA may be sequenced directly out of yeast. This may be achieved by a modified plasmid prep in which the majority of contaminating DNA is removed by treatment with an exonuclease enzyme. Any enzyme with exonuclease activity and free of endonuclease activity may be used in this step. Examples of exonuclease enzymes include but are not limited to: Lambda Exonuclease, RecJf, Exonuclease III (*E. coli*), Exonuclease I (*E. coli*), Exonuclease T, Exonuclease V (RecBCD), Exonuclease VIII, truncated, Exonuclease VII, T5 Exonuclease, and T7 Exonuclease. In some cases, the exonuclease is Exonuclease V. If an exonuclease with activity for single stranded DNA (ssDNA) and not double stranded DNA (dsDNA) is to be used, then the DNA may first be heated to denature the dsDNA. In some cases, the DNA may be treated with a topoisomerase to relax supercoiled plasmids. In some cases, the DNA may not be treated with a topoisomerase. Once the plasmid DNA has been purified by this method sequencing libraries can be prepared. **Figure 10D** demonstrates the increase in purity observed with the exonuclease treatment. Overall, this pipeline has been applied to the sequencing of >1000 clones. Assemblies of up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more than 30 unique DNA fragments can be achieved with high efficiency. **Figure 10E** shows efficient assembly of 2, 3, 4, 5, 6, 8, 10, 12, and 14 DNA fragments. In some cases a strain as described herein allows DNA assembly via homologous recombination with an efficiency of at least 60%, 70%, 80%, 90%, 100%, 110%, 120%, 130%, 140%, 150%, or more than 150% as compared to DNA assembly in BY.

[189] To increase the efficiency of assemblies by yeast homologous repair, the relative efficiencies of BY4743 and BY4743 Δ DNL4 were tested; a strain in which the DNL4 ligase, involved in non-homologous end joining, has been deleted. **Figure 11A** illustrates efficiencies of several plasmid assemblies done in both strains, demonstrating the deletion of the DNL4 DNA ligase does consistently serve as the more efficient assembly background.

[190] The sequencing of plasmid DNA directly out of yeast as in **Figure 11E** (e.g. without transforming into another host such as *E. coli* as in **Figure 11D**) is an advantage of the methods described herein. In establishing these methods, multiple means of preparation for both the plasmid DNA and the next-generation sequencing (NGS) library prep were tested. Shown in **Figure 11B** is a comparison of sequencing efficiency using DNA prepared both from colonies picked from plates

and cell pellets collected from 1 ml of liquid culture. These data show that these approaches generate samples of equivalent purity.

[191] Initially, the platform utilized an NGS library preparation in which the purified, exonuclease treated plasmid DNA was ultrasonically sheared, followed by end repair, A-tailing, and adaptor ligation. In order to decrease labor and increase throughput, a recently published modification of the Illumina NexeraXT transposase based prep was performed (M. Baym, et al., *PLoS One* 10:e01280367, 2015). Ultrasonic shearing necessitated the serial processing of multiple plates of clones while tagmentation allows for parallel processing of multiple plates. **Figure 11C** demonstrates that this modified Nextera preparation provides equivalent efficiency as compared to the standard approach.

[192] This approach may be suitable for multiple DNA preparation methods in multiple strain backgrounds. Additionally, it was shown that this approach is compatible with various library preparations for sequencing on Illumina platforms. It is anticipated that this approach could be easily modified to function in sequencing workflows using alternate sequencing platforms such as those provided by Pacific Bioscience and Oxford Nanopore technologies.

Host cells

[193] The expression vectors may be transfected into a host cell to produce secondary metabolites. The host cell may be any cell capable of expressing the coding sequences from the expression vectors. The host cell may be a cell which can be grown and maintained at a high density. For example the host cell may be one which may be grown and maintained in a bioreactor or fermenter. The host cell may be a fungal cell, a yeast cell, a plant cell, an insect cell or a mammalian cell.

[194] In some cases the host cell is bacterial. The bacteria may be a Proteobacteria such as a Caulobacteria, a phototrophic bacteria, a cold adapted bacteria, a Pseudomonads, or a Halophilic bacteria; an Actinobacteria such as Streptomyces, Nocardia, Mycobacteria, or Coryneform; a Firmicutes bacteria such as a Bacilli, or a lactic acid bacteria. Examples of bacteria which may be used include, but are not limited to: *Caulobacter crescentus*, *Rodhobacter sphaeroides*, *Pseudoalteromonas haloplanktis*, *Shewanella sp. strain Ac10*, *Pseudomonas fluorescens*, *Pseudomonas putida*, *Pseudomonas aeruginosa*, *Halomonas elongata*, *Chromohalobacter salexigens*, *Streptomyces lividans*, *Streptomyces griseus*, *Nocardia lactamdurans*, *Mycobacterium*

smegmatis, *Corynebacterium glutamicum*, *Corynebacterium ammoniagenes*, *Brevibacterium lactofermentum*, *Bacillus subtilis*, *Bacillus brevis*, *Bacillus megaterium*, *Bacillus licheniformis*, *Bacillus amyloliquefaciens*, *Lactococcus lactis*, *Lactobacillus plantarum*, *Lactobacillus casei*, *Lactobacillus reuteri*, and *Lactobacillus gasseri*.

[195] In some cases the host cell is a fungal cell. In some cases, the host cell is a yeast cell. Examples of yeast cells include, but are not limited to *Saccharomyces cerevisiae*, *Saccharomyces pombe*, *Candida albicans*, and *Cryptococcus neoformans*. In some cases the host cell may be a filamentous fungi, such as a mold. Examples of molds include, but are not limited to *Acremonium*, *Alternaria*, *Aspergillus*, *Cladosporium*, *Fusarium*, *Mucor*, *Penicillium*, and *Rhizopus*. In some cases, the host cell may be an *Acremonium* cell. In some cases, the host cell may be an *Alternaria* cell. In some cases, the host cell may be an *Aspergillus* cell. In some cases, the host cell may be an *Cladosporium* cell. In some cases, the host cell may be an *Fusarium* cell. In some cases, the host cell may be a *Mucor* cell. In some cases, the host cell may be a *Penicillium* cell. In some cases, the host cell may be a *Rhizopus* cell.

[196] In some cases, the host cell is an insect cell. In some cases the host cell is a mammalian cell. Examples of mammalian cell lines include HeLa cells, HEK293 cells, B16 melanoma cells, Chinese hamster ovary cells, or HT1080. In some cases, the host cell is a plant cell. In some cases, the host cell may be part of a multicellular host organism.

[197] In some cases, the host cell is a genetically engineered cell. The yeast strain BJ5464 has historically been a workhorse strain for expression of heterologous proteins. BJ5464 lacks two vacuolar proteases genes (PEP4 and PRB1), which makes the strain useful for biochemical studies, owing to reduced protein degradation. However, BJ5464 has several problems that limit its utility. It has a high rate of petite cell formation, which results in offspring that cannot respire (grow on ethanol as a carbon source) and cannot express the respiration-induced promoters used in this project. It is not genetically tractable because it cannot sporulate, and its non-deletion auxotrophic markers prevent facile genome editing. Finally, BJ5464 is slow growing.

[198] This disclosure includes a new yeast super host based on the BY background. BY is a direct descendent of the yeast genome sequence reference strain and contains the complete deletions of auxotrophic markers that facilitate genome editing. It is the basis of the barcoded deletion collection, which has led to a wealth of genetic and chemico-genomic data. However, it also has major

problems limiting its utility. In particular, it has the poorest sporulation frequency and highest petite frequency of all common lab strains.

[199] The petite phenotype arises due to a defect in aerobic respiration. Petite yeasts are unable to grow on non-fermentable carbon sources (for example glycerol or ethanol), and form small anaerobic-sized colonies when grown in the presence of fermentable carbon sources (for example glucose). The phenotype results from mutations in the mitochondrial genome, loss of mitochondria, or mutations in the host cell genome.

[200] The genes and single nucleotide polymorphisms (SNPs) responsible for the sporulation and respiration defects have been identified (**Figure 12**). The sporulation defect may be repaired by a series of genetic crosses to a previously repaired strain. The respiration problem caused by mitochondrial genome instability may also be corrected using genome editing. Genome editing may be performed with any method known in the art, such as the 50:50 method. (J. Horecka and R. W. Davis. *Yeast* 31:103-12, 2014). In some cases, an improved version of Mega 50:50 may be used in which a double stranded break is introduced into the genomic locus to be modified, increasing efficiency by several orders of magnitude (J. D. Smith, et al., *Mol. Syst. Biol.* 13:913, 2017).

[201] In some cases, the host cell may be a cell which has been engineered to repair a sporulation defect. For example, the host cell may be a fungal cell with a repaired sporulation defect. In some cases, the host cell is a yeast cell with a repaired sporulation defect. In some cases, the host cell is a BY yeast cell in which the sporulation defect has been repaired, as in **Figure 12**.

[202] In some cases, the host cell may be a cell which has been engineered to repair a respiratory defect or a mitochondrial genome instability defect. For example, the host cell may be a fungal cell with a repaired mitochondrial stability defect. In some cases, the host cell is a yeast cell with a repaired mitochondrial stability defect. In some cases, the host cell is a BY yeast cell in which the mitochondrial genome instability defect has been repaired, as in **Figure 12**. In some cases, the host cell may be a cell in which both a sporulation defect and a mitochondrial genomic instability defect have been repaired. Using the genetic crosses and genome engineering methods discussed above and the genomic repairs outlined in **Figure 12** the BY strain was engineered to repair the mitochondrial genome instability. An unexpected benefit of repairing the mitochondrial genome instability defect was that the strains grew faster on non-fermentable carbon sources, such as ethanol (see **Figure 13 and 14A**). This is commonly the growth condition of choice for expression of heterologous genes,

which are often linked to production-phase promoters activated by growth on non-fermentable carbon sources.

[203] In some cases, the host cell may be genetically engineered to lack a gene involved in non-homologous end joining. The lack of such a gene may increase the efficacy of homologous recombination in such an engineered cell. The appropriate genes to delete may vary in each host. As an example, an engineered yeast host cell may lack a ligase such as the DNL4 DNA ligase. An engineered bacterial host cell may lack one or both of a Ku homodimer and the multifunctional ligase/polymerase/nuclease LigD. Other genes which may be involved in non-homologous double stranded break repair, depending on species, include: Mre11, Rad50, Xrs2, Nbs1, DNA-PKcs, Ku70, Ku80, DNA ligase IV, XLF, Artemis, XRCC4, Dnl4, Lif1, XLF also known as Cernunnos, Nej1 and Sir2.

[204] DHY strains have utility for expression of heterologous genes for heterologous compound production, as well as ability to perform homologous recombination and DNA assembly. This combination of abilities in one strain allows DNA assembly and production of heterologous compounds in the same strain, whereas previously these two steps were separated in previous types of yeast (BY for DNA assembly and BJ5464 for expression and small molecule production).

[205] These improvements can provide the DHY strain collection with a number of advantages over previous strains, particularly the BY strains: DHY can be faster-growing, result in fewer petite colonies (respiration-deficient), be genetically tractable, allow better expression from ADH2-like promoters, and allow both DNA assembly and production of heterologous products in the same strain (**Figure 14B and 14C**).

[206] In some embodiments, a genetically engineered host cell lacks one or more conditionally essential genes which can be provided by a plasmid or other DNA vector. This allows for the selection of cells which are expressing the DNA vector. Examples of genes which may be used for this are auxotrophic genes which are required for biosynthesis of certain metabolites or genes for resistance to a toxin. Auxotrophic genes are only required when the specific metabolite they are required for is not present in the culture media. Resistance genes are only required when the toxin which they provide protection from is present.

[207] Examples of genetically engineered yeast host cells include genetically engineered DHY super-host strains. In some cases, strains are based on the BY4741 / BY4742 background (C. B.

Brachmann, et al, *Yeast*, 14:115-32, 1998). Strains may also contain any of the following genetic changes from the BY background: sporulation repair (*MKT1(30G)* *RME1(INS-308A)* *TAO3(1493Q)*), and mitochondrial genome stability and function repair (*CAT5(91M)* *MIP1(661T)* *SAL1+ HAP1+*) (see **Figure 12**). It should be noted, as would be understood in by persons having ordinary skill in the art, that any and all these genetic changes can be performed in isolation, in part, or in totality. For example, it is expected that the a single genetic change of either *MKT1(30G)*, *RME1(INS-308A)*, or *TAO3(1493Q)* would result in at least some repair in sporulation activity. Likewise, a single genetic change of either *CAT5(91M)* or *MIP1(661T)* or restoration of function of *SAL1 (SAL1+)* or *HAP1 (HAP1+)* would result in at least some increase in mitochondrial genome stability.

[208] In some cases, a strain may be a prototroph. For example, some strains may require methionine, arginine or lysine in the media. In some cases, a strain may be a full heterozygote for several markers from which any combination of markers can be made by tetrad dissection. For example, heterozygous for genes required for synthesis of histidine, leucine, uracil, lysine and methionine, or heterozygous for genes required for synthesis of histidine, leucine, uracil, lysine and arginine. Some examples of strains are listed in Table 3.

[209] In some cases, the use of a strain as described herein allows for greater expression of BGC proteins and/or greater production of compounds from the BGCs. In some cases expression of heterologous proteins in a strain described herein is accomplished with an efficiency of at least 70%, 80%, 90%, 100%, 110%, 120%, 130%, 140%, or 150% as compared to heterologous protein expression in BJ5464. In some cases production of heterologous compounds in a strain described herein is accomplished with an efficiency of at least 70%, 80%, 90%, 100%, 110%, 120%, 130%, 140%, or 150% as compared to heterologous compound production in BJ5464.

Table 3: Description of strain genotypes

Strain	Parent	Genotype	Reference
BY4741	S288C	MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0	(C. B. Brachmann, et al., 1998, cited supra)

BY4743	S288C	MATa/ α his3 Δ 1/his3 Δ 1 leu2 Δ 0/leu2 Δ 0 LYS2/lys2 Δ 0 met15 Δ 0/MET15 ura3 Δ 0/ura3 Δ 0	(C. B. Brachmann, et al., 1998, cited supra)
BJ5464		MAT α ura3-52 trp1 leu2- Δ 1 his3- Δ 200 pep4::HIS3 prb1- Δ 1.6R can1 GAL	(E. W. Jones, <i>Methods Enzymol.</i> 194:428-53, 1991)
BY4743 Δ D NL4	S288C	MATa/Mata dnl4 Δ /dnl4 Δ	(E. A. Winzeler, et al., <i>Science</i> 285:901-06, 1999)
Y800		MATa ade2-1 leu2- Δ 98 ura3-52 lys2-801 trp1-1 his3- Δ 200 [cir0]	(N. Burns, et al., <i>Genes Dev.</i> 8:1087-105, 1994)
DHY213	BY4741	MATa his3 Δ 1 leu2 Δ 0 ura3 Δ 0 met15 Δ 0 SAL1+ HAP1+ CAT5(91M) MIP1(661T) MKT1(30G) RME1(INS-308A) TAO3(1493Q)	n/a
JHY693	DHY213	MATa his3 Δ 1 leu2 Δ 0 ura3 Δ 0 met15 Δ 0 SAL1+ HAP1+ CAT5(91M) MIP1(661T) MKT1(30G) RME1(INS-308A) TAO3(1493Q) prb1 Δ pep4 Δ	n/a
JHY651	DHY213	MAT α his3 Δ 1 leu2 Δ 0 ura3 Δ 0 met15 Δ 0 SAL1+ HAP1+ CAT5(91M) MIP1(661T) MKT1(30G) RME1(INS-308A) TAO3(1493Q) prb1 Δ pep4 Δ lys2 Δ 0	n/a
JHY692	DHY213	MATa his3 Δ 1 leu2 Δ 0 ura3 Δ 0 met15 Δ 0 SAL1+ HAP1+	n/a

		CAT5(91M) MIP1(661T) MKT1(30G) RME1(INS-308A) TAO3(1493Q) prb1Δ pep4Δ ADH2p-npgA-ACS1t	
JHY705	DHY213	MATα his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SAL1+ HAP1+ CAT5(91M) MIP1(661T) MKT1(30G) RME1(INS-308A) TAO3(1493Q) prb1Δ pep4Δ ADH2p-CPR-ACS1t lys2Δ0	n/a
JHY702	DHY213	MATα/MATα his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 ura3Δ0/ura3Δ0 met15Δ0/met15Δ0 SAL1+/SAL1+ HAP1+/HAP1+ CAT5(91M)/CAT5(91M) MIP1(661T)/MIP1(661T) MKT1(30G)/MKT1(30G) RME1(INS-308A)/RME1(INS-308A) TAO3(1493Q)/TAO3(1493Q)) prb1Δ/prb1Δ pep4Δ/pep4Δ ADH2p-npgA-ACS1t/ADH2p-CPR-ACS1t met15Δ0/+ lys2Δ0/+	n/a

Detection and characterization of novel molecules

[210] Once a host cell is expressing the coding sequences of the identified gene cluster, a secondary metabolite may be synthesized in the host cell. The secondary metabolite may be identified by any method known in the art. In some cases, the secondary metabolite is identified by comparing a host cell expressing the cluster with a host cell which does not express the cluster. This comparison may utilize chromatography methods to separate different small molecules produced in

the cells. For example, column chromatography, planar chromatography, thin layer chromatography, gas chromatography, liquid chromatography, supercritical fluid chromatography, ion exchange chromatography, size exclusion chromatography be done by high performance liquid chromatography (HPLC), mass spectrometry (MS), or by mass spectrometry high performance liquid chromatography (MS-HPLC). Any peaks which appear for the cluster expressing host cell and not from the control host cell indicate the presence of a novel chemical. The comparison between the cluster expressing host cell and the control host cell may comprise a comparison of a cell extract, a culture media, or an extracted cell lysate.

Compounds identified

[211] This disclosure also provides sequences of 43 BGCs, and structures of novel products produced by a subset of these BGCs.

[212] In one embodiment, this disclosure provides sequences of cryptic BGCs which encode various products, **SEQ ID NOs: 67-483**. These BGCs may also be reengineered to provide the coding sequences without the endogenous regulatory sequences. In some examples, the coding sequences may be predicted using known bioinformatics methods, experimental data, or obtained from databases such as default predicted gene coordinates (start, stop, and introns) as deposited in GenBank. Once the coding sequences have been identified the sequences may be isolated and cloned into one or more expression vectors for expression in a model host system such as *S. cerevisiae*.

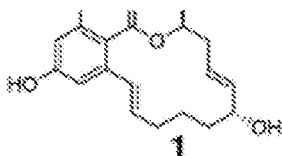
[213] The expression vectors may be plasmids, viruses, linear DNA, bacterial artificial chromosomes or yeast artificial chromosomes. Each of the one or more expression vectors may contain one or more promoters suitable for expression of a heterologous gene in a model host system. Each expression vector may contain a single coding sequence or multiple coding sequences. Multiple coding sequences may be functionally linked to a single promoter, for example via an internal ribosome entry site, or may be linked to multiple promoters. The expression vectors may also contain additional elements to regulate or increase the transcriptional activity, for example enhancers, polyA sequences, introns, and posttranscriptional stability elements. The expression vectors may also contain one or more selectable markers.

[214] The expression vectors may be transfected, or otherwise introduced, into host cells. Examples of host cells include but are not limited to yeast and bacterial cells. For example a host

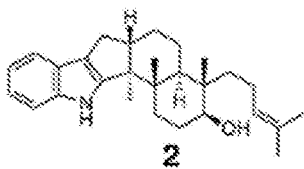
cell may be a *S. cerevisiae* cell or an *E. coli* cell. Incubating the expression vectors in the host cells allows for the transcription and translation of the coding sequences to recreate the proteins of the gene cluster. These proteins may then produce a secondary metabolite which can be isolated from the cells or the media in which the cells are grown.

[215] In another embodiment this disclosure provides host cell extracts containing non-host cell derived products. In some cases, these extracts may be produced by culturing host cells expressing one or more, or all, of the sequences from one of the following groups **SEQ ID NOs: 67-76, 77-81, 82-91, 92-97, 98-106, 107-111, 112-118, 119-127, 128-135, 136-153, 154-157, 158-162, 163-172, 173-181, 182-186, 187-191, 192-199, 200-206, 207-211, 212-224, 225-228, 229-235, 236-240, 241-244, 245-255, 256-267, 268-276, 277-285, 286-289, 290-293, 294-307, 308-313, 314-318, 319-324, 325-329, 330-334, 335-341, 342-350, 351-357, 358-367, 368-372, 373-380, 381-388, 389-395, 396-400, 401-406, 407-413, 414-423, 424-427, 428-439, 440-447, 448-453, 454-462, 463-471, 472-480, or 481-483.** In some cases, a host cell may express all the sequences from one of the following groups **SEQ ID NOs: 67-76, 77-81, 82-91, 92-97, 98-106, 107-111, 112-118, 119-127, 128-135, 136-153, 154-157, 158-162, 163-172, 173-181, 182-186, 187-191, 192-199, 200-206, 207-211, 212-224, 225-228, 229-235, 236-240, 241-244, 245-255, 256-267, 268-276, 277-285, 286-289, 290-293, 294-307, 308-313, 314-318, 319-324, 325-329, 330-334, 335-341, 342-350, 351-357, 358-367, 368-372, 373-380, 381-388, 389-395, 396-400, 401-406, 407-413, 414-423, 424-427, 428-439, 440-447, 448-453, 454-462, 463-471, 472-480, or 481-483.** In some cases the host cell(s) may express one or more sequences selected from **SEQ ID NOs: 67-483.** After culturing the cells, they may be collected, lysed, and the small molecules may be purified from nucleic acid, proteins, complex carbohydrates and lipid containing fractions. The secondary metabolites produced may also be secreted into the cell media. In this case this disclosure also provides a cell media containing secondary metabolites.

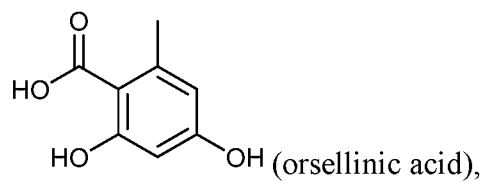
[216] This disclosure also provides compounds isolated from the host cell extracts or media. A compound of this disclosure may be Compound 1:



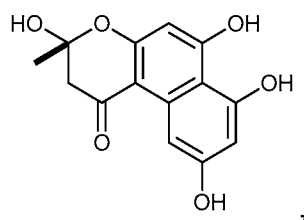
[217] Compound 2:



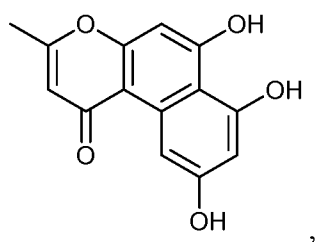
[218] Compound 3:



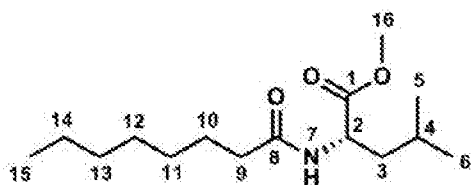
[219] Compound 4:



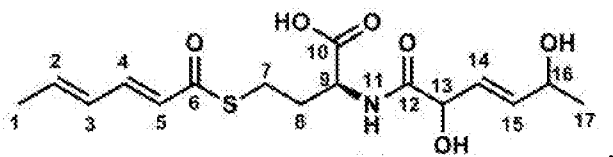
[220] Compound 5:



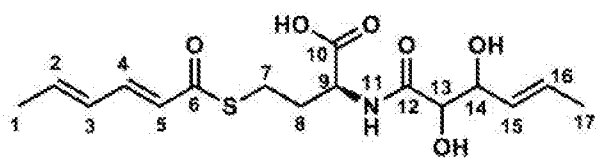
[221] Compound 6:



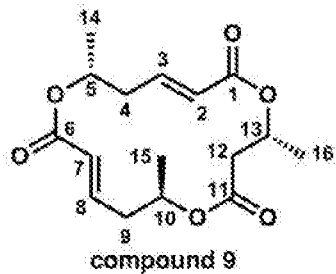
[222] Compound 7:



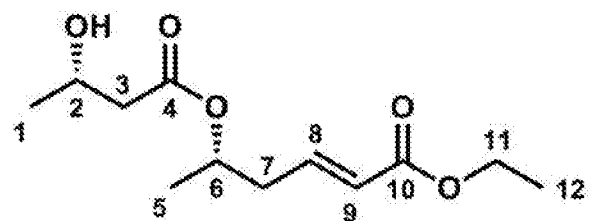
[223] Compound 8:



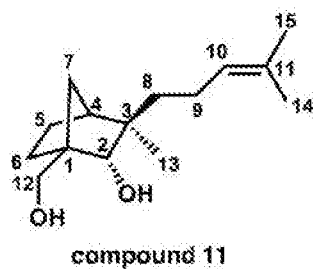
[224] Compound 9:



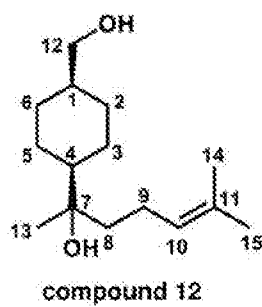
[225] Compound 10:



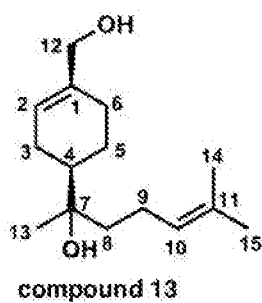
[226] Compound 11:



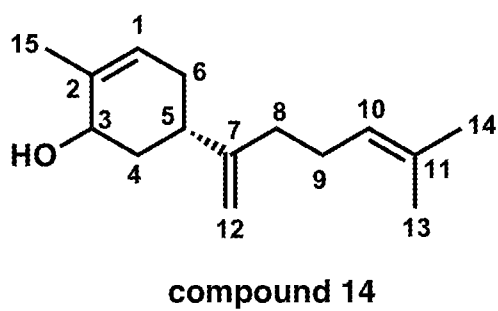
[227] Compound 12:



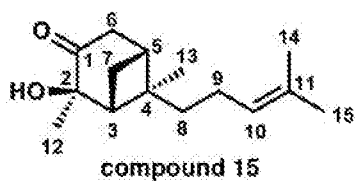
[228] Compound 13:



[229] Compound 14:

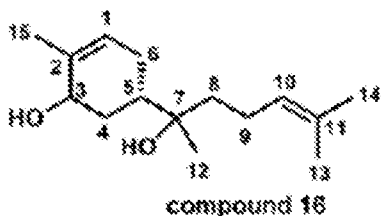


[230] Compound 15:



, and

[231] Compound 16:



[232] Compounds of this disclosure may have useful therapeutic applications, for example in treating or preventing a disease or disorder. Compounds of this disclosure may be used to treat an infection, for example a bacterial, fungal or parasitic infection. Compounds of this disclosure may have antibiotic and/or antifungal activities. Compound 8 and Compound 9 may have antimicrobial, antifungal and/or antibacterial activities. Compounds of this disclosure may have non-medical applications.

[233] In some embodiments this disclosure provides pharmaceutical compositions comprising of a compound of this disclosure. In some cases a pharmaceutical composition contains at least one of Compounds: 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, and 16. Compositions as described herein may comprise a liquid formulation, a solid formulation or a combination thereof. Non-limiting examples of formulations may include a tablet, a capsule, a gel, a paste, a liquid solution and a cream. The compositions of the present disclosure may further comprise any number of excipients. Excipients may include any and all solvents, coatings, flavorings, colorings, lubricants, disintegrants, preservatives, sweeteners, binders, diluents, and vehicles (or carriers). Generally, the excipient is compatible with the therapeutic compositions of the present disclosure. Generally, the excipient is a pharmaceutically acceptable excipient. The pharmaceutical composition may also contain minor amounts of non-toxic auxiliary substances such as wetting or emulsifying agents, pH buffering agents, and other substances such as, for example, sodium acetate, and triethanolamine oleate.

[234] In some embodiments this disclosure provides a method of synthesizing a compound described herein. The method may include steps of providing one or more coding sequences of **SEQ ID NOs: 67-483** in a suitable vector, or vectors, together with regulatory sequences which will drive expression of the coding sequences in a host cell. The vector, or vectors, are then provided to a host cell, such as for example a yeast cell, and the cells are grown under conditions that allow for the expression of the coding sequences. In some cases, a host cell may be provided with 1, 2, 3, 4, 5, 6,

or more than 6 different plasmids. The synthesized compound may be purified from the cell culture by centrifuging the cells to produce a cell pellet and a supernatant. The supernatant and cell pellet may be extracted using either ethyl acetate or acetone, or other suitable organic solvent. For compounds containing carboxylic acid groups, the pH of the supernatant may be adjusted to pH of 4 or less (e.g., 3) with an acid such as HCl prior to extraction. After extraction, both organic phases are combined and evaporated to dryness. The compounds may then be dissolved in the desired solvent and further purified using standard purification methods.

[235] Although the present disclosure has been described in certain specific aspects, many additional modifications and variations would be apparent to those skilled in the art. In particular, any of the various processes described above can be performed in alternative sequences in order to achieve similar results in a manner that is more appropriate to the requirements of a specific application. It is therefore to be understood that embodiments of the present disclosure can be practiced otherwise than specifically described without departing from the scope and spirit of the present disclosure. Thus, embodiments of the present disclosure should be considered in all respects as illustrative and not restrictive.

Example 1: Identification of production phase promoters

[236] Biological data supports the systems and constructs of production-phase promoter DNA vectors and applications thereof. Provided below are several examples of incorporating production-phase promoters into DNA vectors. Some of these vectors were used to produce biosynthetic products from multi-gene clusters derived from various fungal species. Compared to a constitutive promoter system, production-phase promoter systems in accordance with various embodiments produced several-fold greater product.

Production Phase Promoter Expression Analysis

[237] Because the ADH2 promoter (**SEQ ID NO. 1**) has properties of a production-phase promoter, a panel of promoter sequences was compared to the ADH2 promoter to identify other production-phase promoters. To begin, endogenous *S. cerevisiae* genes were identified that appeared co-regulated with ADH2 in a previous genome-wide transcription study (Z. Xu. *et al.*, *Nature* 457:1033-37, 2009, the disclosure of which is incorporated herein by reference). In this study, transcription of yeast genes was quantified during mid-exponential growth in several types of

growth media. Of the 5171 ORFs examined, 35 appeared co-regulated with ADH2, with co-regulation defined as a greater than two-fold increase in expression with a non-fermentable carbon source (ethanol in a yeast-peptone-ethanol (YPE) media) as compared to a fermentable carbon source (dextrose in a yeast-peptone-dextrose (YPD) media). Because these data were collected at a single time point and assessed transcription of genes in their native context, their ability to co-regulate heterologous genes in a production-phase promoter system required further validation and characterization.

[238] A detailed characterization of the ability of 34 selected promoters to control expression of heterologous genes was performed. For this specific purpose, a promoter was defined as the shorter of (a) 500 bp upstream of the start codon, or (b) the entire 5' intergenic region. Each promoter was cloned upstream of the gene for monomeric enhanced GFP (eGFP) and integrated each of the resulting cassettes in a single copy at the *ho* locus of individual strains. Control strains were included in which strong constitutive FBA1 and TDH3 promoters were cloned upstream of eGFP in an identical manner. The 35 promoter sequences can be found in **SEQ ID NOs. 2-35**.

[239] In order to compare the 35 putative production-phase promoters, the expression of eGFP protein was assessed over 72 hours in each strain by flow cytometry in media with both fermentable (YPD) and non-fermentable (YPE) carbon sources (**Figures 16 and 17**). All cultures were started in YPD media and analysis of eGFP expression began when cells were in the midst of exponential fermentative growth ($OD_{600} = 0.4$, 0 hrs). At this point, cells were either left to continue growth in YPD or spun-down and resuspended in YPE. Consistent with previous work, pADH2 was entirely repressed at the point where the experiment commenced (during exponential fermentative growth, 0 hrs) unlike the constitutive promoters pTDH3 and pFBA1, which were expressed at near maximum levels regardless of phase. Moderate expression from pADH2 was observed after a further 6 hours in YPD culture or following a growth media switch to YPE. Within 24 hrs, expression reached levels exceeding those observed in the strong constitutive systems. Cytometry histograms and fluorescence microscopy demonstrated that within 48 hours, >95% of all cells with pADH2 and pPCK1 driven expression were fluorescing above background (**Figure 18**). Protein expression levels spanned 15-50 fold, with most showing little or no expression until 24 hours into the culture (**Figures 16 and 17**). Transgene expression driven by the PCK1, MLS1, and ICL1 promoters (**SEQ ID NOs. 2-4**) not only showed the same timing of expression as pADH2, but also expressed at an

equivalently high level. The promoters of genes YLR307C-A, YGR067C, IDP2, ADY2, GAC1, ECM13 and FAT3 (SEQ ID NOs. 5-11) displayed semi-strong transgene expression (Figure 16). In addition, the promoters of genes PUT1, NQM1, SFC1, JEN1, SIP18, ATO2, YIG1, and FBP1 (SEQ ID NOs. 12-19) displayed weak of transgene expression (Figures 16 and 17). The promoter PHO89 (SEQ ID NO. 20) did not exhibit strong repression in during the growth phase (Figure 16, 0 and 6 hours). The results of the other sequences are also depicted in Figure 16 (SEQ ID NOs. 22-36). The constitutive promoters pTDH3 and pFBA1 (SEQ ID NOs. 50 and 52) were used as controls (Figures 16-18).

[240] The above analysis identified a large set of co-regulated promoters spanning a wide range of expression levels, three of which were as strong as pADH2. However, a more extensive set of strong production-phase promoters is desirable for assembly of constructs having multi-gene pathways, especially pathways having more than four genes. To identify other production-phase promoter candidates, the genomes of five closely related species within the *S. sensu stricto* complex were examined (Figure 19). The promoter region was identified for the closest ADH2 gene homolog in the genomes of *Saccharomyces bayanus*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, and *Saccharomyces castellii*. Multiple sequence alignment of the upstream activation sequences (UAS) revealed that nearly all sequences (except that from *S. castellii*) are highly conserved across this region, suggesting a potential for regulation similar to that of *S. cerevisiae* ADH2 (Figure 20, SEQ ID NOs. 36-40). In order to be used for single-step pathway assembly, all promoter sequences must be sufficiently unique to prevent undesired recombination between each other. Therefore, the pairwise identities for each of the *Saccharomyces sensu stricto* ADH2 promoter pairs were analyzed (Figure 21). The most similar promoter to the *S. cerevisiae* ADH2 promoter is that from *S. paradoxus*, with 83% identity, including a single 40 bp stretch located near the center of the promoter. This homology is significantly less than the 50-100 bp typically used for assembly by yeast homologous recombination, and recombination events between sequences with this level of identity occur at very low frequency, suggesting that these promoters should be compatible with a multi-gene assembly technique utilizing yeast homologous recombination as described above.

[241] As with the endogenous yeast promoter candidates, these other putative *Saccharomyces* promoters required detailed characterization of induction profiles. DNA encoding each of these

promoter sequences was obtained by commercial synthesis and characterized expression of eGFP from each promoter in the same manner as the endogenous yeast promoters (**Figures 22 and 23**). Of the five *Saccharomyces sensu stricto* pADH2s tested (**SEQ ID NOs. 36-40**), the promoters derived from *S. paradoxus*, *S. kudriavzevii*, and *S. bayanus* show timing and strength of expression equivalent to that of *S. cerevisiae* pADH2. In combination with the endogenous yeast promoters, these three additional *Saccharomyces* pADH2s expand the number of strong promoters with the desired induction profile.

Expression of Compound Product Pathways Using the Production-Phase Promoter System

[242] To study the utility of the new promoter set for heterologous expression of a biosynthetic system, production of fungal-derived dehydrozearalenol (1) and indole-diterpene (2) was examined (**Figure 24**, Compounds 1 & 2). The biosynthesis of the indole-diterpene compound resulted from the coordinated expression of four in *Aspergillus tubingensis* genes (**Figure 25**, **SEQ ID NOs. 59-62**). Two versions of each pathway were constructed: one having all production-phase promoters, and the other having all constitutive promoters (**Figure 24**). The production-phase promoter system utilized the pADH2 from *S. cerevisiae* (**SEQ ID NO. 1**), pADH2 from *S. bayanus* (**SEQ ID NO. 38**), and pPCK1 (**SEQ ID NO. 2**) and pMLS1 (**SEQ ID NO. 3**) from *S. cerevisiae*. In the constitutive system, transcription was driven by four frequently used strong constitutive promoters: pTEF1, pFBA1, pPCK1, and pTPI1 (**SEQ ID NOs. 51-54**). Each indole-diterpene system was constructed on a single plasmid harboring four expression cassettes: promoter::GGPPS::tADH2; promoter::PT::tPGI1; promoter::FMO::tENO2; and promoter::Cyc::tTEF1; wherein, the promoter sequences corresponded to either the production-phase or the constitutive promoters (**Figure 24**). Similar constructs were built for the dehydrozearalenol compound with the two genes HR-PKS and NR-PKS (**SEQ ID NOs. 63 and 64**). All plasmids were constructed using yeast homologous recombination. It should be noted that pADH2 sequences from *S. cerevisiae* and *S. bayanus* (61% identity) are sufficiently unique for this type of assembly. The production of compounds 1 and 2 produced by *S. cerevisiae* BJ5464/*npgA*/pRS424 transformed with each of these plasmids were measured over seventy-two hours in YPD batch culture (**Figure 26**). An 80-fold and 4.5-fold increase in titer of compound 1 and 2 was observed for the system using the production-phase promoters as compared to the constitutive system.

Materials and Methods Supporting the Production-Phase Promoter Experiments

[243] General techniques, reagents, and strain information: Restriction enzymes were purchased from New England Biolabs (NEB, Ipswich, 25 MA). Cloning was performed in *E. coli* DH5 α . PCR steps were performed using Q5® high-fidelity polymerase (NEB). Yeast dropout media was purchased from MP Biomedicals (Santa Ana, CA) and prepared according to manufacturer specifications. Promoter characterization experiments were performed in BY4741 (*MAT α , his3 Δ 1 leu2 Δ .0 met15 Δ .0 ura3 Δ 0*) while all experiments involving the production of 1 were performed in BJ5464-*npgA* which is BJ5464 (*MAT α ura3-52 his3 Δ 200 leu2 Δ .1 trp1 pep4::HIS3 prb1. Δ .1.6R can1 GAL*) with two copies of pADH2-*npgA* integrated at δ elements. All Gibson assemblies were performed as previously described using 30 bp assembly overhangs.

[244] Construction and characterization of promoter-eGFP reporter strains: All promoters were defined as the shorter of 500 base pairs upstream of a gene's start codon or the entire 5' intergenic region. All promoters from *S. cerevisiae* were amplified from genomic DNA, while ADH2 promoters from all *Saccharomyces sensu stricto* were ordered as gBlocks from Integrated DNA Technologies (IDT, Coralville, Iowa). Minimal alterations were made to promoters from *S. kudriavzevii* and *S. mikatae* in order to meet synthesis specifications. In all constructs, eGFP was cloned directly upstream of the terminator from the CYC1 gene (tCYC1). pRS415 was digested with SacI and Sall and a NotI-eGFP-tCYC1 cassette was inserted by Gibson assembly generating pCH600. Digestion of pCH600 with AclI and PmlI removed the CEN/ARS origin, which was replaced by 500 bp sequences flanking the ho locus using Gibson assembly to yield plasmid pCH600-HOint. Each of the promoters to be analyzed was amplified with appropriate assembly overhangs and inserted into pCH600-HOint digested with NotI to generate the pCH601 plasmid series. Digestion of the pCH601 plasmid series with AclI generated linear integration cassettes which were transformed into *S. cerevisiae* BY4741 by the LiAc/PEG method. Correct integration was confirmed by PCR amplification of promoters and Sanger sequencing.

[245] For characterization, all strains were initially grown to saturation overnight in 100 μ l of YPD media. These cells were then reinoculated at an OD₆₀₀ of 0.1 into 1ml of fresh YPD and allowed to grow to OD₆₀₀ = 0.4 to reach mid-log phase growth (approximately 6hrs). 500 μ l of each culture was pelleted by centrifugation and resuspended in YPE broth for YPE data while the remaining 500 μ l was used for YPD data. The 0 hour time point was collected immediately after resuspension. For

each time point, 10 µl of culture was diluted in 2 ml of DI water and sonicated for three short pulses at 35% output on a Branson Sonifier. Expression data were collected for 10,000 cells using a FACSCalibur flow cytometer (BD Bioscience) with the FL1 detector. Data were analyzed in R using the flowCore package.

[246] Construction of plasmids to produce compounds in *S. cerevisiae*: The sequences for genes assembled on IDT producing plasmids are contained in the supporting information. Regulatory cassettes of promoters and terminators were fused using overlap extension PCR. All genes and regulatory cassettes were amplified by PCR, ensuring 60 bases of homology between all adjacent fragments. 500 ng of each purified fragment was combined with 100 ng of pRS425 linearized with *NotI* and transformed into *S. cerevisiae* BJ5464/*npG*A. Sixteen clones were picked from each assembly plate and grown to saturation in 5 ml CSM-Leu medium. Plasmids were isolated, transformed into *E. coli* and purified prior to sequence confirmation using the Illumina MiSeq platform. Detailed plasmid maps for pCHIDT-2.1 and pCHIDT-2c are shown in **Figure 27A** illustrates the primers used and the assembly strategy (**SEQ ID NOs. 65 and 66**).

[247] Examining the productivity of indole diterpene generating systems Plasmids pCHIDT-2.1 and pCHIDT-2c were transformed into BJ5464/*npG*A with pRS424 as a source of tryptophan overproduction (see, e.g., **Figure 27B**). Triplicates of each strain were inoculated into CSM – Leu/-Trp medium and grown overnight ($OD_{600} = 2.5-3.0$). Each culture was used to inoculate 20 ml cultures in YPD medium at an $OD_{600} = 0.2$ and incubated with shaking at 30°C for 3 days. Every 24 hrs, 2 mls were sampled from each culture. Supernatants were clarified by centrifugation and extracted with 2 ml ethyl acetate (EtOAc). Cell pellets were extracted with 2 ml 50% EtOAc in acetone. 500 µl each of pellet and supernatant extracts were combined and dried *in vacuo*. Samples were resuspended in 100 µl HPLC grade methanol and LC-MS analysis was conducted on a Shimadzu LC-MS-2020 liquid chromatography mass spectrometer with a Phenomenex Kinetex C18 reverse-phase column (1.7 µm, 100 Å, 100 mm x 2.1 mm) with a linear gradient of 15% to 95% acetonitrile (v/v) in water (0.1% formic acid) over 10 min followed by 95% acetonitrile for 7 min at a flow rate of 0.3 mL/min.

Example 2: Identification of gene clusters which produce compounds that interact with a target protein

[248] Thanks to next-generation sequencing, thousands of bacterial and fungal genomes have been sequenced. These species are known to be rich sources of secondary metabolites, for example penicillin, rapamycin, and the statins. These secondary metabolites are small molecules, enzymatically synthesized by the products of one or more genes, often arrayed contiguously in a “biosynthetic gene cluster”.

[249] This disclosure describes a method for identifying specific biosynthetic gene clusters, where the target of the secondary metabolite is a specific protein, and expressing that secondary metabolite in a host organism.

[250] In certain cases, for example, when a secondary metabolite is being used as a weapon against other organisms, the secondary metabolite may also be toxic to the organism that produces it. In these cases, the producing organism may defend itself against self-harm in a number of ways: by pumping the secondary metabolite out of the cell; by enzymatically negating the secondary metabolite; or by producing an additional version of the target protein that is less sensitive or insensitive to the secondary metabolite.

[251] In those cases where the organism produces an additional version of the target protein, this “protective” version of the gene is often colocalized with the biosynthetic gene cluster. Although different to the gene that produces the target protein, the protective version should maintain detectable homology to the target protein. This method takes advantage of this homology to identify those biosynthetic gene clusters that contain or are adjacent to a protective homolog of the target protein.

[252] The input data required for this method are a list of biosynthetic gene clusters (e.g., polyketide synthase clusters, non-ribosomal peptide synthetase clusters) and a list of target proteins (i.e., proteins whose activity are to be modulated with secondary metabolites). The biosynthetic clusters may be identified based on the presence of certain protein domains, for example by the software program antiSMASH. The target proteins may be chosen based on their quantitative likelihood of being drug targets.

[253] The output of this method is a score for each biosynthetic cluster, where clusters with higher scores represent those that are more likely to produce secondary metabolites targeting specific proteins of interest.

[254] A score was constructed for each biosynthetic cluster based on the following factors:

1. the presence of one or more homologs of a target protein within or adjacent to the cluster, as determined by a homology search (e.g., using the tblastn algorithm, with a maximum score granted when one homolog is found)
2. the confidence in homology of the target to genes in a cluster (e.g., according to the tblastn algorithm, with an increasing score for lower e-values, and an upper bound threshold of $1e-30$)
3. the fraction of the homologous gene that meets a certain threshold of identity (e.g., with an increasing score for more identity, and a lower bound threshold of 25% identity)
4. the total number of genes homologous to the target protein present in the entire genome of the organism (e.g., with a maximum score granted to cases with 2-4 homologs per genome)
5. the homology of the gene in or adjacent to the cluster to the target protein (e.g., using the blastx algorithm, with a maximum score granted when the gene in the biosynthetic gene cluster's closest homolog in the target protein's genome is the target protein itself)
6. the phylogenetic relationship of the target protein to the gene in the cluster (e.g., with an increasing score for homologs in the gene cluster that clade with the target protein, with confidence assigned by a bootstrap test or Bayesian inference of phylogeny, and a lower bound threshold defined as homologs in a phylogenetic context that appear in a clade with bootstrap value of 0.7 or Bayesian posterior probability of 0.8)
7. the expected number of homologs of the target in or adjacent to the biosynthetic cluster (e.g., with a greater score the lower the probability of a homolog of the target being present in or adjacent to a biosynthetic cluster of a certain size, given the number of total homologs in the genome, as determined by a permutation test)

8. the likelihood that the target protein is essential for viability, growth, or other cellular processes in the native environment, e.g., through evidence that deletion of homologs in related organisms (such as *S. cerevisiae*) render the organism inviable
 9. synteny of the gene cluster with related species (e.g., with a maximum score if the entire cluster, including the target homolog, is conserved across several species)
 10. the functional class of the target homolog (e.g., with a greater score if the gene is in a protein complex already known to be targeted by secondary metabolites)
 11. the presence of specific promoters adjacent to the target homolog (e.g., with a greater score when there is a bidirectional promoter upstream of the target homolog and a biosynthetic gene)
 12. the presence of specific regulatory elements in the biosynthetic gene cluster (e.g., with a greater score when there is a transcription factor binding site that is shared between target genes and/or biosynthetic genes in the cluster)
 13. the presence of target homologs outside the cluster (including on other chromosomes) that are co-regulated with some or all of the genes in the biosynthetic cluster (e.g., with a greater score when biosynthetic gene clusters are co-regulated with putative target homologs)
 14. the presence of protein- and DNA-sequence-derived features within the clusters that have successfully been shown to produce secondary metabolites (e.g., with a greater score when a gene in a cluster shares a domain — as determined by a Hidden Markov Model (HMM) — with a cluster that has produced a secondary metabolite in one of the host organisms)
- [255] The above score was calibrated with reference to a set of “true positives” (i.e., cases where there are one or more known targets in or adjacent to a biosynthetic gene cluster that produces a small molecule known to target that protein).
- [256] This algorithm has been programmed in the Python programming language and has been applied to a set of more than 1,000 fungal genomes (and more than 10,000 biosynthetic gene clusters) to produce a list of potentially relevant biosynthetic clusters.

Expressing biosynthetic clusters in a host organism

[257] Given the highest scoring biosynthetic clusters as defined by the algorithm above, DNA was synthesized for each of the genes in those clusters. The DNA was cloned into a host organism (e.g., *S. cerevisiae*, also known as baker's yeast) for expression. The host organism synthesized the proteins from the gene cluster, which produces a secondary metabolite. Using HPLC and mass spectrometry, the secondary metabolite-expressing strain can be compared to an unmodified strain and affirm the presence of a new secondary metabolite.

[258] This method has been successfully applied to the production of several secondary metabolites where there is evidence, based on the above method, for what the target protein of the secondary metabolite should be:

- An secondary metabolite derived from a biosynthetic gene cluster containing a homolog of the human gene SOS1;
- An secondary metabolite derived from a biosynthetic gene cluster containing a homolog of the human gene BRSK1; and
- An secondary metabolite derived from a biosynthetic gene cluster containing a homolog of the human gene DDX41.

[259] A further example of a gene cluster which produces a product, for which there is evidence suggesting the target, is shown in **Figure 15A**.

Example 3: Prioritization of novel biosynthetic gene clusters by phylogenetic analysis

[260] Two classes of fungal BGCs; those with either a polyketide synthase (PKS), or an UbiA-type sesquiterpene cyclase (UTC) as their core enzyme were chosen for analysis.

[261] A computational pipeline was developed to prioritize PKS and UTC containing BGCs for heterologous expression. 581 sequenced fungal genomes were analyzed from the publicly available GenBank database of the National Center for Biotechnology Information (NCBI, as of July 2015). Each genome was analyzed for BGCs using antiSMASH2, identifying 3512 BGCs harboring an iterative type 1 PKS (iPKS) and 326 BGCs harboring a UTC homologue. Phylogenetic trees of each of these enzyme types were generated with identified characterized homologs from the MIBiG database²⁰. BGCs were primarily selected from clades having few characterized members (**Figure 28A, Figure 29A**). The selected BGCs were found in the genomes of both ascomycetes and basidiomycetes. Basidiomycetes are, in general, more difficult to culture with fewer tools for genetic

manipulation available as compared to ascomycetes. As a result, BGCs from basidiomycetes are under-studied, with few PKS-containing clusters deposited in MIBig, suggesting that these organisms represent a reservoir of BGCs capable of producing compounds with interesting new structures.

[262] The coding sequences of all BGCs were ordered as synthetic constructs according to the default predicted gene coordinates (start, stop, and introns) as deposited in GenBank and clusters are described in **Table 4**.

[263] Shown in **Figure 28A** is a cladogram of the ketosynthase sequences of the 3512 iPKS sequences identified in this study. Of these, 28 were selected and the associated BGC containing the selected iPKS was analyzed using heterologous expression. Selected BGCs met the following criteria: (a) genetic structure was conserved across 3 or more species, (b) exhibited canonical domain architecture, and (c) contained an in- cis or proximal in- trans protein capable of releasing the polyketide from the carrier protein of the PKS (**Figure 30A**). Seven of these clusters were derived from distinct clades comprised entirely of sequences from basidiomycetes (**Figure 28A**).

[264] The 28 selected PKS clusters were edited according to the methods described here to form expression vectors suitable for expression of the cluster coding sequences in yeast cells. The host cells were incubated and analyzed for the presence of novel chemical compounds by HPLC, as described in the methods section below. Of the PKS clusters selected from ascomycetes, 13 produce compounds. The most notable is the PKS1 cluster, which only contains an iPKS, a hydrolase, and the genes for three tailoring enzymes: a Cytochrome p450 (P450), a Flavin-dependent monooxygenase (FMO), and a Short-chain dehydrogenase/reductase (SDR).

[265] For the study of fungal UTCs, the phylogenetic tree shown in **Figure 29A** was constructed based on the UbiA-type sesquiterpene cyclase, Fma-TC, from the fumagillin biosynthetic pathway. Moreover, the P450, Fma-P450, from the same pathway was shown to be a powerful enzyme catalyzing the 8 e oxidation of bergamotene to generate a highly oxygenated product. UTC BGCs spanning the entirety of the cladogram were selected in **Figure 29A** where a cytochrome P450 was proximal to the UTC gene (**Figure 30B**). Ultimately, 13 UTC BGCs from both ascomycetes and basidiomycetes were selected for analysis.

[266] Screening of strains expressing these clusters by LC/HRMS revealed novel spectral features consistent with oxidized sesquiterpenoids being produced by five clusters (**Figure 29A**). These

results demonstrate that the membrane-bound UTCs represent a general class of terpene cyclase encoded by the genomes of diverse fungi. Several clusters and compounds produced are shown in **Figures 28B, 28C, 28D, 28E and 28F**.

[267] Including both PKS and UTC BGCs, 24 of the 41 clusters produced measurable compounds, see **Table 4** for a summary of the type, species of origin, and productivity of the clusters. Gene annotation errors introduced by incorrect intron prediction may have contributed to this failure rate. Manual inspection of one UTC (TC5) that initially had yielded no products suggested an incorrect intron prediction at the 5' terminus of the gene. Correction of this intron led to a C-terminal protein sequence that aligned well with known functional UTCs. When tested by heterologous expression in a host cell, the version with the corrected intron produced a compound confirming that incorrect intron prediction is a failure mode in approaches that rely on publicly available gene annotations, (**Figure 29B**). These results illustrate the importance of careful gene curation and the need for improved eukaryotic gene prediction, particularly with sequences from taxa with few well-studied members.

[268] The results summarized in **Table 4** demonstrate the utility of the methods herein for the selection of cryptic fungal BGCs. With the tools developed here, strains were built expressing 41 such clusters with 22 (54%) producing detectable levels of products not native to *S. cerevisiae*. While both basidiomycetes and ascomycetes are known to be prolific producers of bioactive compounds, to date, the bulk of research on the biosynthesis of fungal natural products has been undertaken in ascomycetes. In this study, heterologous expression allowed a large-scale survey of cryptic fungal BGCs from both ascomycetes and basidiomycetes, a less studied and more difficult to culture division of fungi with fewer tools for genetic manipulation. Using this platform, a panel of new products produced by the selected PKS and UTC clusters was identified.

Methods

[269] antiSMASH2 software was applied to 581 public fungal genomes deposited in the Genbank database of the National Center for Biotechnology Information (NCBI), to search for type 1 PKS and UbiA-like terpene cyclase gene clusters. This analysis identified 3,512 type 1 PKS gene clusters and 326 UbiA-like terpene gene clusters in 538 fungal genomes.

[270] Phylogenetic analysis of both sequence sets was performed by building multiple sequence alignments of all protein sequences using MAFFT and building phylogenetic trees as shown in **Figure 28A** and **Figure 29A** using FastTree 2.

[271] 28 of the 3,512 sequenced type 1 PKS gene clusters and 13 of the 326 terpene gene clusters were selected for expression in yeast as described above.

Construction and culture of production strains:

[272] Production strains were constructed by transforming plasmid DNA isolated out of *E. coli* (Qiagen miniprep 27106) into the appropriate expression host (JHY692 for PKS containing plasmids, JHY705 for all others) using the Frozen-EZ Yeast Transformation II kit (Zymo Research T2001) followed by plating on the appropriate SDC dropout media (CSM –Leu for PKS containing plasmids, CSM –Ura for all others). For BGCs encoded on at least two plasmids, three biological replicates for each haploid transformant were mated on YPD plates and incubated at 30°C for 4-16 hrs prior to streaking for single colonies on CSM –Ura/-Leu and incubated at 30°C.

[273] Small-scale cultures for analysis were begun by picking three biological replicates of each production strain along with empty vector controls into 500 µL of the appropriate SDC dropout medium in a 1 ml deep-well block and grown for approximately 24 hrs at 30°C. 50 µL of overnight culture was used to inoculate 500 µL of each of the production media to be tested in the experiment (generally both YPD and YPEG) in 1 ml deep well blocks. All blocks were covered with gas-permeable plate seals (Thermo Scientific AB-0718) and incubated at 30°C for 72 hrs with shaking at 1000 rpm. Supernatants were clarified by centrifugation for 20 mins at 2800 g and a minimum of 100 µL of clarified supernatant was stored for future analysis. The remainder of the supernatant was discarded and the cell pellets extracted by mixing with 400 µL of 1:1 ethyl acetate:acetone. Cell debris was precipitated by centrifugation for 20 mins at 2800 g and 200 µL of the extraction solvent pipetted to a fresh block and evaporated in a speedvac.

[274] Prior to analysis, all supernatants were passed through a 0.2 µm filter plate while all cell pellet extracts were resuspended in 200 µL of HPLC grade methanol prior to filtering.

Analysis of small scale cultures:

[275] LC-MS analysis was conducted on an Agilent 6545 quantitative time-of-flight mass spectrometer interfaced to an Agilent 1290 HPLC system. The ion source for most analyses was an electrospray ionization source (dual-inlet Agilent Jet Stream or “dual AJS”). In some analyses, an

Agilent Multimode Ion Source was also used for atmospheric pressure chemical ionization. The parameters used for both ionization sources are outlined in **Table 5**.

[276] The HPLC column for all analyses was a 50 mm × 2.1 mm Zorbax RRHD Eclipse C18 column with 1.8 µm beads (Agilent, 959757-902). No guard column was used.

[277] Gradient conditions were isocratic at 95% A from 0 to 0.2 min, with a gradient from 95% A to 5% A from 0.2 to 4.2 minutes, followed by isocratic conditions at 5% A from 4.2 to 5.2 minutes, followed by a gradient from 5% A to 95% A from 5.2 to 5.2 minutes, followed by isocratic re-equilibration at 95% A from 5.2 to 6 minutes. For electrospray analyses, A was 0.1% v/v formic acid in water and B was 0.1 % v/v formic acid in acetonitrile. For APCI analyses, B was substituted by 0.1% v/v formic acid in methanol.

[278] Data analysis by untargeted metabolomics was performed with xcms, using optimal parameters determined by IPO25. For PKS containing clusters, automated analyses were set to generate extracted ion chromatograms (EICs) for the top 50 spectral features as defined by both fold-change and p-value. These EICs were then manually inspected to identify the subset of automatically identified features that appear specific to the expressed BGC as defined by presence in each of three biological replicates of the production strain and absence from three biological replicates of a negative control strain (**FIG. 31**). EICs of all BGC specific features are illustrated in **FIGS. 32-49**.

Example 4: Construction of Yeast strains

[279] In the current example, yeast strains are based on the BY4741/BY4742 background, which is in turn based on S288c (C. B. Brachmann, et al., 1998, cited supra). The strains were made in two stages: 1) creation of a core DHY set with restored sporulation and mitochondrial genome stability and 2) creation of JHY derivatives modified for other benefits, which may include protein production. All changes introduced in this study were confirmed by diagnostic PCR and sequencing.

[280] A sporulation-restored strain set was built by crossing BY4710 (C. B. Brachmann, et al., 1998, cited supra) to a haploid derivative of YAD373 (A. M. Deutschbauer and R. W. Davis, *Nat. Genet.* 37:133-40, 2005), a BY-based diploid that contains three QTLs that restore sporulation: *MKT1*(30G), *RME1*(INS-308A), and *TAO3*(1493Q). A spore clone from the resulting diploid was repaired for *HAP1*, which encodes a zinc-finger transcription factor localized to mitochondria and the nucleus. *HAP1* is important for mitochondrial genome stability (see J. R. Matoon, E. Caravajal, and D. Gurthrie *Curr. Genet.* 17:179-83, 1990) and likely also important for sporulation. S288c and derivatives contain a Ty1 insertion in the 3' end of *HAP1* that inactivates function. The transposon was excised using the Delitto Perfetto method (F. Storici and M. A.

Resnick, *Methods Enzymol.* 409:329-45, 2006) and confirmed repaired *HAP1* function based on transcription of a *CYC1p-lacZ* reporter (M. Gaisne, et al., *Curr. Genet.* 36:195-200, 1999). The sporulation-restored, *HAP1*-repaired strain and its auxotrophic and prototrophic derivatives were then used to create the DHY set of strains that were additionally restored for mitochondrial genome stability.

[281] The above sporulation-restored strains were used to repair the poor mitochondrial genome stability known to be a problem with S288c and BY derivatives. Mitochondrial genome stability is likely to improve growth and *ADH2p*-like gene expression under conditions of respiration, and for reducing the frequency of petite cells (slow-growing, respiration-defective cells that cannot grow on non-fermentable carbon sources). For a detailed description of the “mito-repair” method, see construction of JHY650 (J.D. Smith, 2017, cited supra). Briefly, the 50:50 genome editing method was used to introduce the wild-type alleles of three genes shown to be important for mitochondrial genome stability by QTL analysis³¹. The repaired QTLs are: *SAL1*⁺ (repair of a frameshift), *CAT5*(91M) and *MIP1*(661T). Crosses with prototrophic and auxotrophic strains completed the DHY core set of about a dozen sporulation and mitochondrial genome stability restored strains that can be further modified as needed. DHY213 (see **Table 3**) is one such strain: it contains the seven desired changes described above, is otherwise congenic with BY4741, and was used in this study to create derivatives for the HEx platform (see **Table 3**).

[282] Marker-free, seamless deletion of the complete *PRB1* and *PEP4* ORFs was performed using the 50:50 method (J. Horecka and R. W. Davis, 2014, cited supra). Integration of a 1609 bp *ADH2p-npgA-ACS1t* expression cassette on the chromosome was performed using a similar method used to integrate DNA segments with the REDI method (J. D. Smith, et al., 2017, cited supra), except that *URA3*, not *FCY1*, was used as the counter-selectable marker. For an integration site, an 1166 bp cluster of three transposon LTRs located centromere-distal to YBR209W on chromosome II was replaced (deletion of chrII 643438 to 644603). Two DNA segments were simultaneously inserted via homologous recombination at the integration site that had been cut with *SceI* to create double strand breaks. One inserted segment was *ADH2p-npgA* (1448 bp) PCR amplified from a BJ5464/*npgA* expression strain (*npgA* from *A. nidulans*) (K. K. M. Lee, N. A. Da Silva, and J. T. Kealey, *Anal. Biochem.*, 394:75-80, 2009). The *npgA* 3' end was repaired to wildtype using a reverse PCR primer that replaced the *npgA* intron included previously with the wildtype *npgA* 3' sequence. To preclude recombination of the expression cassette with the native *ADH2* locus, the 161 bp *ACS1* terminator was used as the second DNA segment (not *ADH2t*) and PCR amplified from BY4741. The resulting strain (JHY692) was used in a similar fashion to replace only *npgA* with the *CPR* ORF (cytochrome P450 reductase, ATEG_05064 from *A. terreus*). Finally, a strain with both *npgA* and *CPR* expression cassettes (JHY702) was created by mating JHY692 and JHY705.

Example 5: Determination of Chemical Structures

[283] For compound isolation, large-scale fermentation was carried out with the strains and clusters of Example 3. The yeast strains were first struck out onto the appropriate SDC dropout agar plates and incubated for 48 hrs at 30°C. A colony was then inoculated into 40 mL SDC dropout medium and incubated at 28°C for two days with shaking at 250 rpm. This seed culture was used to inoculate 4 L of YPD medium (1.5% Glucose) and cultured for 3 days at 28°C and 250 rpm. Supernatants were then clarified by centrifugation and extracted with equal volume of ethyl acetate. Cell pellets were extracted with 1 L of acetone. For compounds containing carboxylic acid groups, the pH value of the supernatant was adjusted to 3 by adding HCl prior to extraction. The organic phases were combined and evaporated to dryness. The residue was purified by ISCO-CombiFlash® Rf 200 (Teledyne Isco, Inc) with a gradient of hexane and acetone. After analysis by LC-MS, the fractions containing the target compounds were combined and further purified by semi-preparative HPLC using C18 reverse-phase column. The purity of each compound was confirmed by LC-MS, and the structure was solved by NMR (**Figure 49-59**).

[284] All NMR spectra including ¹H, ¹³C, COSY, HSQC, HMBC and NOESY spectra were obtained on Bruker AV500 spectrometer with a 5 mm dual cryoprobe at the UCLA Molecular Instrumentation Center. The NMR solvents used for these experiments were purchased from Cambridge Isotope Laboratories, Inc.

Table 4: Summary of control and cryptic fungal BGCs examined in this study.

Cluster ID	Type	Native Locus				Species of origin	Division	Productive?
		Genbank ID	Start	End	Length			
IDT	Ctl					<i>Aspergillus tubingensis</i>	Ascomycota	Y
DHZ	Ctl					<i>Hypomyces subiculosus</i>	Ascomycota	Y

PKS1	PKS	KV441552	530394	54672 3	16329	<i>Coniothyrium sporulosum</i>	Ascomycota	Y
PKS2	PKS	KV441551	60641	83767	23126	<i>Coniothyrium sporulosum</i>	Ascomycota	Y
PKS3	PKS	Deposition pending			9892	<i>Acremonium Sp. KY4917</i>	Ascomycota	N
PKS4	PKS	AM270992	578654	60336 7	24713	<i>Aspergillus niger</i>	Ascomycota	Y
PKS5	PKS	CP003009	873083 1	87535 60	22729	<i>Thielavia terrestris</i>	Ascomycota	N
PKS6	PKS	ABDF0200 0052	9685	28459	18774	<i>Trichoderma virens</i>	Ascomycota	Y
PKS7	PKS	JPJY01000 093	2671	30026	27355	<i>Pseudogymno ascus pannorum</i>	Ascomycota	N
PKS8	PKS	JOWA010 00110	160785 7	16432 05	35348	<i>Scedosporium apiospermum</i>	Ascomycota	Y
PKS9	PKS	KE384750	270098	29275 5	22657	<i>Metarhizium anisopliae</i>	Ascomycota	N
PKS10	PKS	KB445572	91380	11519 9	23819	<i>Cochliobolus heterostrophu s</i>	Ascomycota	Y
PKS11	PKS	JPKB0100 1000	12749	37136	24387	<i>Pseudogymno ascus pannorum</i>	Ascomycota	N
PKS12	PKS	JPJU01000 852	23274	38823	15549	<i>Pseudogymno ascus pannorum</i>	Ascomycota	N

PKS13	PKS	JPJR01000 396	848	19416	18568	<i>Pseudogymno ascus pannorum</i>	Ascomycota	Y
PKS14	PKS	KN847553	300827	32648 9	25662	<i>Verruconis gallopava</i>	Ascomycota	Y
PKS15	PKS	AWSO010 00045	179446	20357 7	24131	<i>Moniliophthor a roreri</i>	Basidiomycot a	Y
PKS16	PKS	JH687542	431482	46533 2	33850	<i>Punctularia strigosoazonata</i>	Basidiomycot a	Y
PKS17	PKS	KN839868	143119	18801 3	44894	<i>Hydnomeruliu s pinastri</i>	Basidiomycot a	Y
PKS18	PKS	DS989828	138944 9	14074 99	18050	<i>Arthroderma gypseum</i>	Ascomycota	Y
PKS19	PKS	KB908593	409171	44306 9	33898	<i>Setosphaeria turcica</i>	Ascomycota	N
PKS20	PKS	GL532685	4724	17845	13121	<i>Pyrenophora teres</i>	Ascomycota	Y
PKS21	PKS	AMGW01 000002	129425 1	13220 17	27766	<i>Cladophialop hora yegresit</i>	Ascomycota	N
PKS22	PKS	DF933843	225991	25430 7	28316	<i>Talaromyces cellulolyticus</i>	Ascomycota	Y
PKS23	PKS	KE720645	48853	80314	31461	<i>Endocarpon pusillum</i>	Ascomycota	Y
PKS24	PKS	DF933834	523551	55801 8	34467	<i>Talaromyces cellulolyticus</i>	Ascomycota	Y

PKS25	PKS	AWSO010 00633	5071	30440	25369	<i>Moniliophthora roreri</i>	Basidiomycot a	N
PKS26	PKS	KN817529	130968	15224 0	21272	<i>Hypholoma sublateritium</i>	Basidiomycot a	N
PKS27	PKS	KB445800	136224 8	14033 85	41137	<i>Ceriporiopsis subvermispora</i>	Basidiomycot a	N
PKS28	PKS	AWSO010 00632	8804	37857	29053	<i>Moniliophthora roreri</i>	Basidiomycot a	Y
TC1	UTC	ABDF0200 0086	384012	36886 8	15144	<i>Trichoderma Virens</i>	Ascomycota	Y
TC2	UTC	ABDF0200 0083	37740	49247	11507	<i>Trichoderma Virens</i>	Ascomycota	N
TC3	UTC	FQ790293	97750	13129 6	33546	<i>Botrytonia cinerea</i>	Ascomycota	Y
TC4	UTC	JH717969	858521	86922 6	10705	<i>Formitiporia mediterranea</i>	Basidiomycot a	Y
TC5	UTC	KI925459	241299 2	24323 65	19373	<i>Heterobasidio n annosum</i>	Basidiomycot a	Y
TC6	UTC	KB445798	581049	61882 3	37774	<i>Ge/atoporia subvermispora</i>	Basidiomycot a	N
TC7	UTC	JH719450	83620	11427 0	30650	<i>Dichomitius squalens</i>	Basidiomycot a	N
TC8	UTC	KL198014	171843 8	17465 40	28102	<i>Pleurotus ostreatus</i>	Basidiomycot a	N
TC9	UTC	GL377319	205808	21293 5	7127	<i>Schizophyllum commune</i>	Basidiomycot a	Y

TC10	UTC	JH687394	99692	11428 1	14589	<i>Stereum hirsutum</i>	Basidiomycot a	N
TC11	UTC	JH687396	33983	48346	14363	<i>Stereum hirsutum</i>	Basidiomycot a	N
TC12	UTC	JH719415	280135	30047 0	20335	<i>Dichomitus squalens</i>	Basidiomycot a	N
TC13	UTC	JH795868	73132	97338	24206	<i>Dacryopinax primogenitus</i>	Basidiomycot a	N
							Total	43
							Productive	24

Table 5: Ion source parameters used in this study

Ion source parameter	Dual AJS	MMI
gas temperature	250 °C	350 °C
drying gas	12 L / min	7.5 L / min
nebulizer	10 psig	20 psig
sheath gas temp.	400 °C	-
sheath gas flow	12 L / min	-
vaporizer	-	250 °C
capillary voltage (V _{cap})	3500 V	1500 V
nozzle voltage	1400 V	-
corona discharge	-	4 µA
fragmentor	100 V	120 V
skimmer	50 V	50 V
octopole 1 RF V _{pp}	750 V	750 V
charging voltage	-	1000 V

Table 6: Description of promoter sequences

SEQ ID NO.	Description
1	S. cerevisiae pADH2
2	S. cerevisiae pPCK1

3	<i>S. cerevisiae</i> pMLS1
4	<i>S. cerevisiae</i> pICL1
5	<i>S. cerevisiae</i> pYLR307C-A
6	<i>S. cerevisiae</i> pYGR067C
7	<i>S. cerevisiae</i> pIDP2
8	<i>S. cerevisiae</i> pADY2
9	<i>S. cerevisiae</i> pGAC1
10	<i>S. cerevisiae</i> pECM13
11	<i>S. cerevisiae</i> pFAT3
12	<i>S. cerevisiae</i> pPUT1
13	<i>S. cerevisiae</i> pNQM1
14	<i>S. cerevisiae</i> pSFC1
15	<i>S. cerevisiae</i> pJEN1
16	<i>S. cerevisiae</i> pSIP18
17	<i>S. cerevisiae</i> pAT02
18	<i>S. cerevisiae</i> pYIG1
19	<i>S. cerevisiae</i> pFBP1
20	<i>S. cerevisiae</i> PHO89
21	<i>S. cerevisiae</i> CAT2
22	<i>S. cerevisiae</i> CTA1
23	<i>S. cerevisiae</i> ICL2
24	<i>S. cerevisiae</i> ACS1
25	<i>S. cerevisiae</i> PDH1
26	<i>S. cerevisiae</i> REG2
27	<i>S. cerevisiae</i> CIT3
28	<i>S. cerevisiae</i> CFRC1
29	<i>S. cerevisiae</i> RGI2

30	<i>S. cerevisiae</i> PUT4
31	<i>S. cerevisiae</i> NCA3
32	<i>S. cerevisiae</i> STL1
33	<i>S. cerevisiae</i> ALP1
34	<i>S. cerevisiae</i> NDE2
35	<i>S. cerevisiae</i> QNQ1
36	<i>S. paradoxus</i> pADH2
37	<i>S. kudriavzevii</i> pADH2
38	<i>S. bayanus</i> pADH2
39	<i>S. mikitae</i> pADH2
40	<i>S. castellii</i> pADH2
41	<i>S. paradoxus</i> pPCK1
42	<i>S. kudriavzevii</i> pPCK1
43	<i>S. bayanus</i> pPCK1
44	<i>S. paradoxus</i> pMLS1
45	<i>S. kudriavzevii</i> pMLS1
46	<i>S. bayanus</i> pMLS1
47	<i>S. paradoxus</i> pICL1
48	<i>S. kudriavzevii</i> pICL1
49	<i>S. bayanus</i> pICL1
50	<i>S. cerevisiae</i> pTDH3
51	<i>S. cerevisiae</i> pTEF1
52	<i>S. cerevisiae</i> pFBA1
53	<i>S. cerevisiae</i> pPDC1
54	<i>S. cerevisiae</i> pTPI1
55	<i>S. cerevisiae</i> tADH2

56	S. cerevisiae tPGI1
57	S. cerevisiae tENO2
58	S. cerevisiae tTEF1
59	A. tubingensis GGPPS
60	A. tubingensis PT
61	A. tubingensis FMO
62	A. tubingensis Cyc
63	H. subiculosus hpm8
64	H. subiculosus hpm3
65	pCHIDT-2.1
66	pCHIDT-2c

Table 7: Description of gene sequences

Cluster ID NO.	SEQ ID NO.	Description
AFU3G	67	AFOC1
	68	AFOC9
	69	AFOC6
	70	AFOC5
	71	AFOC8
	72	AFOC4
	73	AFOC7
	74	AFOC5N
	75	AFOC2_PKS
	76	AFOC3
Afulg17740	77	A1OC1_TF
	78	A1OC2_serine_hydrolase
	79	A1OC3_aldose_epimerase

	80	A1OC4_P450
	81	Afulg17740_2_PKS
Ca157	82	Ca157_1_SDR
	83	Ca157_2_Acyl_CoA_oxidase
	84	Ca157_3_P450
	85	Ca157_4_FMO
	86	Ca157_5_PKS
	87	Ca157_6_transferase
	88	Ca157_7_PfpI
	89	Ca157_8_hyp
	90	Ca157_9_AB_hydrolase
	91	Ca157_10_hyp
Ca2032	92	Ca2032_1_3_GHMP_kinase
	93	Ca2032_1_PKS
	94	Ca2032_3_MT
	95	Ca2032_4_P450
	96	Ca2032_5_Ca_uniporter
	97	Ca2032_6_GNAT_acetyltransferase
KU14	98	KU14_SC3_4774_Cyclase
	99	KU14_SC3_4776_P450
	100	KU14_SC3_4773_polyprenyl_synthetase
	101	KU14_SC3_4771_AK_reductase
	102	KU14_SC3_4770_polyprenyl_synthetase
	103	KU14_SC3_4768_AK_reductase
	104	KU14_SC3_4772_DNA_repair
	105	KU14_SC3_47775_QacA_drug_transporter
	106	KU14_SC3_4769_Hyp
KU18	107	KU18_SH9_7287_Cyclase
	108	KU18_SH9_7288_P450

	109	KU18_SH9_7289_P450
	110	KU18_SH9_7286_P450
	111	KU18_SH9_7285_DH
KU26	112	KU26_ETS81063.1_metallo_hydrolase
	113	KU26_ETS81064.1_hyp
	114	KU26_ETS81065.1_esterase
	115	KU26_ETS81066.1_PKS
	116	KU26_ETS81067.1_hyp
	117	KU26_ETS81068.1_SDR
	118	KU26_ETS81069.1_SDR
KU29	119	KU29_BO23_6166_ADH
	120	KU29_BO23_6167_aminotransferase_3
	121	KU29_BO23_6168_fasciclin
	122	KU29_BO23_6169_hyp
	123	KU29_BO23_6170_esterase
	124	KU29_BO23_6171_glycoside_hydrolase
	125	KU29_BO23_6172_hyp
	126	KU29_BO23_6173_P450
	127	KU29_BO23_6174_PKS
KU40	128	KU40_KFA56048.1_NmrA_like
	129	KU40_KFA56160.1_hyp
	130	KU40_KFA56046.1_phytanoly-CoA_dioxygenase
	131	KU40_KFA56190.1_NmrA_like
	132	KU40_KFA56035.1_PKS
	133	KU40_KFA56227.1_hyp
	134	KU40_KFA56040.1_hyp
	135	KU40_KFA56229.1_alkaline_serine_protease
KU41	136	KU41_KIL85236.1_phenylalanine_specific_permease
	137	KU41_KIL85237.1_aldehyde_DH

	138	KU41_KIL85238.1_DH
	139	KU41_KIL85239.1_hyp
	140	KU41_KIL85240.1_hyp
	141	KU41_KIL85241.1_hyp
	142	KU41_KIL85242.1_1-amino_cyclopropane-1-carboxylate_oxidase
	143	KU41_KIL85243.1_hyp
	144	KU41_KIL85244.1_PKS
	145	KU41_KIL85245.1_hyp
	146	KU41_KIL85246.1_aromatic_dioxygenase
	147	KU41_KIL85247.1_peptidase
	148	KU41_KIL85248.1_kinase
	149	KU41_KIL85249.1_metalloprotease
	150	KU41_KIL85250.1_BLA
	151	KU41_KIL85251.1_SDR
	152	KU41_KIL85252.1_AK_reductase
	153	KU41_KIL85253.1_metallopeptidase
KU44	154	KU44_SL06_4460_TPR_repeat
	155	KU44_SL06_4461_PKS
	156	KU44_SL06_4462_aminotransferase_V
	157	KU44_SL06_4463_MFS_transporter
PKS1	158	CS100GC1_CDS1_PKS
	159	CS100GC1_CDS2_serine_hydrolase
	160	CS100GC1_CDS3_p450
	161	CS100GC1_CDS4_short_chain_dehydrogenase
	162	CS100GC1_CDS5_FAD_dehydrogenase
PKS10	163	KU34_CH10_1770_epimerase_DH
	164	KU34_CH10_1802_P450
	165	KU34_CH10_1821_PKS
	166	KU34_CH10_1888_metallo_bla

	167	KU34_CH10_1909_hyp
	168	KU34_CH10_1922_DUF1772
	169	KU34_CH10_1937_NTF2_like
	170	KU34_CH10_1951_NmrA_like
	171	KU34_CH10_1975_SDR
	172	KU34_CH10_2002_ABC_transporter
PKS11	173	KU35_KFY73936.1_SDR
	174	KU35_KFY73937.1_DUF_3425
	175	KU35_KFY73938.1_Zn_finger
	176	KU35_KFY73939.1_OMT
	177	KU35_KFY73940.1_metallo_lactamase
	178	KU35_KFY73941.1_PKS
	179	KU35_KFY73942.1_acetyl_transferase
	180	KU35_KFY73943.1_NmrA_like
	181	KU35_KFY73944.1_FAD_linked_oxygenase
PKS12	182	KU36_KFY14209.1_hyp
	183	KU36_KFY14210.1_OMT
	184	KU36_KFY14211.1_Metallo_BLA
	185	KU36_KFY14212.1_PKS
	186	KU36_KFY14213.1_FAD_linked_oxidase
PKS13	187	KU37_KFY01907.1_DH
	188	KU37_KFY01908.1_ABC_tranporter
	189	KU37_KFY01909.1_PKS
	190	KU37_KFY01910.1_MFS
	191	KU37_KFY01911.1_PKc_like
PKS14	192	KU38_KIW01747.1_metallo_hydrolase
	193	KU38_KIW01748.1_halogenase
	194	KU38_KIW01749.1_P450
	195	KU38_KIW01750.1_cupin_like

	196	KU38_KIW01751.1_OMT
	197	KU38_KIW01752.1_MHR_TF
	198	KU38_KIW01753.1_monooxygenase
	199	KU38_KIW01754.1_PKS
PKS15	200	KU39_ESK96608.1_monocarboxylate_permease
	201	KU39_ESK96609.1_NAD_epimerase_DH
	202	KU39_ESK96610.1_hyp
	203	KU39_ESK96611.1_carbonyl_reductase
	204	KU39_ESK96612.1_phenol_2_monooxygenase
	205	KU39_ESK96613.1_PKS
	206	KU39_ESK96614.1_SDR
PKS16	207	NW_006767437_-_cystathionine_beta-synthase_CDS
	208	NW_006767437_-_hypothetical_protein_1_CDS
	209	NW_006767437_-_hypothetical_protein_2_CDS
	210	NW_006767437_-_hypothetical_protein_3_CDS
	211	NW_006767437_-_Pkinase-domain-containing_protein_CDS
PKS17	212	KU43_KIJ60838.1
	213	KU43_KIJ60843.1_P450
	214	KU43_KIJ60845.1
	215	KU43_KIJ60839.1_isoprenylcysteine_carboxyl_methyltransferase
	216	KU43_KIJ60847.1_isoprenylcysteine_carboxyl_methyltransferase
	217	KU43_KIJ60848.1_ABC_transporter
	218	KU43_KIJ60844.1
	219	KU43_KIJ60846.1_P450
	220	KU43_KIJ60842.1
	221	KU43_KIJ60840.1_ABC_transporter
	222	KU43_KIJ60837.1_P450
	223	KU43_KIJ60841.1_P450
	224	KU43_KIJ60886.1_PKS

PKS18	225	SU62_EFR04826.1_hypothetical_protein
	226	SU62_EFR04827.1_fatty-acid-CoA_ligase
	227	SU62_EFR04828.1_pks
	228	SU62_EFR04829.1_esterase
PKS19	229	SU64_EOA86426.1_YfhR_like
	230	SU64_EOA86427.1_ER
	231	SU64_EOA86425.1_FAD-hydroxylase
	232	SU64_EOA86421.1_Drug_resistance_transporter
	233	SU64_EOA86423.1_OMT
	234	SU64_EOA86422.1_pks
	235	SU64_EOA86424.1_esterase
PKS2	236	CS163GC1_CDS1_serine_hydrolase
	237	CS163GC1_CDS2_p450
	238	CS163GC1_CDS3_PKS
	239	CS163GC1_CDS4_short_chain_dehydrogenase
	240	CS163GC1_CDS_FAD_dehydrogenase
PKS20	241	SU65_EFQ95559.1_BLA
	242	SU65_EFQ95560.1_PKS
	243	SU65_EFQ95561.1_hyp
	244	SU65_EFQ95562.1_KR
PKS21	245	SU67_EXJ61964.1_drug_resistance_transporter
	246	SU67_EXJ61965.1_scytalone_dehydratase
	247	SU67_EXJ61966.1_versicolorin_reductase
	248	SU67_EXJ61967.1_AMP_ligase
	249	SU67_EXJ61968.1_hypothetical_protein
	250	SU67_EXJ61969.1_FAD_monooxygenase
	251	SU67_EXJ61970.1_pks
	252	SU67_EXJ61971.1_metallo_BLA
	253	SU67_EXJ61972.1_hyp

	254	SU67_EXJ61973.1_SDR
	255	SU67_EXJ61974.1_AifR_reg
PKS22	256	SU68_GAM43180.1_beta-lactamase_family_protein
	257	SU68_GAM43181.1_NAD-dependent_epimerase_dehydratase
	258	SU68_GAM43183.1_oxidoreductase
	259	SU68_GAM43185.1_benzoate_4-monooxygenase_cytochrome_P450
	260	SU68_GAM43187.1_scytalone_dehydratase
	261	SU68_GAM43176.1_riboflavin_biosynthesis_protein
	262	SU68_GAM43177.1_sugar_transport_protein
	263	SU68_GAM43178.1_short-chain_dehydrogenase
	264	SU68_GAM43179.1_pks
	265	SU68_GAM43182.1_halogenase
	266	SU68_GAM43184.1_NAD-dependent_epimerase_dehydratase
	267	SU68_GAM43186.1_SDR
PKS23	268	SU70_ERF77218.1_SDR
	269	SU70_ERF77219.1_Fungal_TF
	270	SU70_ERF77220.1_p450
	271	SU70_ERF77221.1_pks
	272	SU70_ERF77222.1_DABB_superfamily
	273	SU70_ERF77223.1_FAD_monooxygenase
	274	SU70_ERF77224.1_alcohol_DH
	275	SU70_ERF77225.1_OMT
	276	SU70_ERF77226.1_alcohol_DH
PKS24	277	SU71_GAM40295.1_sulfhydrolase
	278	SU71_GAM40298.1_AMP_ligase
	279	SU71_GAM40299.1_MT
	280	SU71_GAM40301.1_carnosine_synthase
	281	SU71_GAM40303.1_carnitine_acetyl-CoA_transferase
	282	SU71_GAM40296.1_aminotransferase

	283	SU71_GAM40297.1_ammonia_lyase
	284	SU71_GAM40300.1_pks
	285	SU71_GAM40302.1_ABC_tranporter
PKS25	286	SU72_ESK88623.1_pks
	287	SU72_ESK88624.1_carotenoid_cleavage_dioxygenase_1
	288	SU72_ESK88625.1_long-chain-fatty-acid-ligase
	289	SU72_ESK88626.1_amino_acid_permease
PKS26	290	SU73_KN817529.1_pks
	291	SU73_KN817529.1_hyp
	292	SU73_KN817529.1_AMP_ligase
	293	SU73_KN817529.1_8_amino_7_oxanoate_synthase
PKS27	294	SU74_EMD35673.1_NAD_DH
	295	SU74_EMD35676.1_sulfhydrylase
	296	SU74_EMD35664.1_hyp
	297	SU74_EMD35669.1_halogenase
	298	SU74_EMD35663.1_AB_hydrolase
	299	SU74_EMD35666.1_alcohol_DH
	300	SU74_EMD35667.1_Drug_resistance_transporter
	301	SU74_EMD35668.1_hyp
	302	SU74_EMD35670.1_P450
	303	SU74_EMD35665.1_hyp
	304	SU74_EMD35671.1_pks
	305	SU74_EMD35672.1_hyp
	306	SU74_EMD35674.1_hyp
	307	SU74_EMD35675.1_hyp
PKS28	308	SU75_ESK88629.1_pks
	309	SU75_ESK88630.1_drug_resistance_subfamily
	310	SU75_ESK88631.1_hypothetical_protein
	311	SU75_ESK88632.1_dead-box_protein_abstrakt

	312	SU75_ESK88633.1_hyp
	313	SU75_ESK88634.1_nadh-ubiquinone_oxidoreductase
PKS3	314	AK_C24701GC76_CDS1_class_II_aminotransferase
	315	AK_C24701GC76_CDS2_p450
	316	AK_C24701GC76_CDS3_PKS
	317	AK_C24701GC76_CDS4_ferric_chelate_reductase
	318	AK_C24701GC76_CDS5_DUF4243
PKS4	319	KU27_AN22_1464_esterase
	320	KU27_AN22_1465_PKS
	321	KU27_AN22_1466_hyp
	322	KU27_AN22_1467_A_TD
	323	KU27_AN22_1468_hyp
	324	KU27_AN22_1469_amino_oxidase
PKS5	325	KU28_TT08_2721_AK_reductase
	326	KU28_TT08_2722_ABC_transporter
	327	KU28_TT08_2723_esterase
	328	KU28_TT08_2724_PKS
	329	KU28_TT08_2725_sugar_transport
PKS6	330	KU30_TV43_5580_esterase
	331	KU30_TV43_5581_P450
	332	KU30_TV43_5582_PKS
	333	KU30_TV43_5583_OMT
	334	KU30_TV43_5584_P450
PKS7	335	KU31_KFY69032.1_P450
	336	KU31_KFY69033.1_hyp
	337	KU31_KFY69034.1_esterase
	338	KU31_KFY69035.1_P450
	339	KU31_KFY69036.1_PKS
	340	KU31_KFY69037.1_glycoside_hydrolase

	341	KU31_KFY69038.1_thymine_dioxygenase
PKS8	342	KU32_KEZ41287.1_nucleotidyltransferase
	343	KU32_KEZ41288.1_OMT
	344	KU32_KEZ41289.1_AB_hydrolase
	345	KU32_KEZ41290.1_mito_phos_carrier
	346	KU32_KEZ41291.1_hyp
	347	KU32_KEZ41292.1_crotonyl_CoA_reductase
	348	KU32_KEZ41293.1_PKS
	349	KU32_KEZ41294.1_Drug_resistance_transporter
	350	KU32_KEZ41295.1_NADB_monooxygenase
PKS9	351	KU33_KJK75348.1_alkaline_serine_hydrolase
	352	KU33_KJK75349.1_LysM_containing_protein
	353	KU33_KJK75350.1_PKS
	354	KU33_KJK75351.1_drug_resistance_transporter
	355	KU33_KJK75352.1_phytanoyl_CoA_dioxygenase
	356	KU33_KJK75353.1_SDR
	357	KU33_KJK75354.1_amino-7_oxonolate_synthase
SU61	358	SU61_ENH82084.1_choline_oxidase
	359	SU61_ENH82085.1_fungal_specific_transcription_factor
	360	SU61_ENH82086.1_short-chain_dehydrogenase
	361	SU61_ENH82087.1_hypothetical_protein
	362	SU61_ENH82088.1_hypothetical_protein
	363	SU61_ENH82089.1_pks
	364	SU61_ENH82090.1_esterase
	365	SU61_ENH82091.1_ABC_multidrug_transporter_mdr1
	366	SU61_ENH82092.1_cytochrome_b5_type_b
	367	SU61_ENH82093.1_sulfite_reductase_subunit_alpha
SU63	368	SU63_KJZ74253.1_hyp_signalling_protein
	369	SU63_KJZ74254.1_mtRNA_formyl_transferase

	370	SU63_KJZ74255.1_esterase
	371	SU63_KJZ74256.1_PKS
	372	SU63_KJZ74257.1_choline_DH_halogenase
SU66	373	SU66_EMF17384.1_ras-domain-containing_protein
	374	SU66_EMF17385.1_phosphoinositide_phosphatase
	375	SU66_EMF17387.1_hypothetical_protein
	376	SU66_EMF17389.1_versicolorin_reductase
	377	SU66_EMF17390.1_scytalone_dehydratase
	378	SU66_EMF17386.1_pks
	379	SU66_EMF17388.1_Metallo-hydrolase_oxidoreductase
	380	SU66_EMF17391.1_StcQ-like_protein
SU69	381	SU69_KFY04761.1_aldehyde_oxidase
	382	SU69_KFY04762.1_SDR
	383	SU69_KFY04763.1_hyp
	384	SU69_KFY04764.1_reg_protein
	385	SU69_KFY04765.1_OMT
	386	SU69_KFY04766.1_metallo_BLA
	387	SU69_KFY04767.1_pks
	388	SU69_KFY04768.1_FAD-linked_oxidase
TC1	389	Tv86_130_CDS
	390	Tv86_132_CDS
	391	Tv86_133_CDS
	392	Tv86_134_CDS
	393	Tv86_135_ORF
	394	Tv86_136_CDS
	395	Tv86_137_CDS
TC10	396	KU19_SH16_10821_Cyclase
	397	KU19_SH16_10820_P450
	398	KU19_SH16_10822_P450

	399	KU19_SH16_10823_AK_reductase
	400	KU19_SH16_10819_QacA_MFS
TC11	401	KU_20_SH18_11663_Cyclase
	402	KU_20_SH18_11664_P450
	403	KU_20_SH18_11665_P450
	404	KU_20_SH18_11662_P450
	405	KU_20_SH18_11660_NMT
	406	KU_20_SH18_11661_MFS
TC12	407	KU21_DS19_7112_Cyclase
	408	KU21_DS19_7113_P450
	409	KU21_DS19_7108_P450
	410	KU21_DS19_7109_MT
	411	KU21_DS19_7111_GMC_oxido
	412	KU21_DS19_7114_AK_reductase
	413	KU21_DS19_7110_Hyp
TC13	414	KU22_DDJM14_6568_Cyclase
	415	KU22_DDJM14_6570_P450
	416	KU22_DDJM14_6567_OMT
	417	KU22_DDJM14_6565_MT
	418	KU22_DDJM14_6564_MT
	419	KU22_DDJM14_6562_P450
	420	KU22_DDJM14_6563_MT
	421	KU22_DDJM14_6561_GST
	422	KU22_DDJM14_6569_V-type_ATPase
	423	KU22_DDJM14_6566_TF
TC2	424	Tv83_13_CDS
	425	Tv83_14_CDS
	426	Tv83_15_CDS
	427	Tv83_16_CDS

TC3	428	BFT4_1_FAD_binding_domain_protein
	429	BFT4_2_Aldo_keto_reductase_oxidoreductase
	430	BFT4_3_cytochrome_P450
	431	BFT4_4_UbiA_cyclase
	432	BFT4_6_hyp_protein_4
	433	BFT4_7_hyp_protein_5
	434	BFT4_8_hyp_protein_6
	435	BFT4_10_FAD_FMN_isoamyl_alcohol_oxidase
	436	BFT4_11_hyp_protein_8
	437	BFT4_14_hyp_protein_11
	438	BFT4_15_glutathione_S_transferase
	439	BFT4_16_D_isomer_specific_2_hydroxyacid_dehydrogenase
TC4	440	KU11_FM3_3034_Cyclase
	441	KU11_FM3_3033_P450
	442	KU11_FM3_3032_P450
	443	KU11_FM3_3031_FAD
	444	KU11_FM3_3037_Hydrox
	445	KU11_FM3_3027_AMP_SDR
	446	KU11_FM3_3030_PHOS
	447	KU11_FM3_3035_Hyp
TC5	448	KU12_HI6_11661_Cyclase
	449	KU12_HI6_11655_P450
	450	KU12_HI6_11638_P450
	451	KU12_HI6_11667_AMP_ligase
	452	KU12_HI6_11646_monocarbox_MFS
	453	KU12_HI6_11632_MFS
TC6	454	KU13_CeS8_5906_Cyclase
	455	KU13_CeS8_5905_P450
	456	KU13_CeS8_5911_P450

	457	KU13_CeS8_5904_choline_DH
	458	KU13_CeS8_5910_halogenase
	459	KU13_CeS8_5907_AMP_ligase
	460	KU13_CeS8_5909_superoxide_dismutase
	461	KU13_CeS8_5908_MFS
	462	KU13_CeS8_5912_MFS
TC7	463	KU15_DS54_10337_Cyclase
	464	KU15_DS54_10336_P450
	465	KU15_DS54_10340_P450
	466	KU15_DS54_10341_P450
	467	KU15_DS54_10338_polyprenyl_synthetase
	468	KU15_DS54_10333_GMC_oxido
	469	KU15_DS54_10334_GMC_oxido
	470	KU15_DS54_10339_Hyp
	471	KU15_DS54_10335_Hyp_transmembrane
TC8	472	KU16_PO11_11845_Cyclase
	473	KU16_PO11_11844_FAD_oxidase
	474	KU16_PO11_11843_P450
	475	KU16_PO11_11841_2_P450
	476	KU16_PO11_11840_FAD_oxidase
	477	KU16_PO11_11847_FAD_oxidase
	478	KU16_PO11_11846_SDR
	479	KU16_PO11_11848_AMP_ligase
	480	KU16_PO11_11839_AMP_ligase
TC9	481	KU17_SC18_16687_Cyclase
	482	KU17_SC18_16686_P450
	483	KU17_SC18_16688_SDR

[285] While preferred embodiments of the present disclosure have been shown and described herein, those skilled in the art will understand that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the disclosure.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for screening a plurality of compounds, the method comprising:
identifying a gene cluster that includes or is within 20 kilobases of a region that encodes for a protein that is identical with or homologous to a first target protein, wherein the gene cluster comprises a region that encodes for a protein selected from the group consisting of (1) polyketide synthases and (2) non-ribosomal peptide synthetases, (3) terpene synthetases, (4) UbiA-type terpene cyclases, and (5) dimethylallyl transferases;
introducing a plurality of genes from the gene cluster into a vector;
introducing the vector into a host cell;
expressing the proteins encoded by the plurality of genes in the host cell;
determining whether a compound that is formed or modified by the expressed proteins modulates the first target protein.
2. The method of claim 1, wherein the host cell is a yeast cell.
3. The method of claim 2, wherein the yeast cell is a yeast cell that has been modified to increased sporulation frequency and increased mitochondrial stability.
4. The method of any one of claims 1–3, wherein each gene of the plurality of genes is under the control of a different promoter.
5. The method of claim 4, wherein the promoters are designed to increase expression when the host cell is in the presence a nonfermentable carbon source.
6. The method of any one of claims 1–5, wherein the plurality of genes are introduced into the vector via homologous recombination.

7. The method of claim 6, wherein introducing the plurality of genes into the vector via homologous recombination comprises combining a first plurality of nucleotides with a second plurality of nucleotides, wherein:

each polynucleotide of the first plurality of DNA polynucleotides encodes for a promoter and terminator, wherein each promoter and terminator is distinct from the promoter and terminator of other polynucleotides of the first plurality of nucleotides; and

each polynucleotide of the second plurality of nucleotides includes a coding sequence, a first flanking region on the 5' side of the polynucleotide and a second flanking region on the 3' side of the polynucleotide; and

introducing the polynucleotides into a host cell that includes machinery for homologous recombination, wherein the host cell assembles the expression vector via homologous recombination that occurs in the flanking regions of the second plurality of polynucleotides;

wherein the expression vector is configured to facilitate simultaneously production of a plurality of proteins encoded by the second plurality of nucleotides.

8. The method of claim 7, wherein the first flanking region and the second flanking region are each between 15 and 75 base pairs in length.

9. The method of claim 8, wherein the first flanking region and the second flanking region are each between 40 and 60 base pairs in length.

10. A method for identifying a gene cluster capable of producing a small molecule for modulating a first target protein, the method comprising:

selecting, from a database comprising a list of biosynthetic gene clusters, one or more gene clusters that include or are positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein.

11. The method of claim 10, wherein the one or more gene clusters are selected from the group consisting of (1) clusters that comprise one or more polyketide synthases and (2) clusters that comprise one or more non-ribosomal peptide synthetases, (3) clusters that comprise one or more

terpene synthases, (4) clusters that comprises one or more UbiA-type terpene cyclases, and (5) clusters that comprise one or more dimethylallyl transferases.

12. The method of any one of claims 10–11, wherein the protein that is encoded by the region that is included in or positioned proximal to the biosynthetic gene cluster is identical to or has greater than 30% homology to the first target protein.

13. The method of any one of claims 10–12, wherein the region that encodes the protein that is identical with or homologous to the first target protein is within 20,000 base pairs of a region of a portion of the gene cluster that encodes a polyketide synthase, a non-ribosomal peptide synthetase, a terpene synthetase, a UbiA-type terpene cyclase, or a dimethylallyl transferase.

14. The method of any one of claims 10–13, wherein the first target protein is BRSK1.

15. The method of any one of claims 10–14, wherein selecting the one or more gene clusters comprises operating a computer, wherein operation of the computer comprises running an algorithm that takes into account both an input sequence for the first target protein and sequence information from a database that includes sequence information from a plurality of species such that the computer returns information corresponding to one or more gene clusters.

16. The method of claim 15, wherein the algorithm takes into account the phylogenetic relationship between gene clusters in the database.

17. The method of any one of claims 10–16, wherein the one or more gene clusters include a coding sequence for a protein that is an extracellular protein, a membrane tethered protein, a protein involved in a transport or secretion pathway, a protein homologous to a protein involved in a transport or secretion pathway, a protein with a peptide targeting signal, a protein with a terminal sequence with homology to a targeting signal, an enzyme that degrades small molecules, or a protein with homology to an enzyme that degrades small molecules.

18. A method for producing a compound that modulates the first target protein, the method comprising:

identifying a gene cluster via the method of any one of claims 10–17;

expressing the gene cluster or a plurality of genes from the gene cluster in a host cell; and

isolating a compound produced by the gene cluster.

19. The method of claim 18, further comprising screening the isolated compound for modulation of an activity of the first target protein.

20. The method of claim 18 or claim 19, wherein the cluster-encoded protein that is homologous to the first target protein is resistant to modulation by the isolated compound when compared to modulation of the first target protein.

21. The method of claim 20, wherein the compound is not toxic to the species from which the cluster originates due to one or more of (1) sequence differences between target protein and the cluster-encoded protein, (2) spatial separation of the compound from the cluster-encoded protein and (3) high expression levels for the cluster-encoded protein.

22. A method for making a DNA vector, the method comprising:

identifying a gene cluster comprising a plurality of genes that are capable of producing a small molecule for modulating a first target protein;

introducing two or more genes of the plurality of genes into a vector, wherein the vector is configured to facilitate expression of the two or more genes in a host organism;

wherein (1) the gene cluster encodes one or more proteins selected from the group consisting of a polyketide synthase, a non-ribosomal peptide synthetase, a terpene synthetase, a UbiA-type terpene cyclase, and a dimethylallyl transferase and (2) the gene cluster includes or is positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein.

23. The method of claim 22, wherein the DNA vector is a circular plasmid.

24. The method of any one of claims 22–23, wherein the DNA vector comprises a plurality of promoters, wherein each promoter of the plurality of promoters is configured to, when the vector is introduced into a *Saccharomyces cerevisiae* cell, promote a lower level of heterologous expression when the cell exhibits predominantly anaerobic energy metabolism than when the cell exhibits aerobic energy metabolism.

25. The method of claim 24, wherein each promoter of the plurality of promoters differs in sequence from one another.

26. The method of any one of claims 24–25, wherein each promoter of the plurality of promoters has a sequence selected from the group consisting of: SEQ ID NOs: 1–66.

27. The method of claim 26, wherein each promoter of the plurality of promoters has a sequence selected from the group consisting of: SEQ ID NOs: 20–35, and SEQ ID NOs: 41–50.

28. The method of any one of claims 24–27, wherein, when the *Saccharomyces cerevisiae* cell is exhibiting anaerobic energy metabolism, the cell is catabolizing a fermentable carbon source selected from glucose or dextrose; and when the *Saccharomyces cerevisiae* cell is exhibiting aerobic energy metabolism, the cell is catabolizing a non-fermentable carbon source selected from ethanol or glycerol.

29. A method for the heterologous expression of a plurality of genes in a yeast strain, the method comprising:

obtaining a yeast strain that includes a vector for expressing a plurality of genes from a single gene cluster of a non-yeast organism; and

inducing expression of the plurality of genes;

wherein the gene cluster in the non-yeast organism includes or is positioned proximal to a region that encodes a protein that is at least 30% homologous to a target protein.

30. The method of claim 29, further comprising introducing the plurality of genes from the single gene cluster into the vector.

31. The method of any one of claims 29–30, further comprising introducing the vector into the yeast strain.

32. The method of any one of claims 29–31, wherein expression of the plurality of genes results in the formation of small molecule, wherein the small molecule modulates the activity of the target protein.

33. The method of any one of claims 29–32, wherein the gene cluster is a gene cluster of a non-yeast fungus.

34. The method of any one of claims 29–33, wherein the yeast strain is from *Saccharomyces cerevisiae*.

35. The method of any one of claims 29–34, wherein the target protein is a human protein.

36. A system for identifying a gene cluster for introduction of a plurality of genes from the gene cluster into a host organism, the system comprising:

a processor;

a non-transitory computer-readable medium comprising instructions that, when executed by the processor, cause the processor to perform operations, the operations comprising:

loading the identity or sequence of a first target protein into memory;

loading the identity or sequence of a plurality of biosynthetic gene clusters into memory;

identifying, from the plurality of biosynthetic gene clusters, one or more gene clusters that encode or are positioned proximal to a region that encodes a protein that is identical with or homologous to the first target protein; and

scoring the one or more gene clusters based on the likelihood of each gene cluster being capable of use to produce a small molecule that modulates the first target protein.

37. The system of claim 36, wherein scoring the one or more gene clusters comprises comparing the sequence of the first target protein (or a DNA sequence encoding the first target protein) to a sequence of a protein encoded in or proximal to the gene cluster (or to a DNA sequence encoding the protein that is in or proximal to the gene cluster).

38. A system for identifying one or more biosynthetic gene clusters for introduction into a host organism to produce one or more compounds that modulate a specific target protein, the system comprising:

- a processor;

- a memory containing a gene cluster identification application;

- wherein the gene cluster identification application directs the processor to:

- load data describing at least one target protein into the memory;

- load data describing a plurality of biosynthetic gene clusters into the memory;

- score each of the plurality of biosynthetic gene clusters based upon:

- performing a homolog search for each biosynthetic gene cluster to determine a presence of at least one homolog of a target protein within or adjacent the biosynthetic gene cluster;

- confidence of homology of the at least one target protein to at least one gene in a biosynthetic gene cluster;

- a fraction of a homologous gene that meets an identity threshold;

- a total number of genes homologous to the at least one target protein present in the entire genome of an organism;

- homology of the at least one homolog of at least one target protein within or adjacent the biosynthetic gene cluster to genes in the target protein's genome;

- phylogenetic relationship of the at least one target protein to a gene in a cluster;

expected number of homologs of the at least one target protein in or adjacent to a biosynthetic cluster; and
a likelihood that at least one target protein is essential for cellular process in the natural environment; and
output a report identifying one or more biosynthetic gene clusters that are most likely to produce a compound that modulates the at least one target protein.

39. A method for selecting a biosynthetic gene cluster that produces a secondary metabolite, the method comprising:

obtaining a list of gene clusters;
performing a phylogenetic analysis of the genes within the clusters compared to known genes from known biosynthetic gene clusters; and
selecting the biosynthetic gene cluster based on its phylogenetic relationship with the known genes.

40. The method of claim 39, wherein the biosynthetic gene cluster with the most distant phylogenetic relationship from the known genes is selected.

41. A method for identifying a gene cluster that produces a compound that binds a protein of interest, the method comprising:

obtaining sequence information for a plurality of contiguous sequences, wherein each contiguous sequence includes a biosynthetic gene cluster and flanking genomic sequences;
analyzing the contiguous sequences for the presence of a gene that encodes a protein with homology to the protein of interest, and
selecting a biosynthetic gene cluster which includes, or is proximal to, a gene that encodes a protein that is homologous to the protein of interest.

42. The method of claim 41, wherein the contiguous nucleotide sequence is less than 40,000 base pairs in length.

43. A modified yeast cell having a BY background, wherein relative to unmodified BY4741 and BY4742, the modified yeast cell has both (1) increased sporulation frequency and (2) increased mitochondrial stability.

44. The modified yeast cell of claim 43, wherein the modified yeast cell grows faster on non-fermentable carbon sources than unmodified BY4741 and BY4742.

45. The modified yeast cell of any one of claims 43–44, wherein the yeast cell comprises one or more of the following genotypes: *MKT1(30G)*, *RME1(INS-308A)*, and *TAO3(1493Q)*.

46. The modified yeast cell of any one of claims 43–45, wherein the yeast cell comprises one or more of the following genotypes: *CAT5(91M)*, *MIP1(661T)*, *SAL1+*, and *HAP1+*.

47. A method of forming an expression vector, the method comprising:
combining a first plurality of DNA polynucleotides with a second plurality of polynucleotides, wherein:

each polynucleotide of the first plurality of DNA polynucleotides encodes for a promoter and terminator, wherein each promoter and terminator is distinct from the promoter and terminator of other polynucleotides of the first plurality of nucleotides; and

each polynucleotide of the second plurality of nucleotides includes a coding sequence, a first flanking region on the 5' side of the polynucleotide and a second flanking region on the 3' side of the polynucleotide, wherein each flanking region is between 15 and 75 base pairs in length; and

introducing the polynucleotides into a host cell that includes machinery for homologous recombination, wherein the host cell assembles the expression vector via homologous recombination that occurs in the flanking regions of the second plurality of polynucleotides;

wherein the expression vector is configured to facilitate simultaneously production of a plurality of proteins encoded by the second plurality of nucleotides.

48. The method of claim 47, wherein the host cell is a yeast cell.

49. The method of any one of claims 47–48, wherein each flanking region is between 40 and 60 base pairs in length.

50. The method of any one of claims 47–49, wherein at least one polynucleotide of the first plurality of nucleotides encodes a selection marker.

51. A system for generating a synthetic gene cluster via homologous recombination, the system comprising 1 through N unique promoter sequences, 1 through N unique terminator sequences, and 1 through N unique coding sequences, wherein:

coding sequence 1 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 1 and a second end portion is identical to the first 30–70 base pairs of terminator 1;

coding sequence 2 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 2 and a second end portion is identical to the first 30–70 base pairs of terminator 2;

and coding sequence N is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter N and a second end portion is identical to the first 30–70 base pairs of terminator N.

52. The system of claim 51, wherein terminator 1 and promoter 2 are portions of the same double-stranded oligonucleotide.

53. A method for assembling a synthetic gene cluster, the method comprising:

- a. obtaining 1 through N unique promoters, 1 through N unique terminators, and 1 through N unique coding sequences, wherein:
 - i. coding sequence 1 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 1 and a second end portion is identical to the first 30–70 base pairs of terminator 1;

- ii. coding sequence 2 is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter 2 and a second end portion is identical to the first 30–70 base pairs of terminator 2; and
 - iii. and coding sequence N is attached to an additional 30–70 base pair sequence on each end such that a first end portion is identical to the last 30–70 base pairs of promoter N and a second end portion is identical to the first 30–70 base pairs of terminator N;
- b. transforming the 1 through N promoters, terminators and coding sequences into a yeast cell;
 - c. isolating a plasmid containing the 1 through N promoters, terminators and coding sequences from the yeast cell.
54. The method of claim 53, further comprising a coding sequence for a selection marker.
55. The method of claim 54, wherein the selection marker is an auxotrophic marker.
56. The method of any one of claims 53–55, wherein the yeast cell has a deficiency in a DNA ligase gene.
57. A yeast strain which allows for both (1) homologous DNA assembly and (2) production of heterologous genes in the same strain.
58. The yeast strain of claim 57, wherein the yeast strain is a DHY strain.
59. The yeast strain of and one of claims 57–58, wherein the strain allows DNA assembly via homologous recombination with an efficiency of at least 80% as compared to DNA assembly in BY.

60. The yeast strain of any one of claims 57–59, wherein production of heterologous compounds in the strain is accomplished with an efficiency of at least 80% as compared to heterologous compound production in BJ5464.

61. The yeast strain of any one of claims 57–60, wherein the strain allows production of heterologous proteins with an efficiency of at least 80% as compared to heterologous protein production in BJ5464.

62. A method for isolating a plasmid from a yeast cell, the method comprising:

- a. isolating total DNA from a yeast cell that includes a plasmid;
- b. incubating the DNA with an exonuclease such that the exonuclease degrades substantially all of the linear DNA in the isolated total DNA from the yeast cell;
- c. optionally inactivating the exonuclease; and
- d. recovering the plasmid DNA.

63. The method of claim 62, wherein the isolated plasmid DNA is of sufficient purity for use in a sequencing reaction.

64. The method of any one of claims 62–63, wherein the plasmid DNA is further prepared for a sequencing reaction.

65. A pharmaceutical composition comprising Compound 6 and a pharmaceutically acceptable excipient.

66. A pharmaceutical composition comprising Compound 7 and a pharmaceutically acceptable excipient.

67. A pharmaceutical composition comprising Compound 8 and a pharmaceutically acceptable excipient.

68. A pharmaceutical composition comprising Compound 9 and a pharmaceutically acceptable excipient.

69. A pharmaceutical composition comprising Compound 10 and a pharmaceutically acceptable excipient.

70. A pharmaceutical composition comprising Compound 11 and a pharmaceutically acceptable excipient.

71. A pharmaceutical composition comprising Compound 12 and a pharmaceutically acceptable excipient.

72. A pharmaceutical composition comprising Compound 13 and a pharmaceutically acceptable excipient.

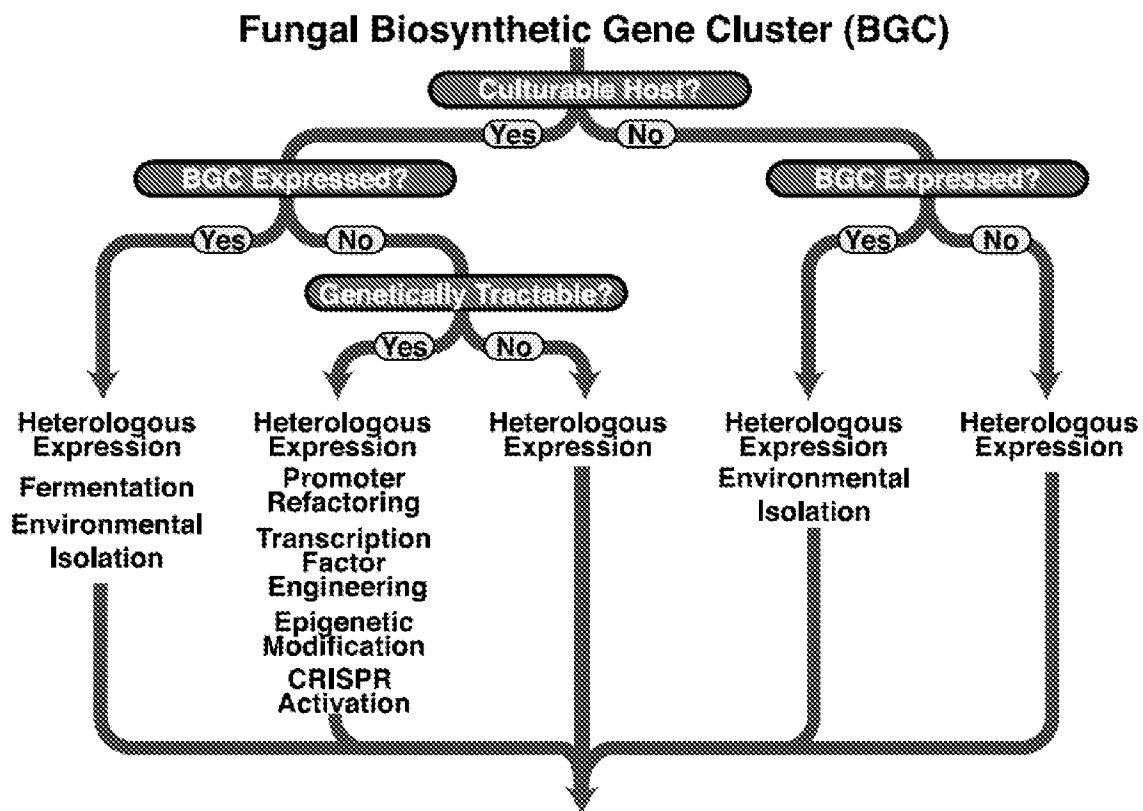
73. A pharmaceutical composition comprising Compound 14 and a pharmaceutically acceptable excipient.

74. A pharmaceutical composition comprising Compound 15 and a pharmaceutically acceptable excipient.

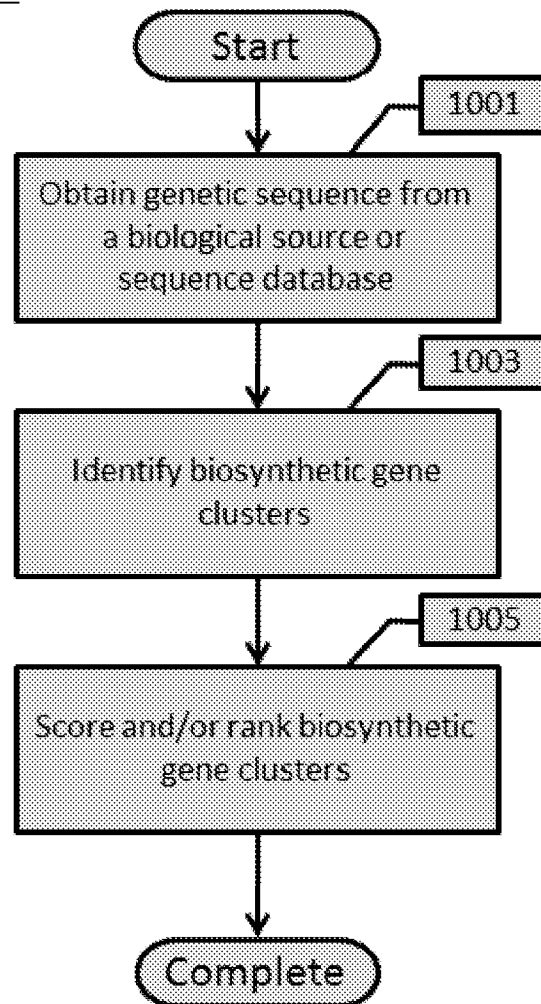
75. A pharmaceutical composition comprising Compound 16 and a pharmaceutically acceptable excipient.

76. A method of producing Compound 3, the method comprising:
providing a vector or vectors comprising the coding sequences of **SEQ ID NOs: 200–206** ;
transforming a host cell with the vector or vectors;
incubating the host cell in culture media under conditions suitable for the expression of the coding sequences; and
isolating the compound produced by the host cell.

1 / 87



2 / 87

Process 1000**FIG. 1B**

3 / 87

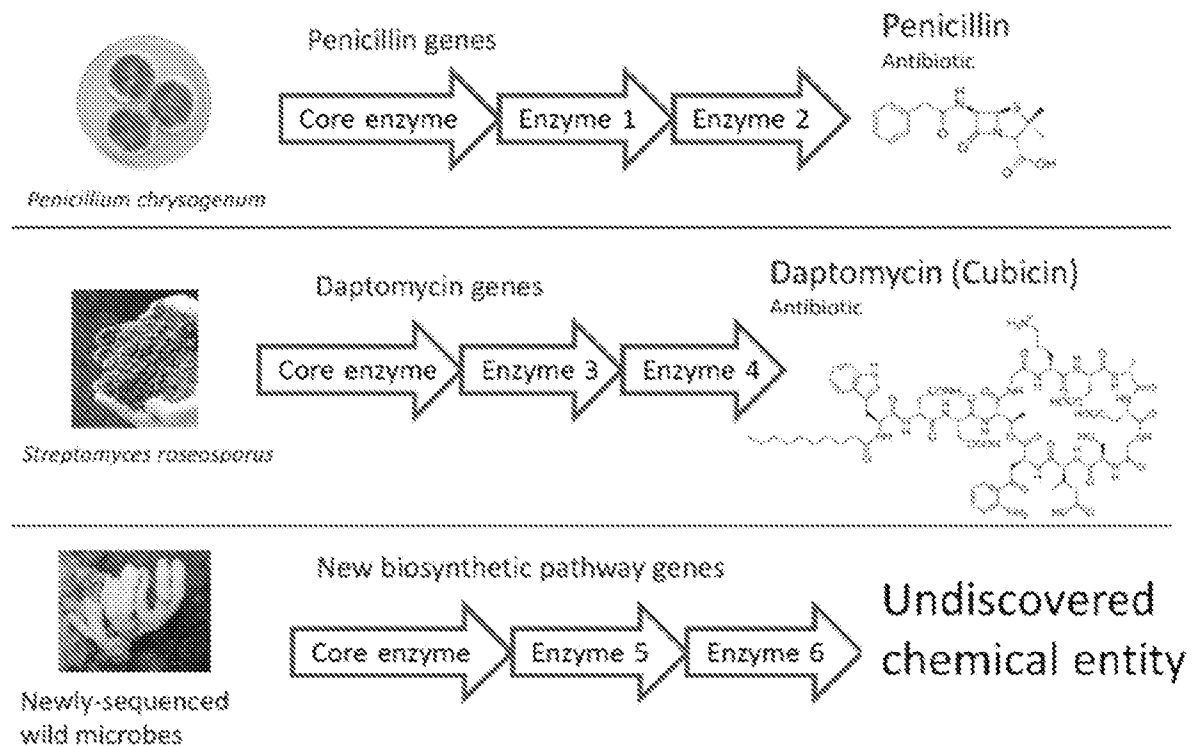
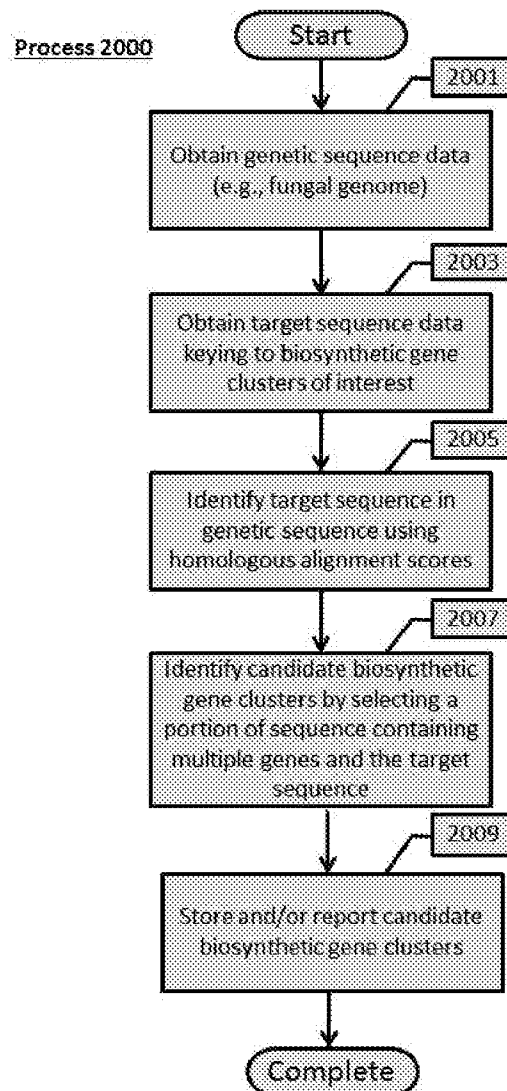


FIG. 1C

4 / 87

**FIG. 1D**

5 / 87

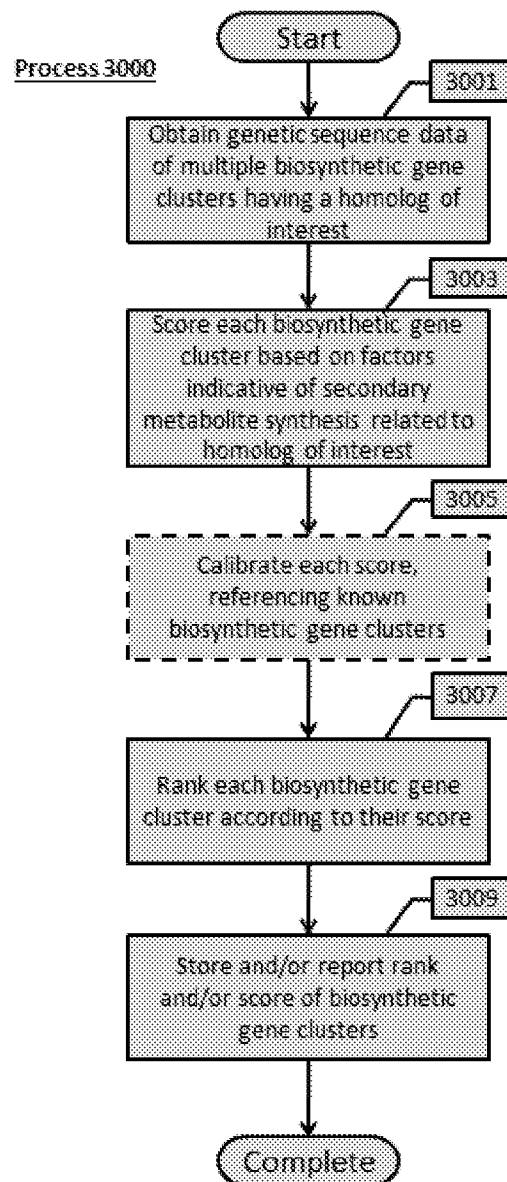


FIG. 1E

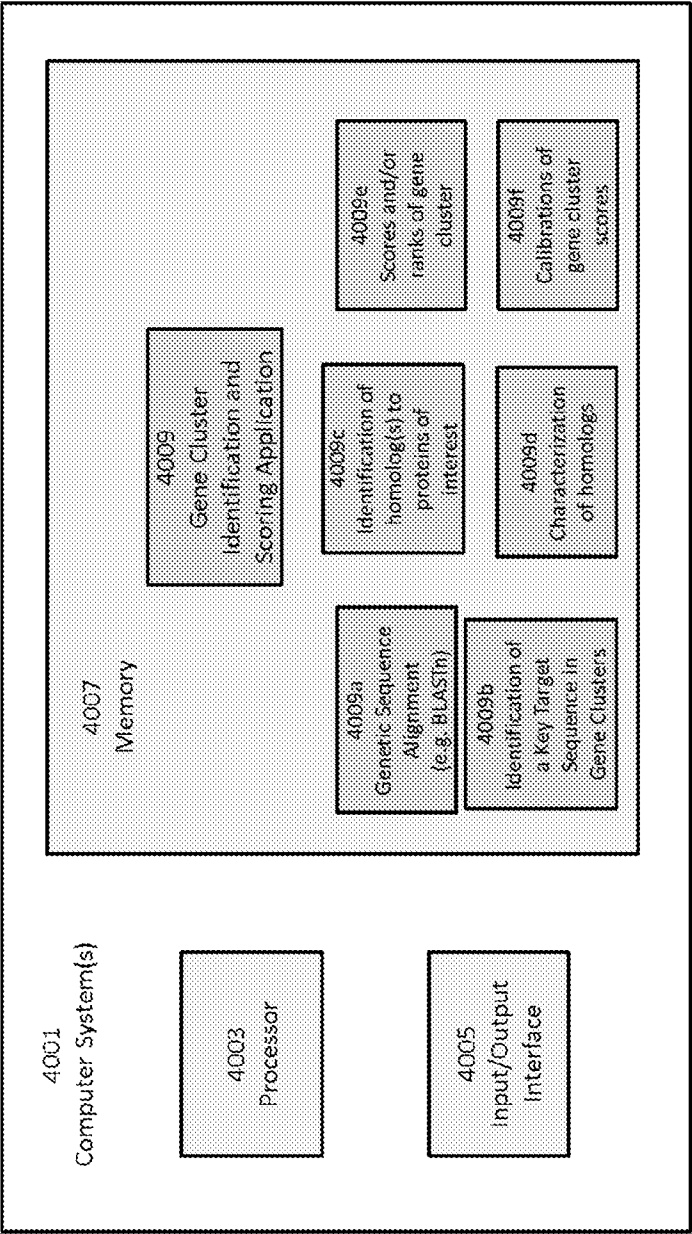


FIG. 1F

7 / 87

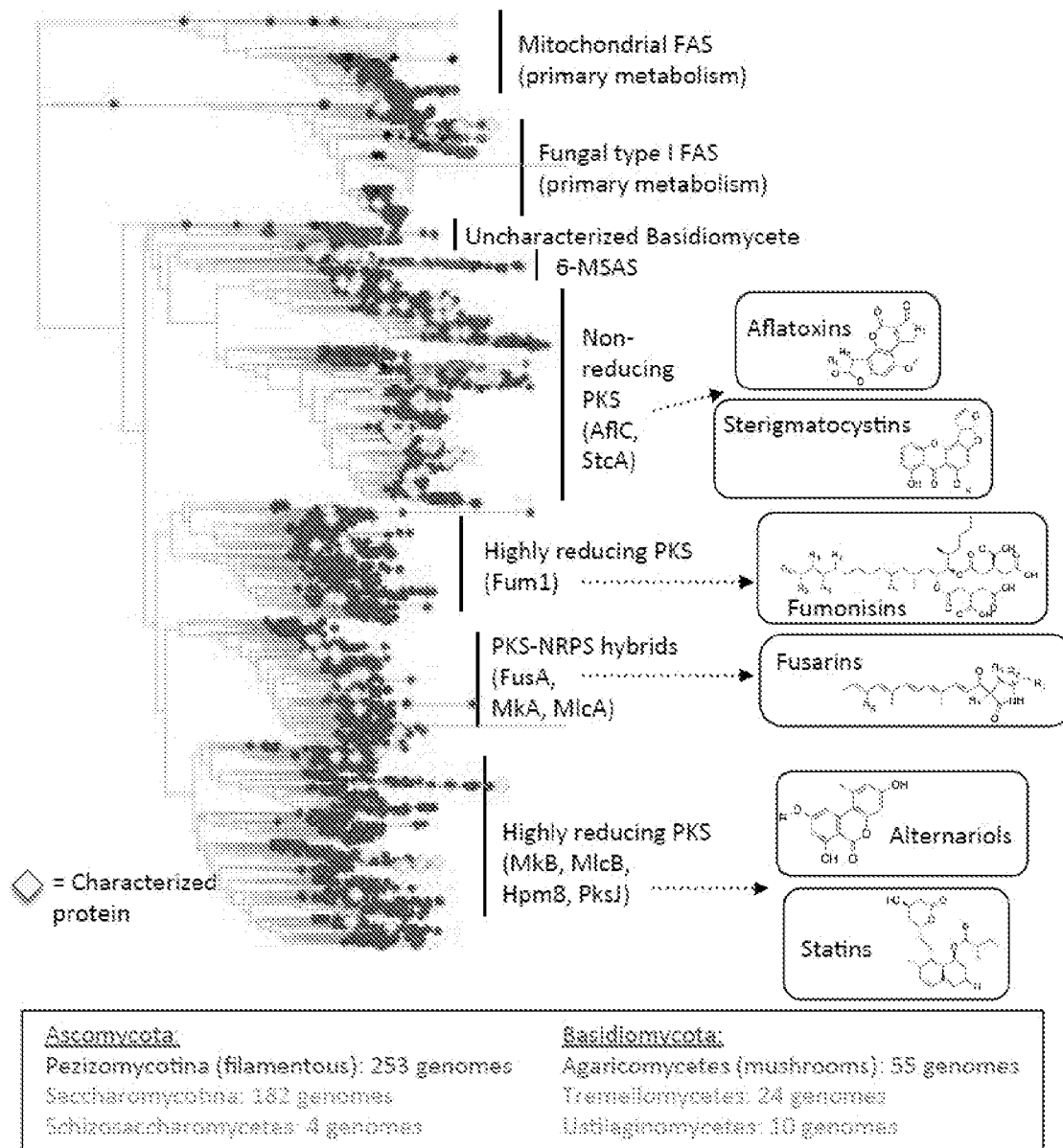
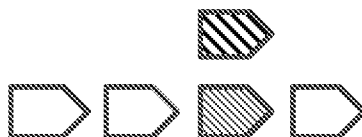


FIG. 2

8 / 87

Self-resistance mechanisms to toxic natural products



The hatched gene encodes mutated target protein, which is resistant to natural product

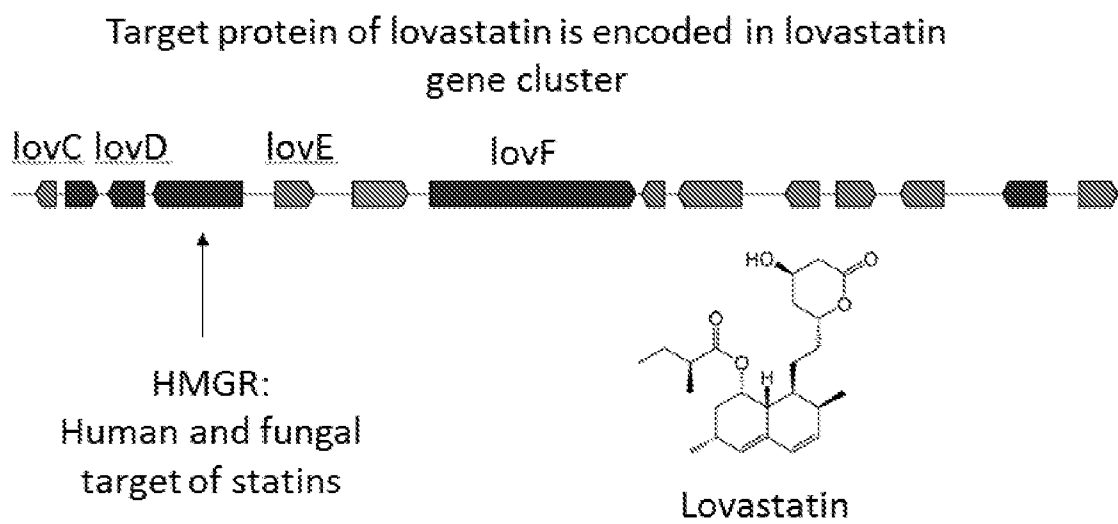


Striped gene encodes a protein that performs enzymatic modification of toxic small molecule into a nontoxic form



Dotted gene encodes a transporter protein that pumps the natural product out of the host cell

FIG. 3

**FIG. 4**

10 / 87

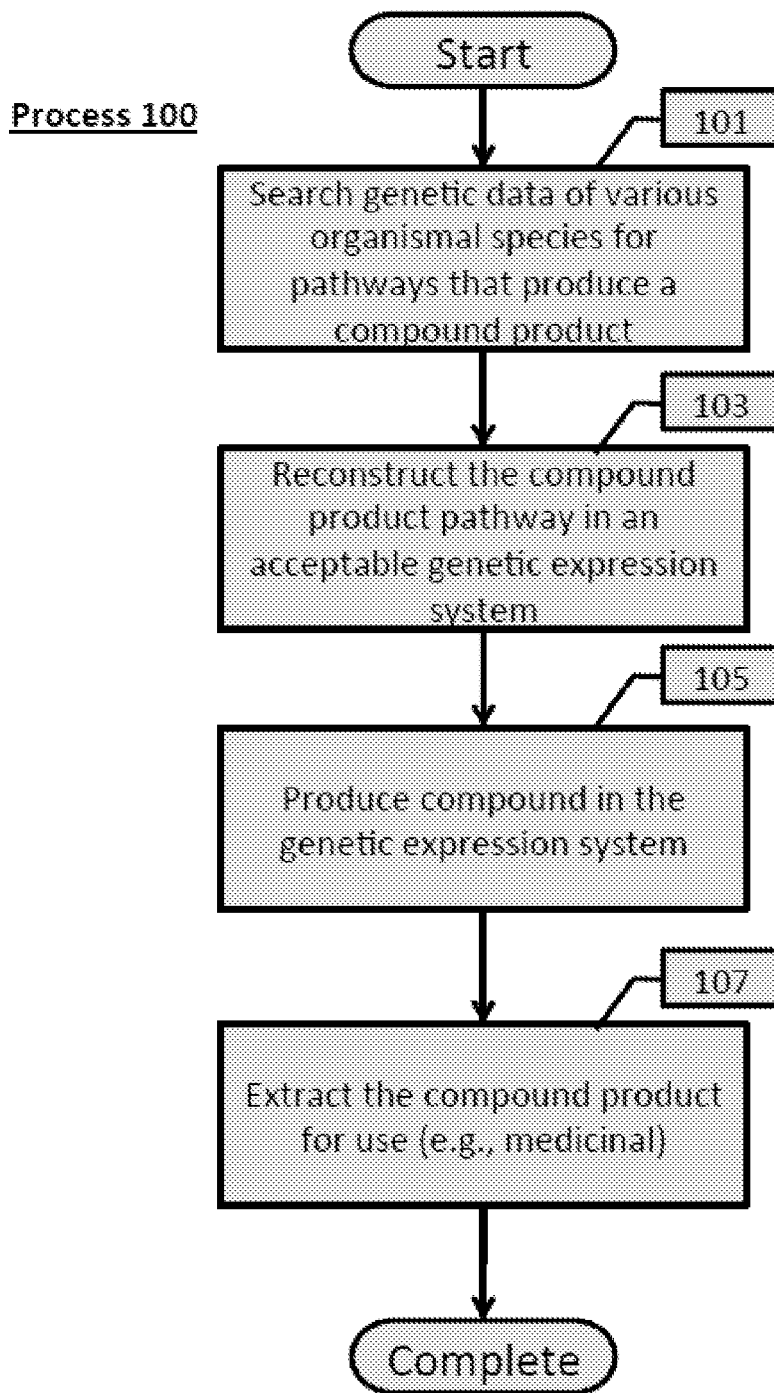


FIG. 5A

11 / 87

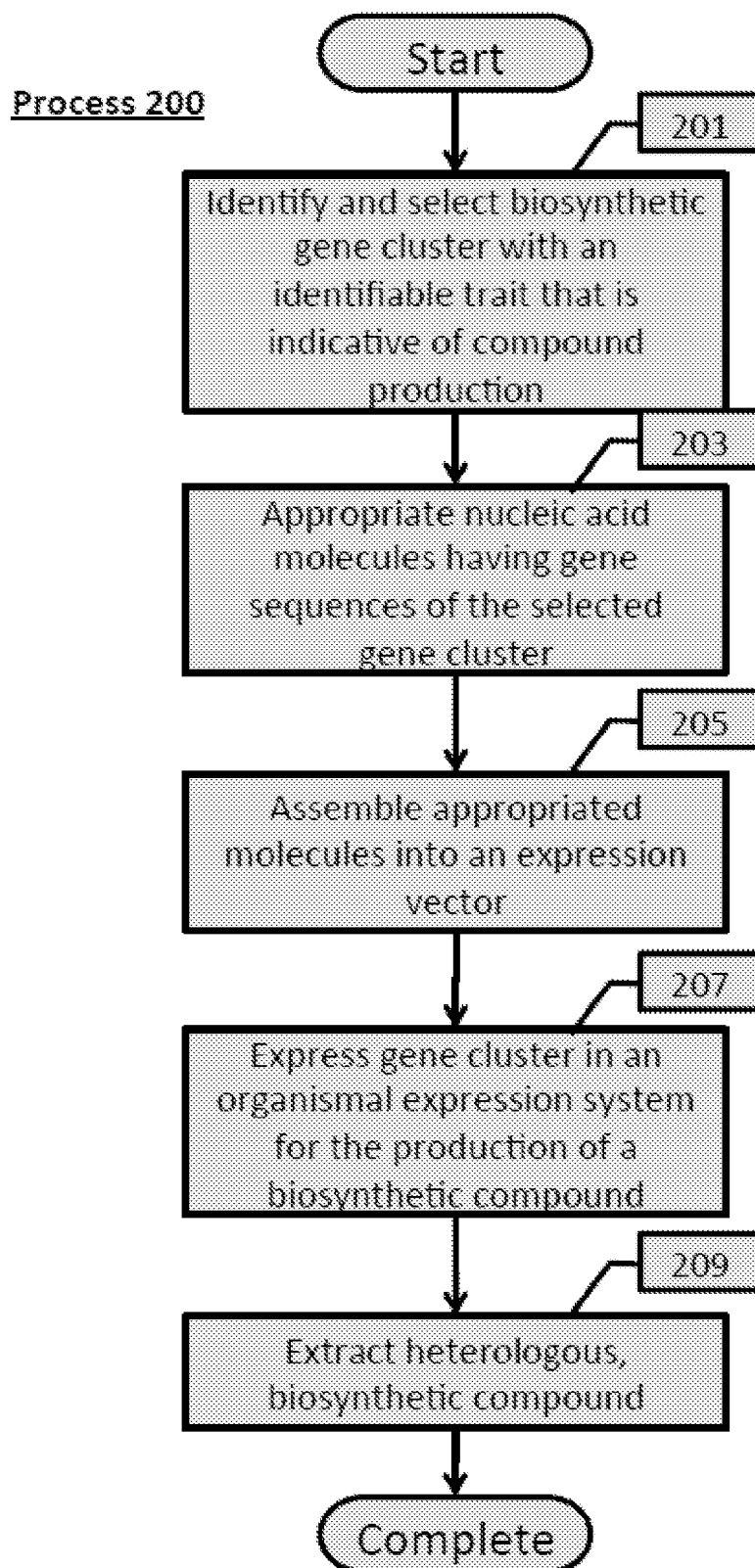
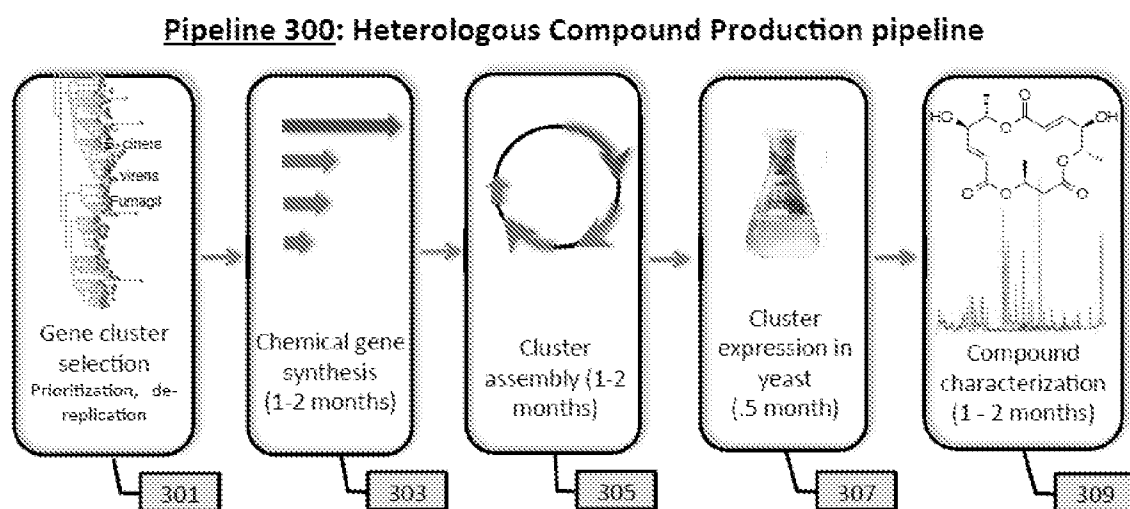


FIG. 5B

**FIG. 5C**

13 / 87

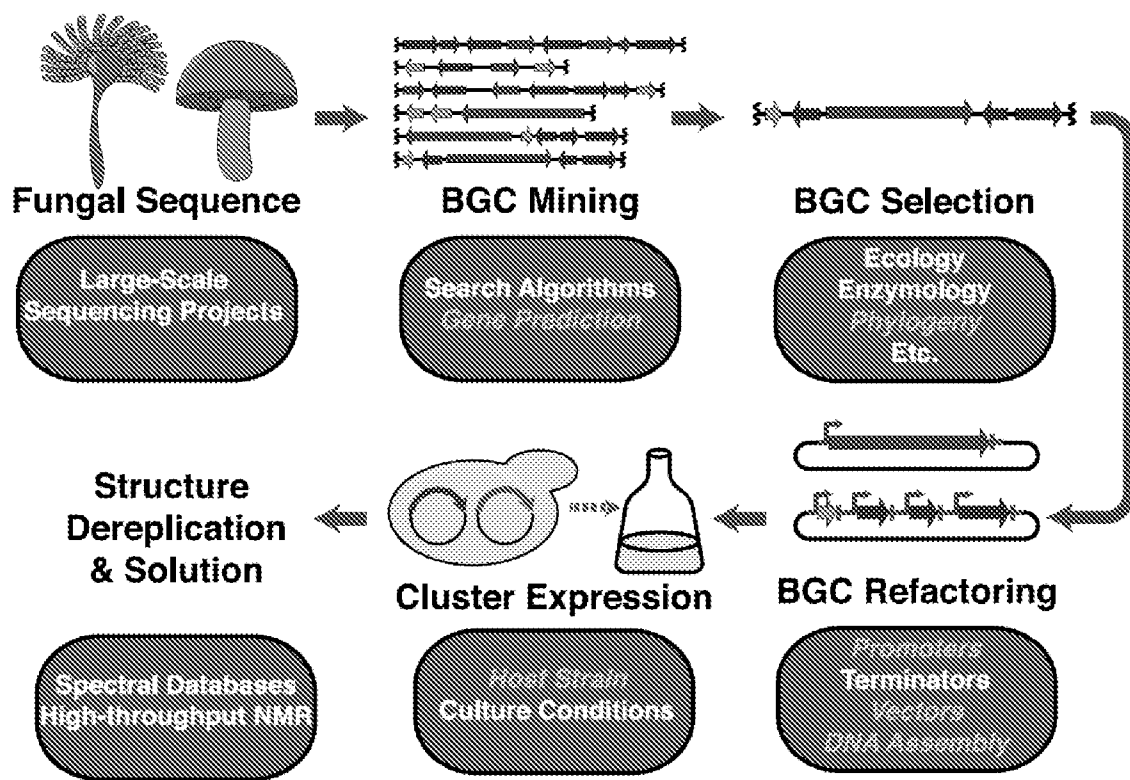


FIG. 5D

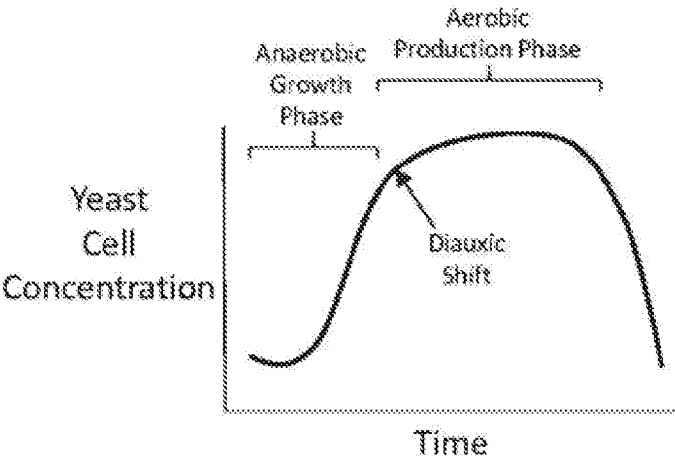


FIG. 6A

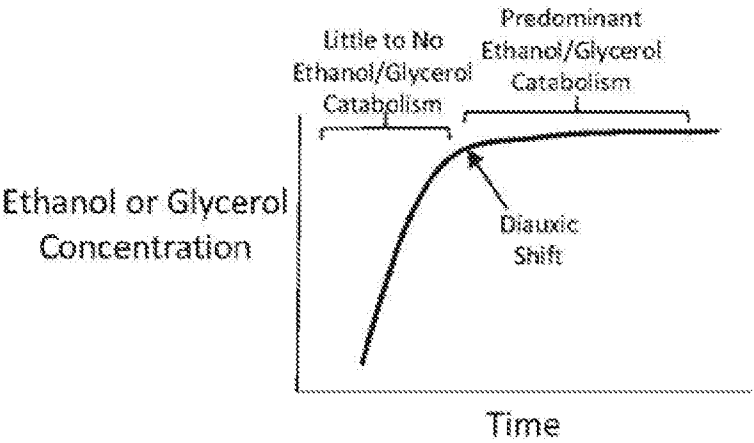


FIG. 6B

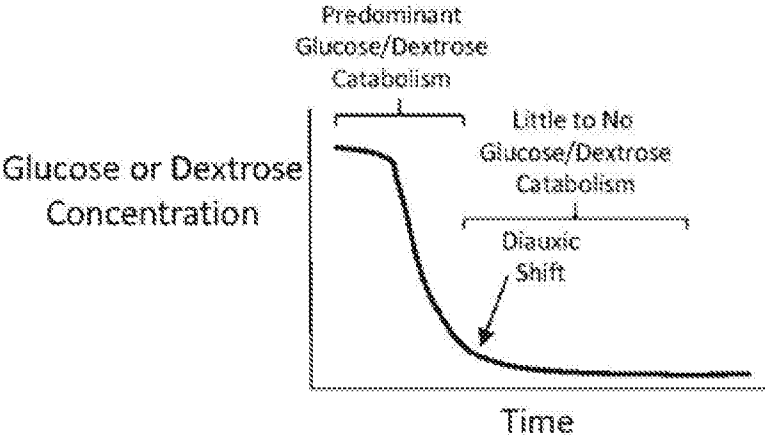


FIG. 6C

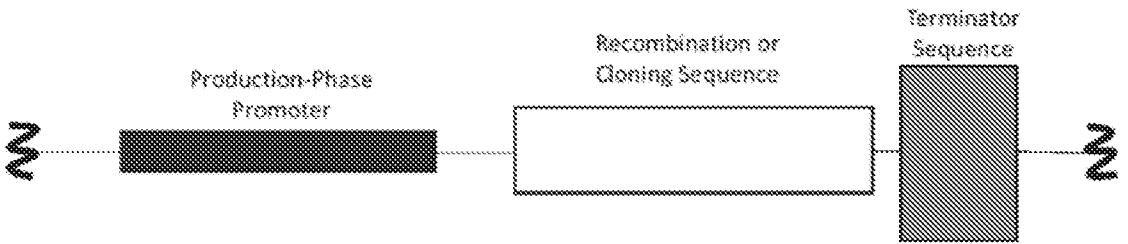


FIG. 7A

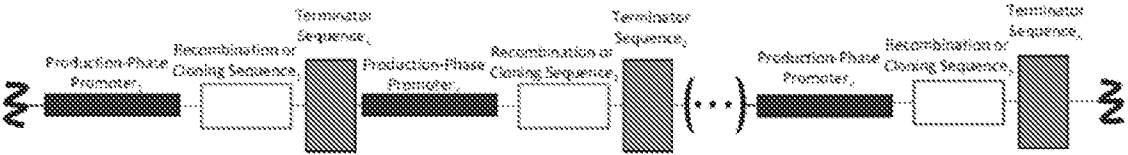


FIG. 7B

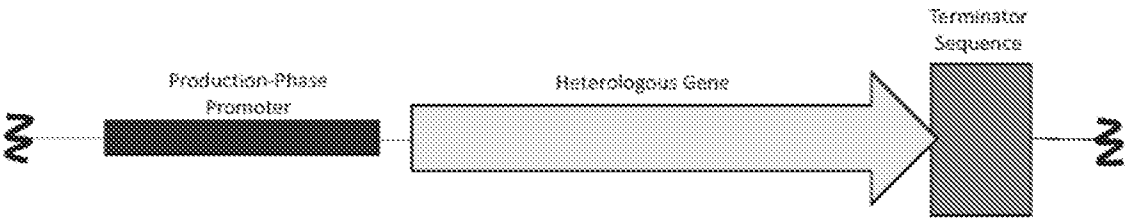


FIG. 8A

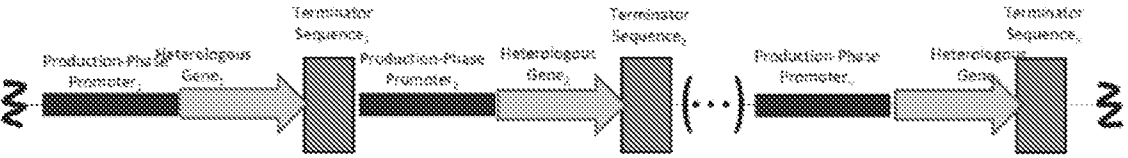
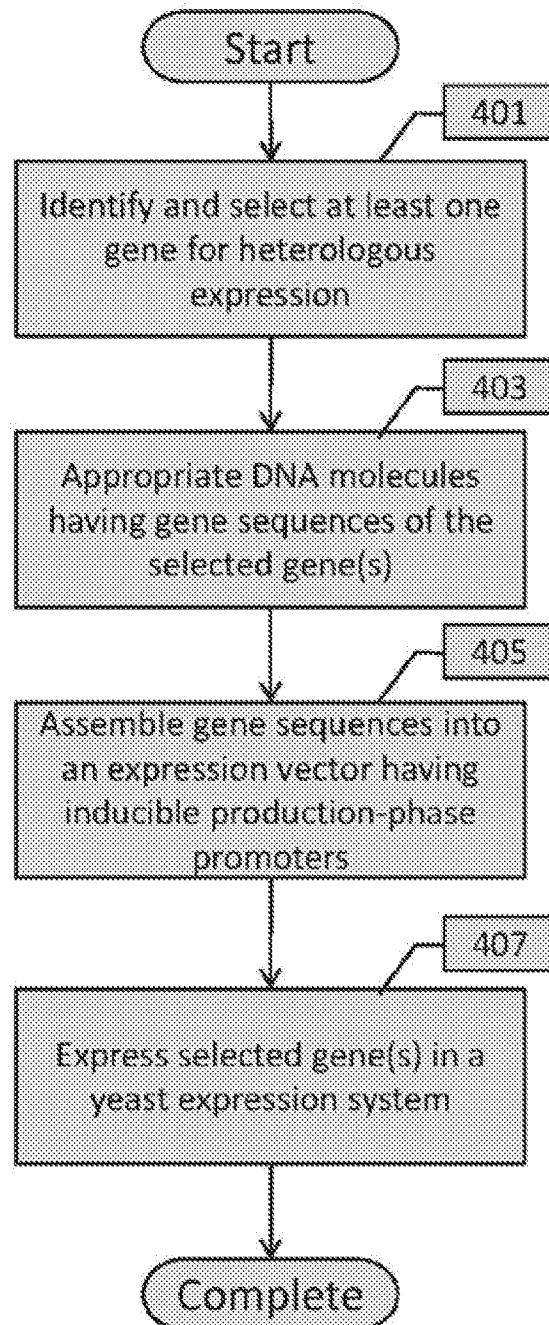


FIG. 8B

17 / 87

Process 400**FIG. 9**

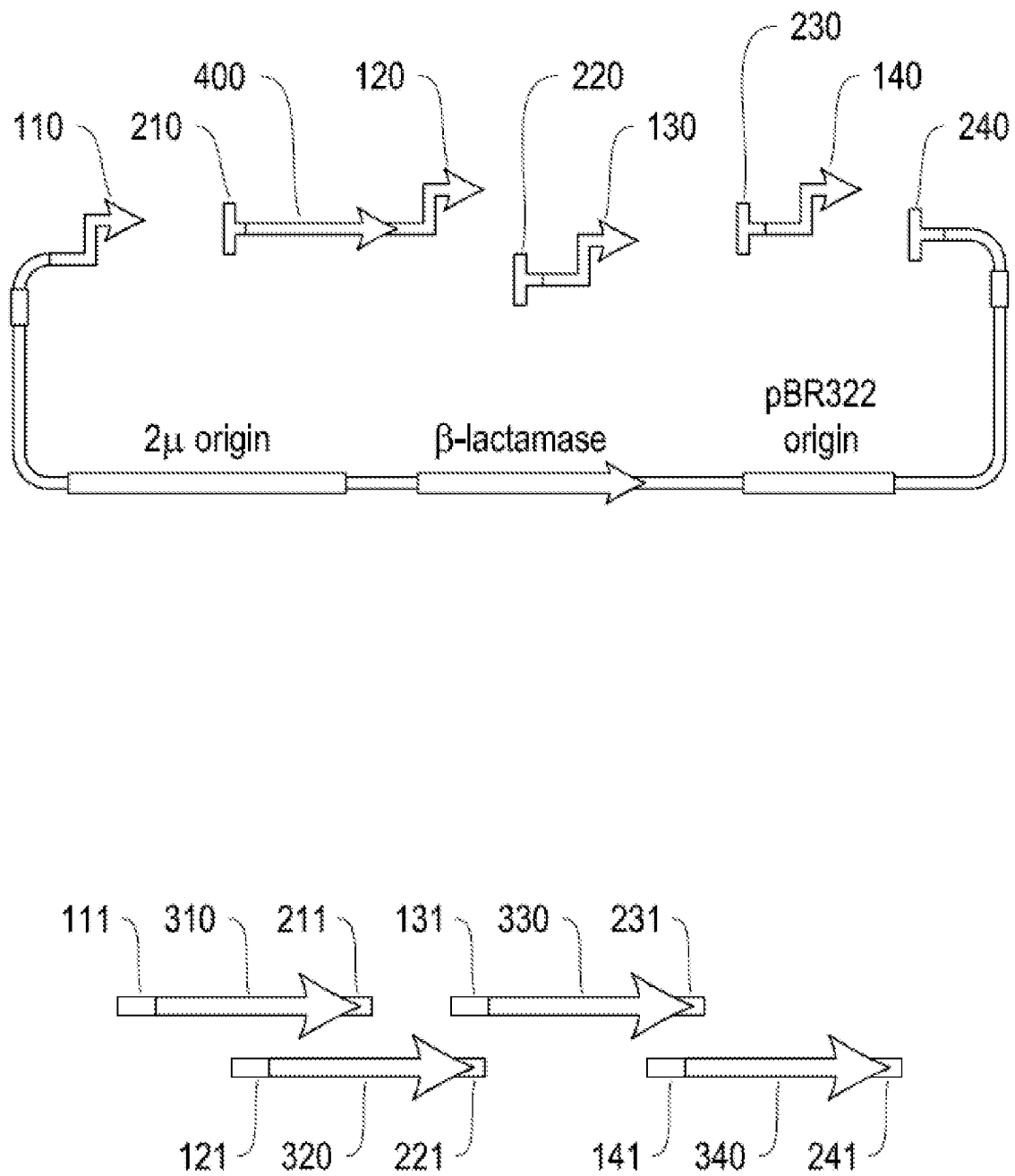
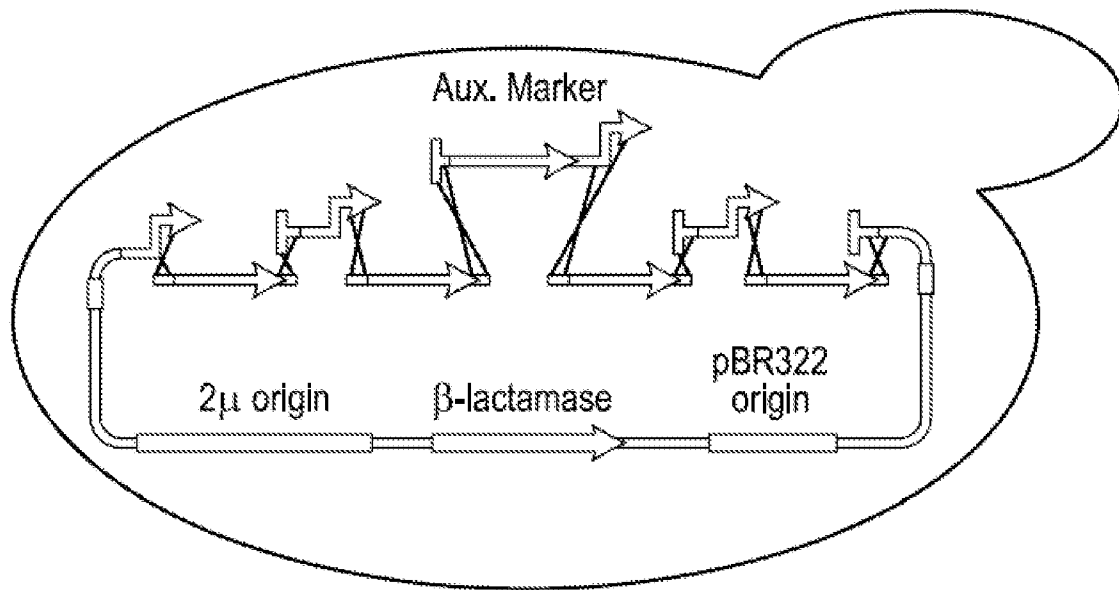


FIG. 10A

19 / 87



Assembly by YHR

FIG. 10B

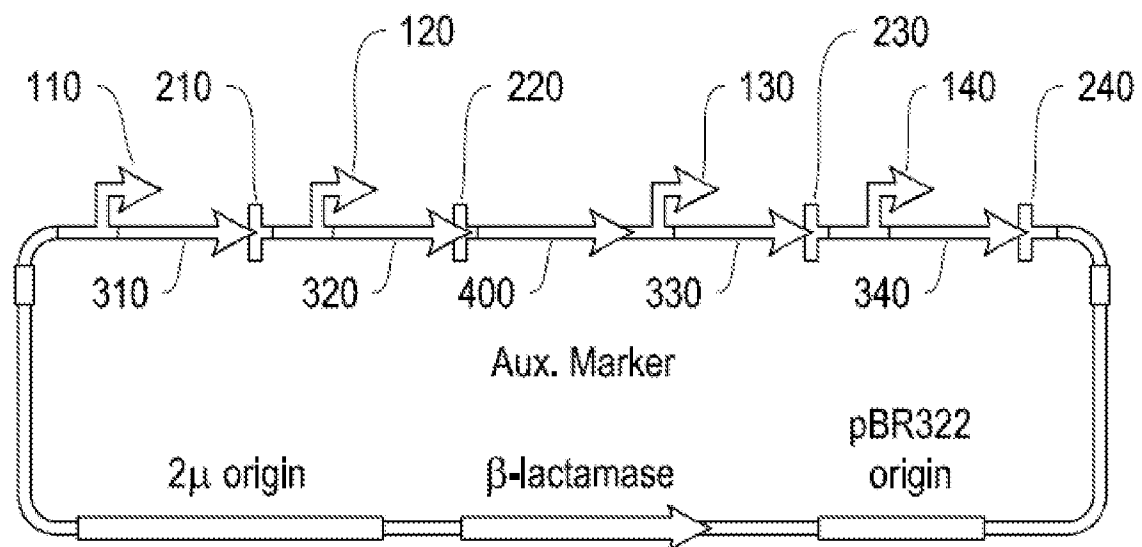


FIG. 10C

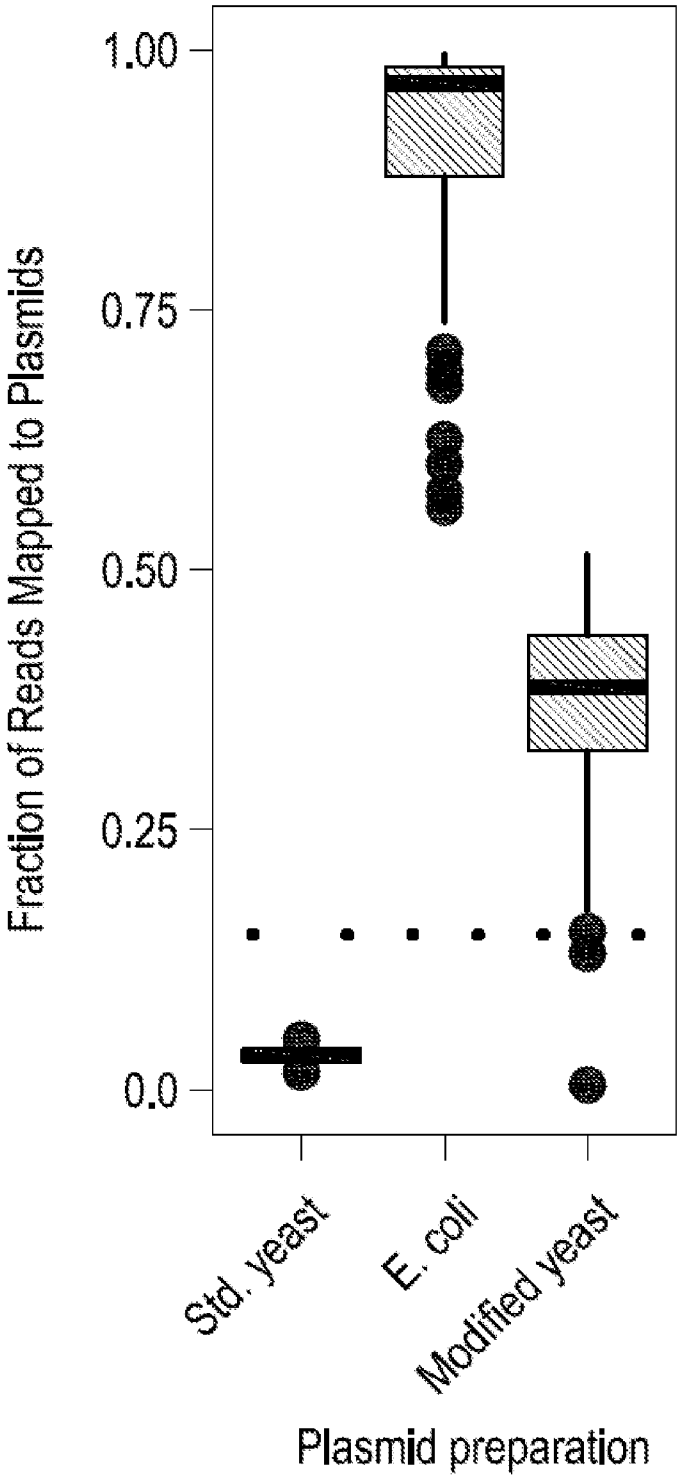


FIG. 10D

21 / 87

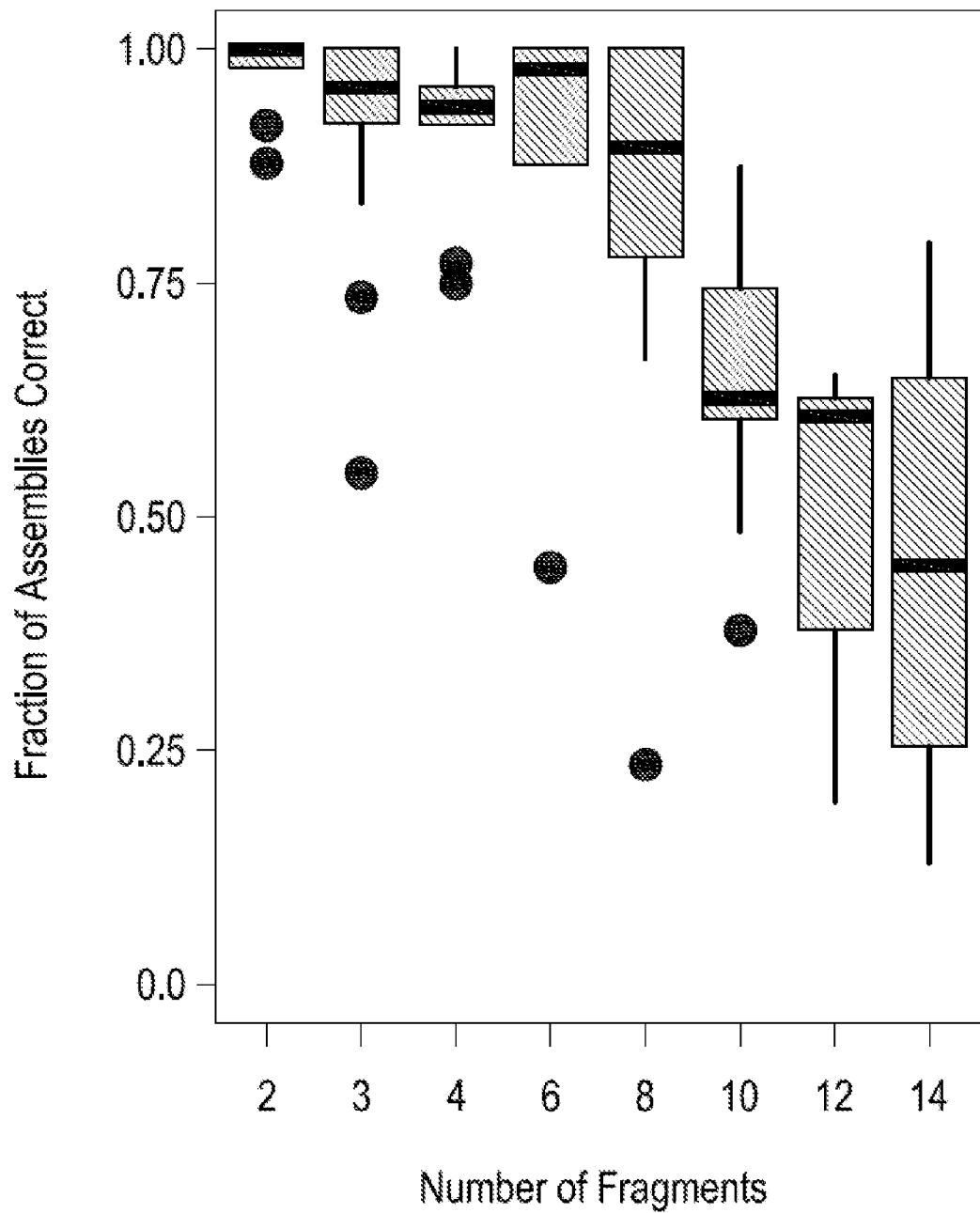


FIG. 10E

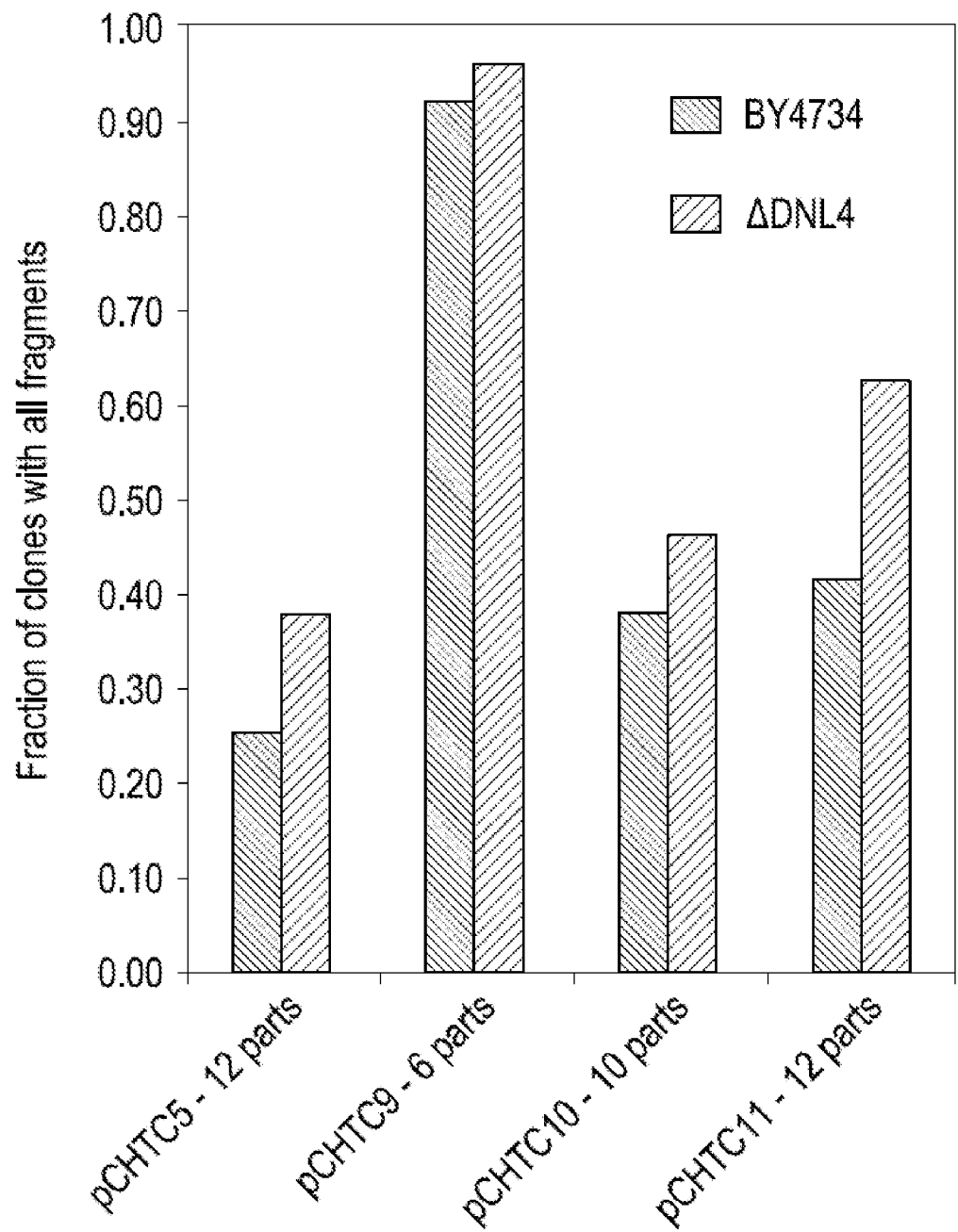


FIG. 11A

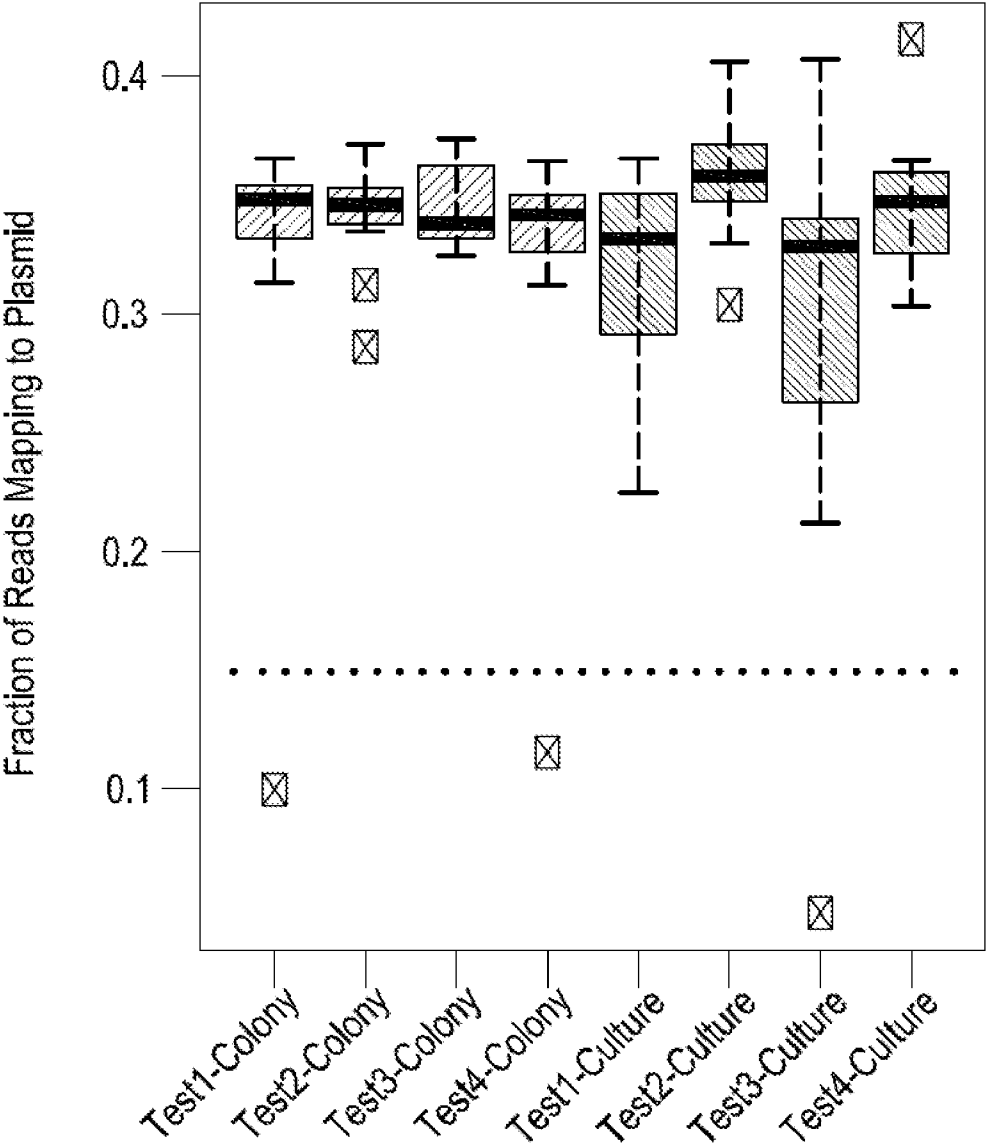


FIG. 11B

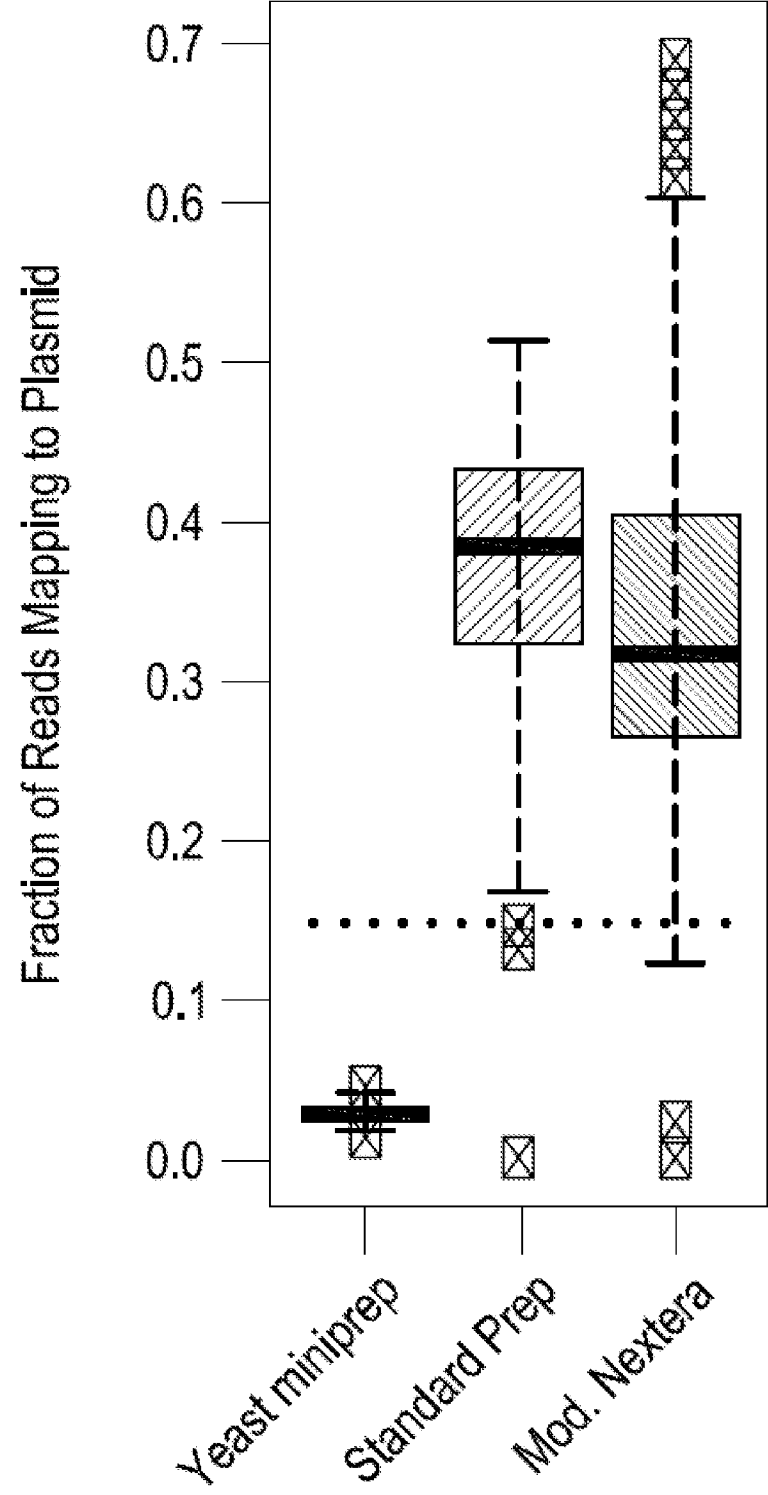


FIG. 11C

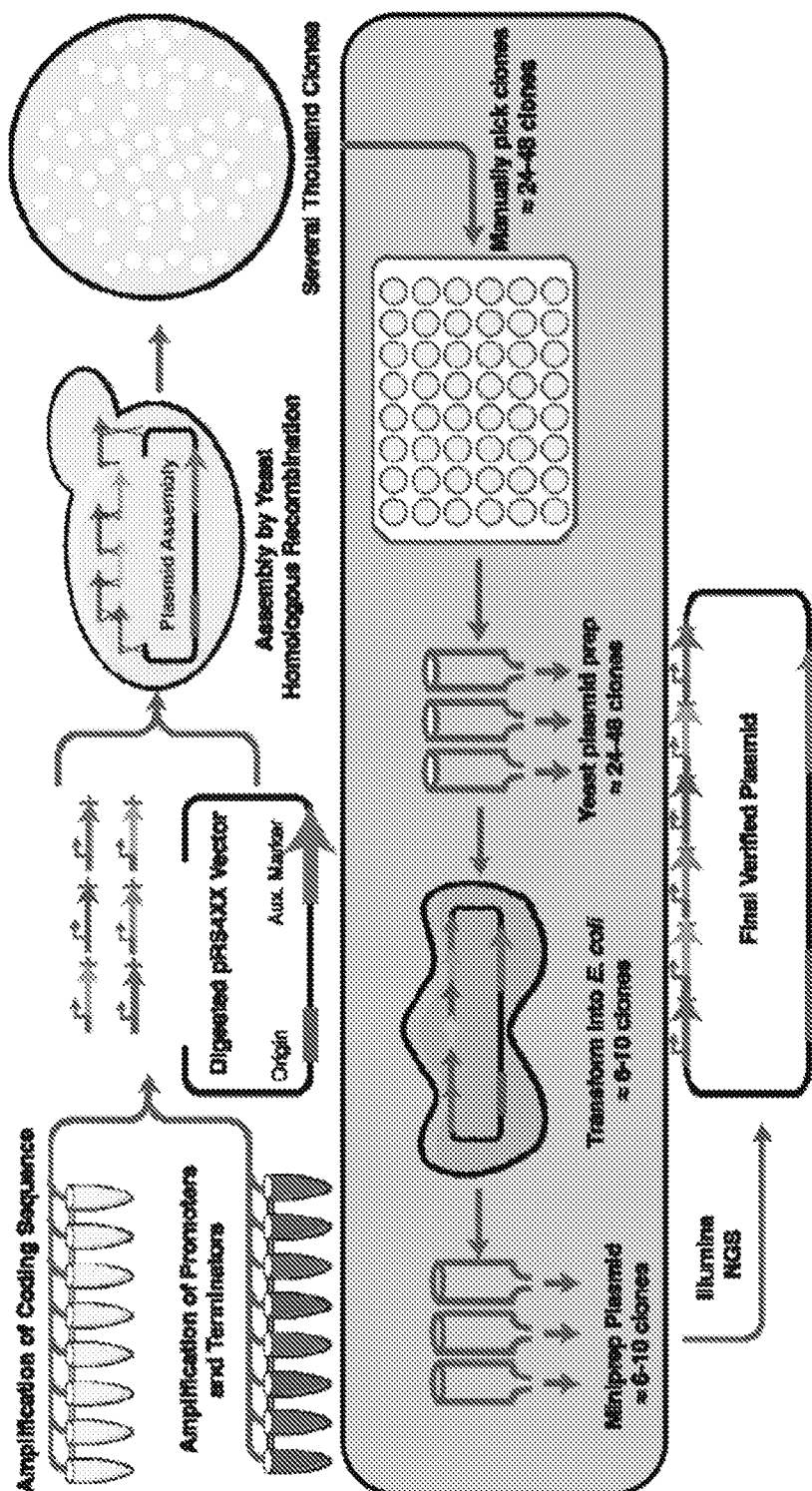


FIG. 11D

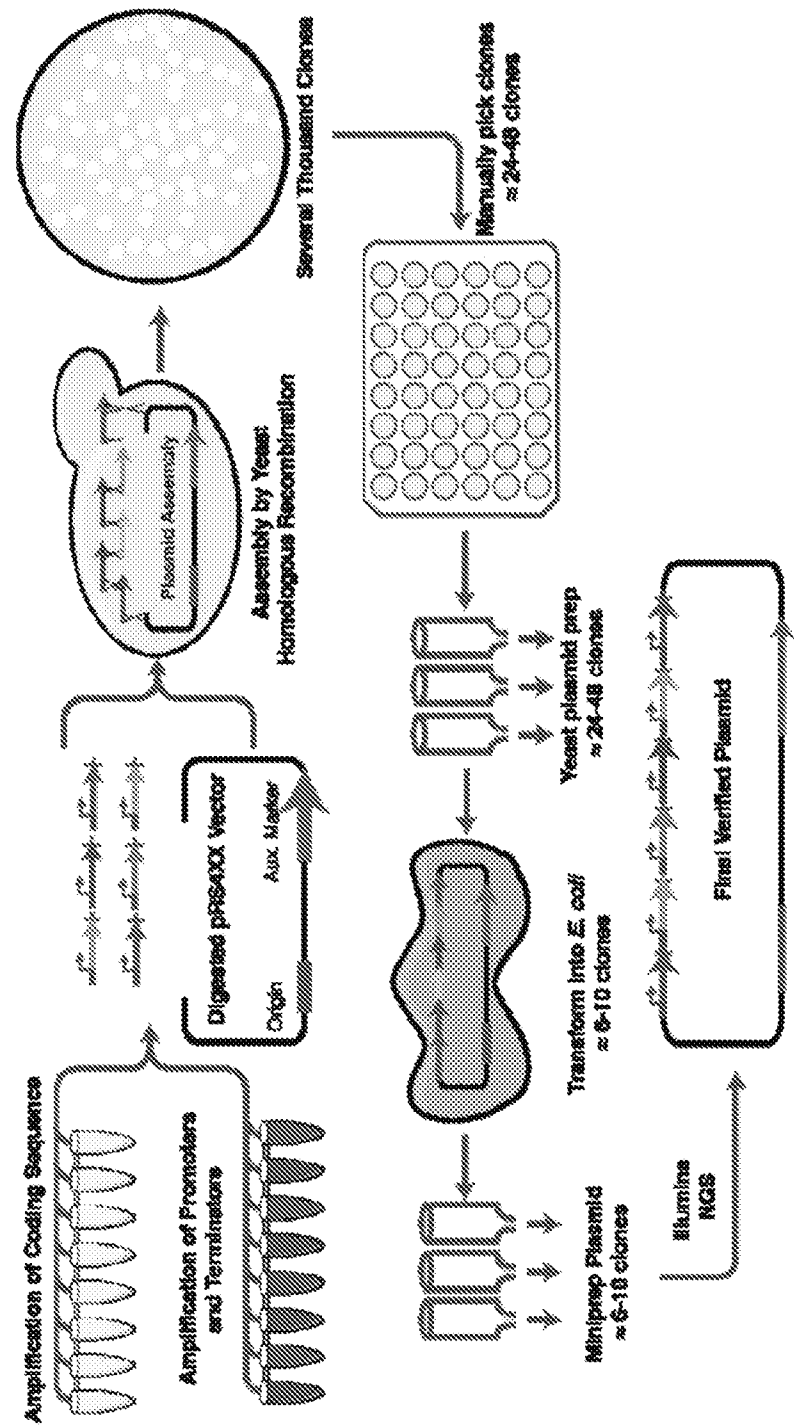


FIG. 11E

27 / 87

DHY Genotype

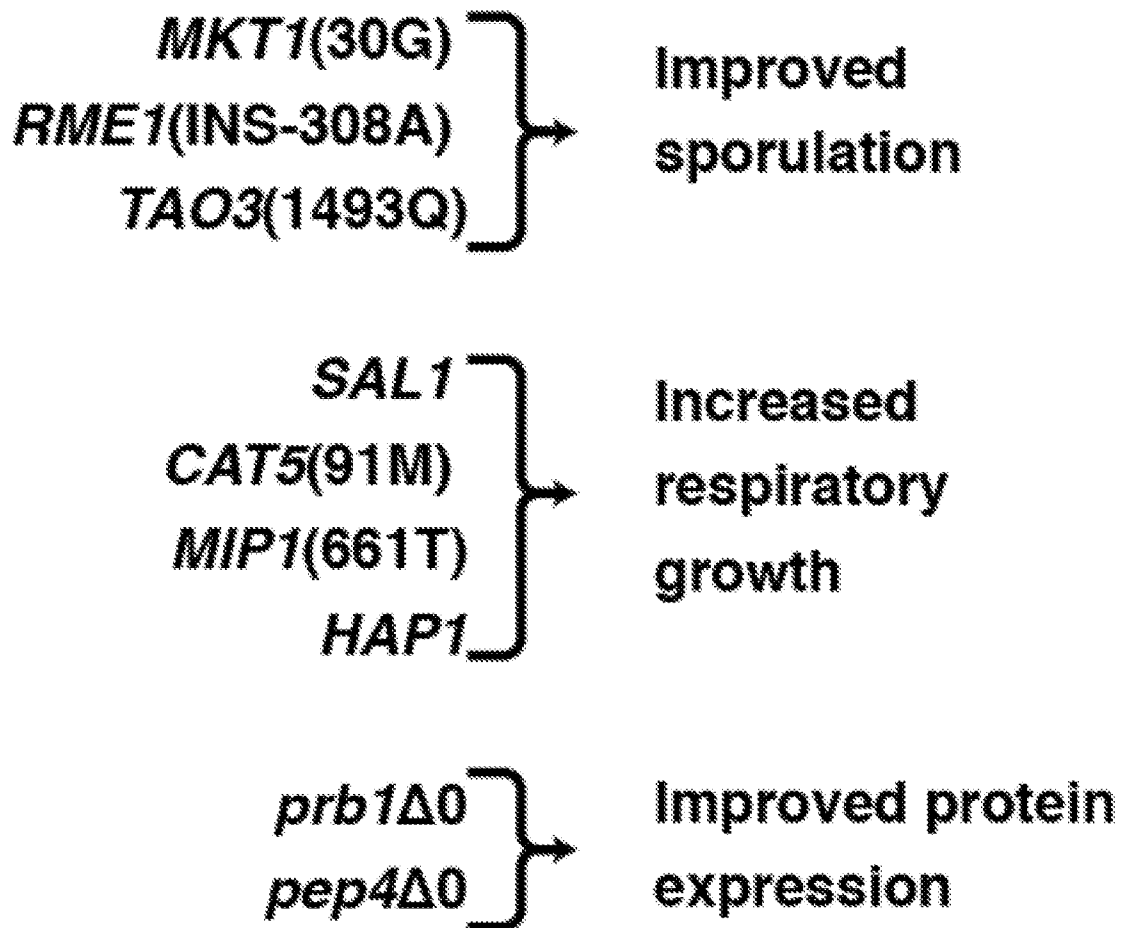


FIG. 12

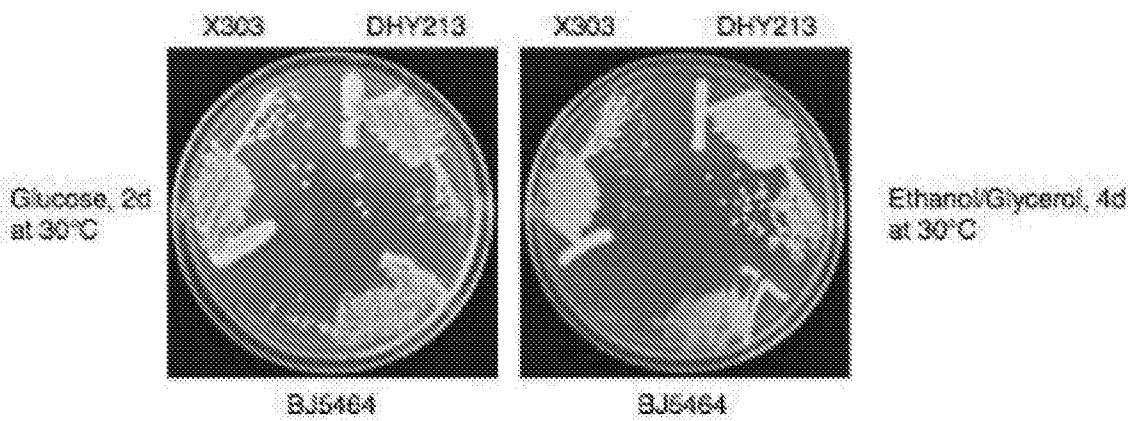


FIG. 13

Strain Growth

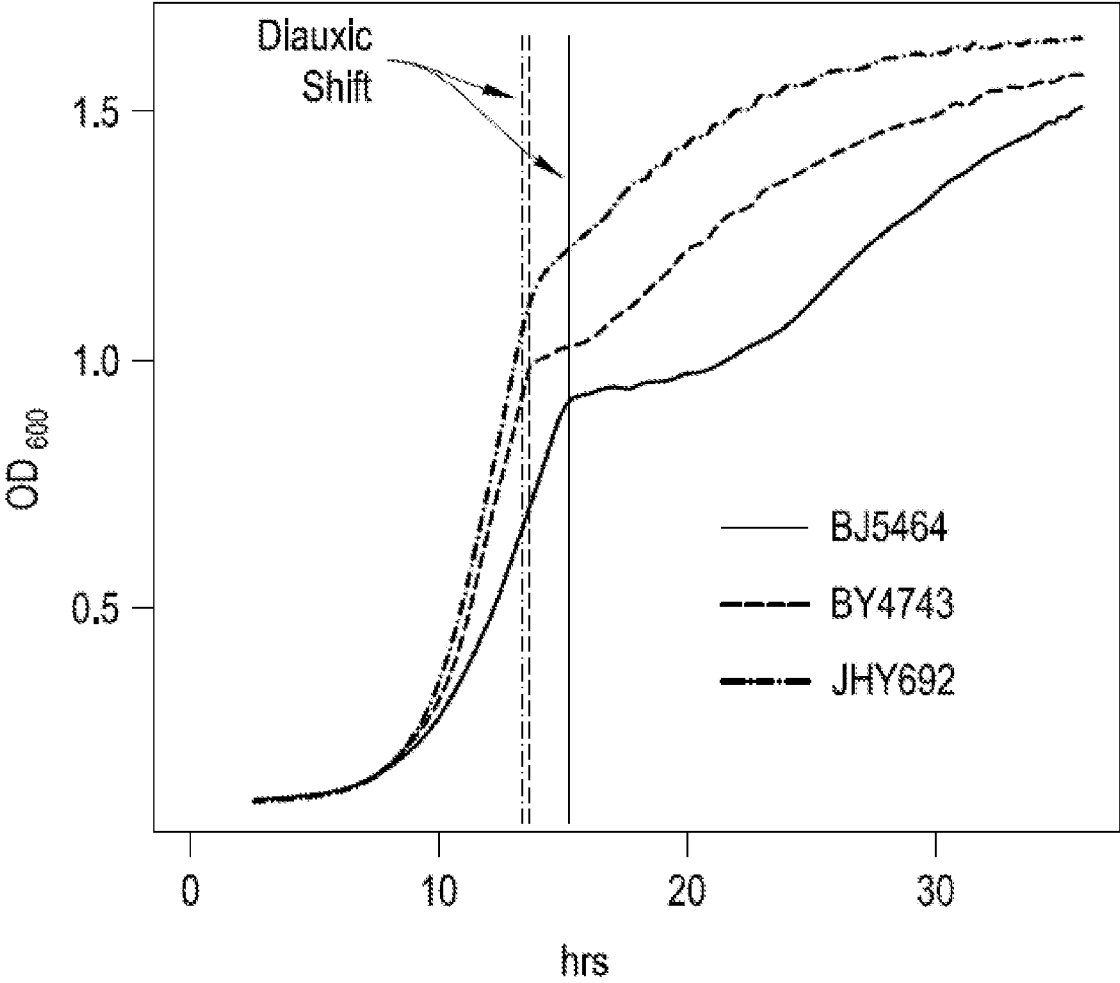


FIG. 14A

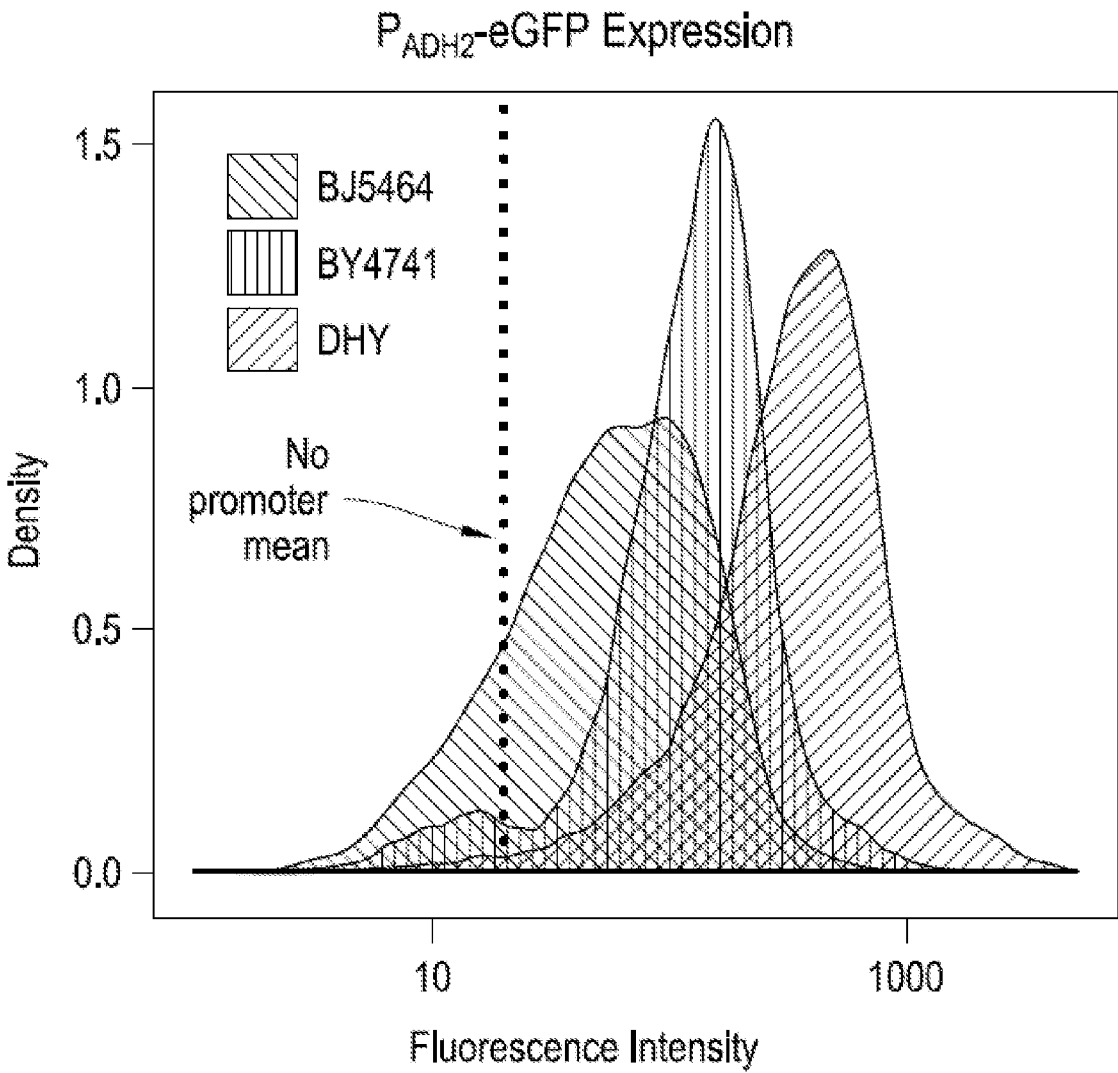


FIG. 14B

31 / 87

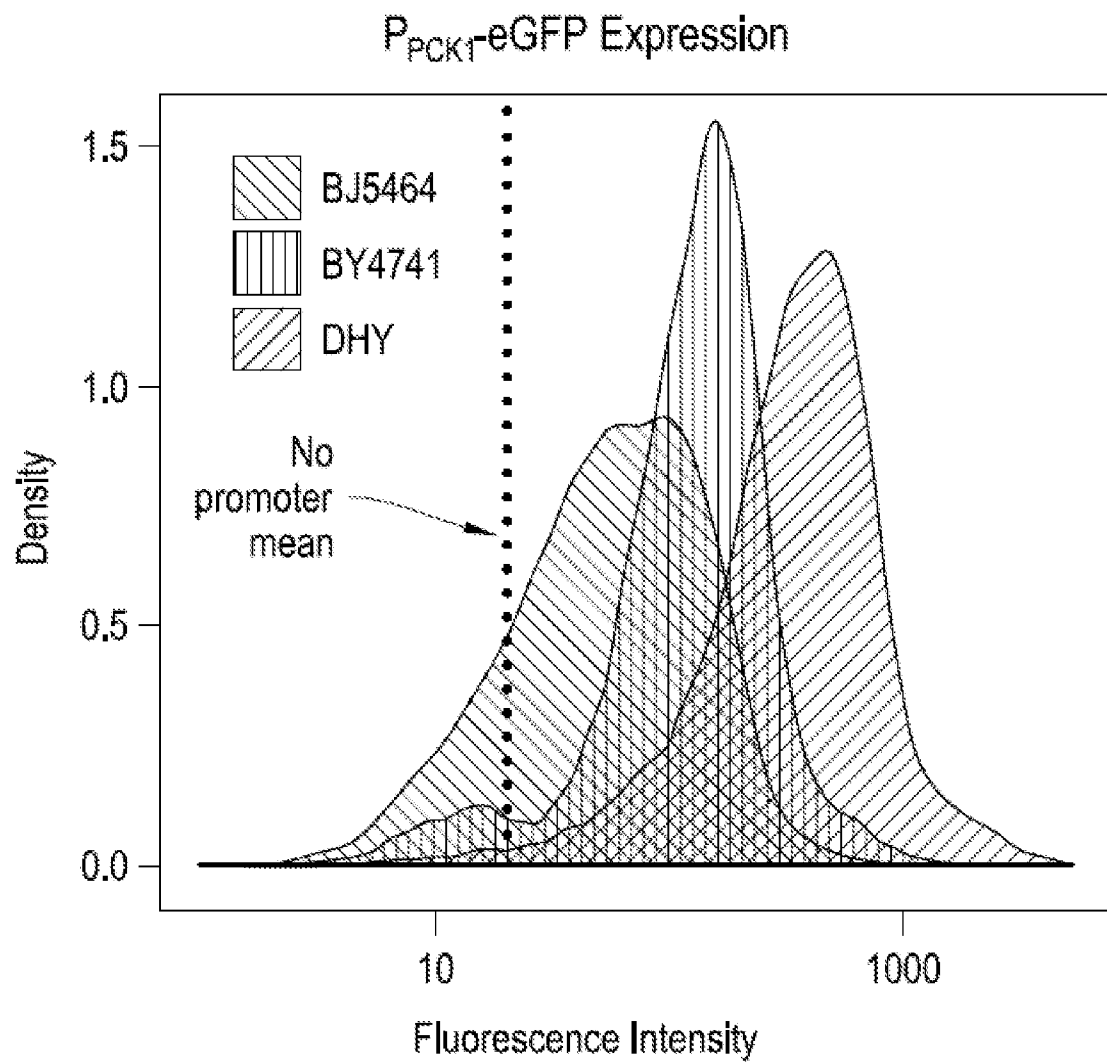


FIG. 14C

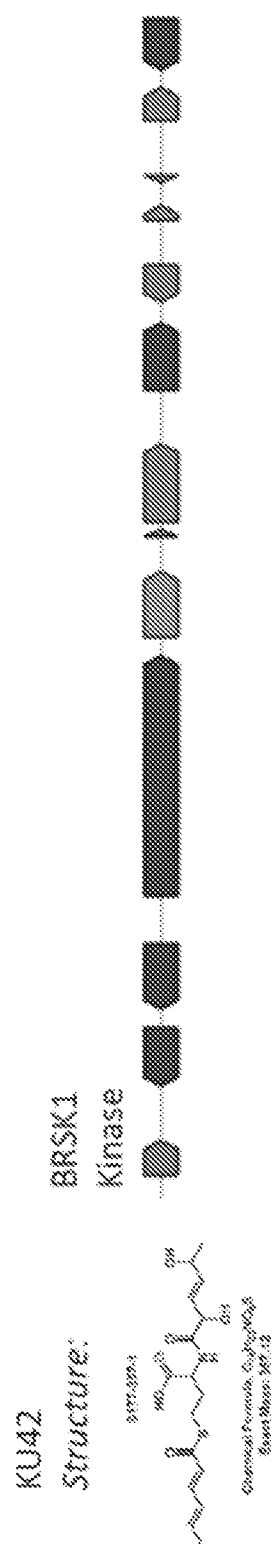
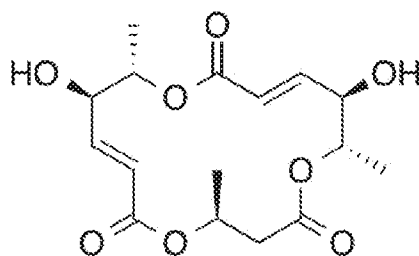
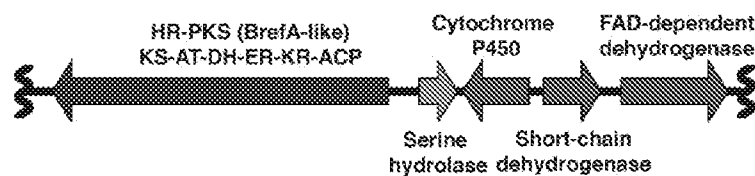


FIG. 15A

33 / 87

Example result: 5-gene pathway expressed
in yeast results in polyketide production



Ascomycota; Pezizomycotina;
Dothideomycetes; Pleosporomycetidae;
Pleosporales; Massarineae;
Didymosphaeriaceae; Paraphaeosphaeria

Microsphaeropsis

Ascomycota; Pezizomycotina; Dothideomycetes;
Pleosporomycetidae; Pleosporales; Pleosporineae;
Pleosporineae incertae sedis; *Microsphaeropsis*.

FIG. 15B

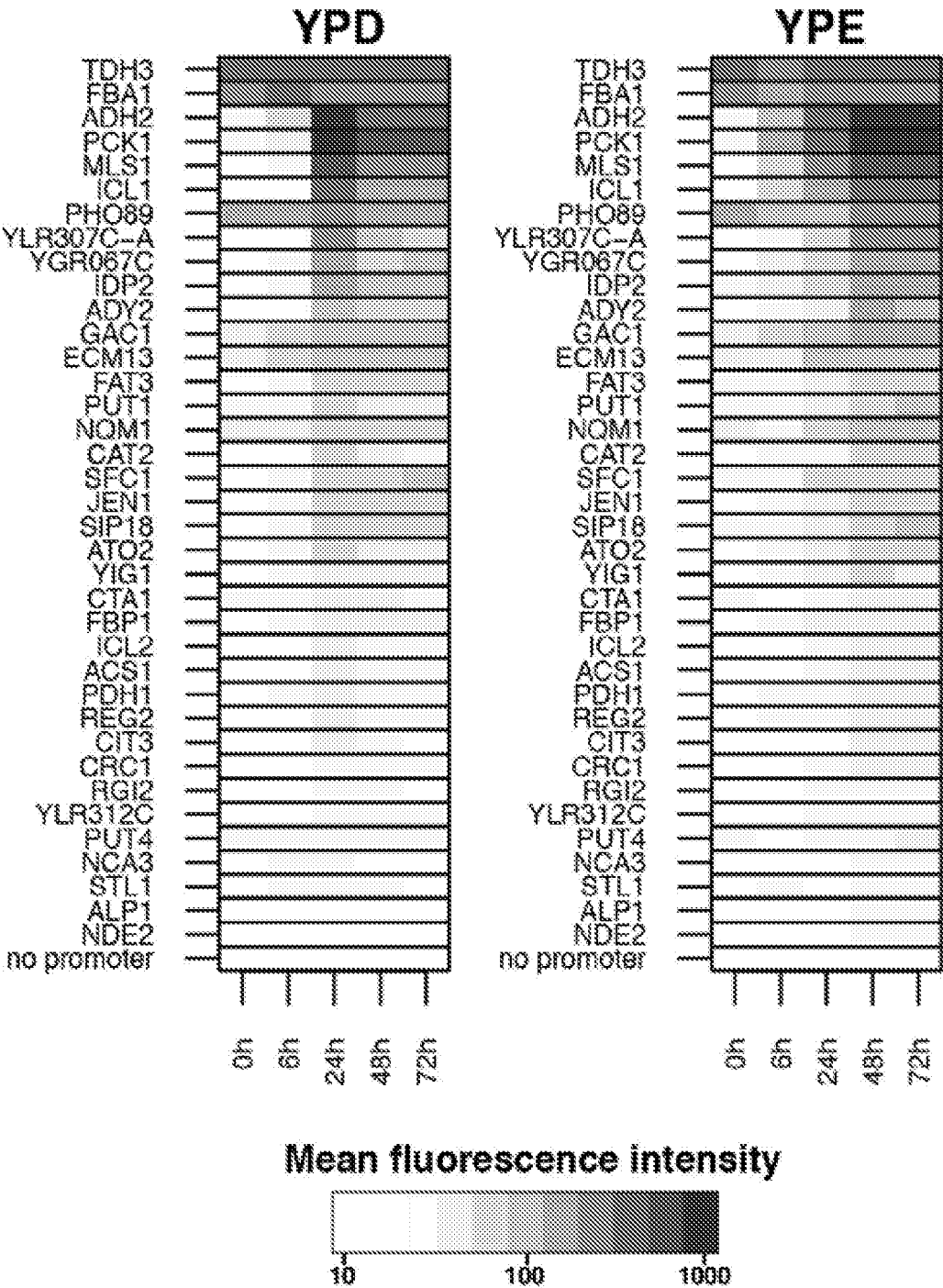


FIG. 16

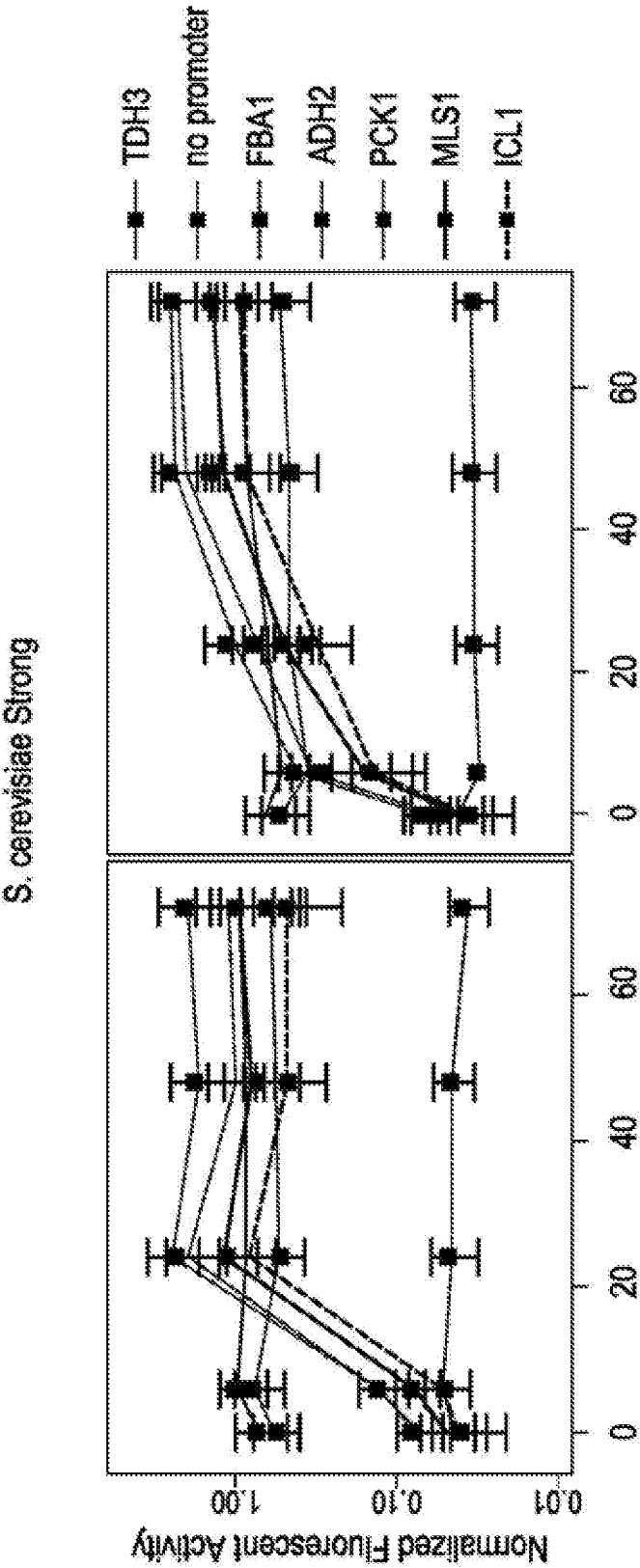


FIG. 17 (part 1/3)

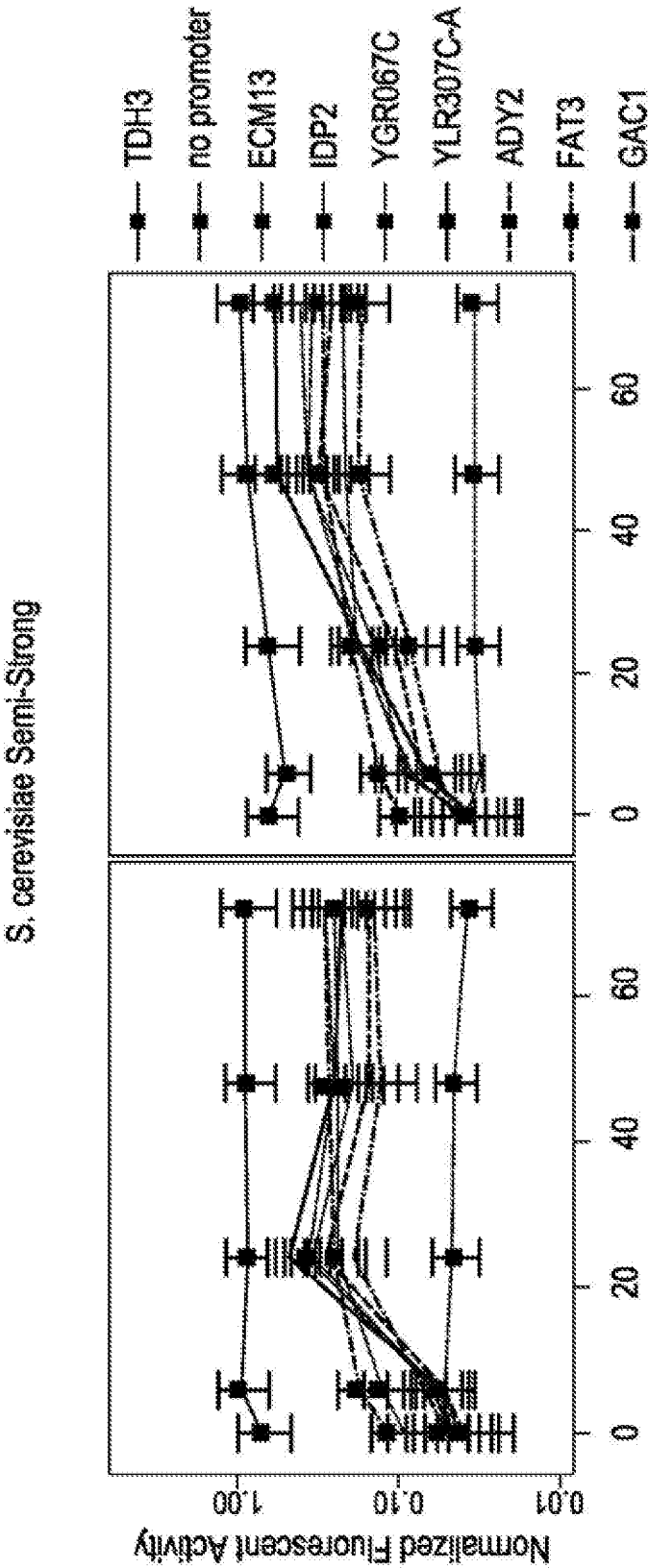


FIG. 17 (part 2/3)

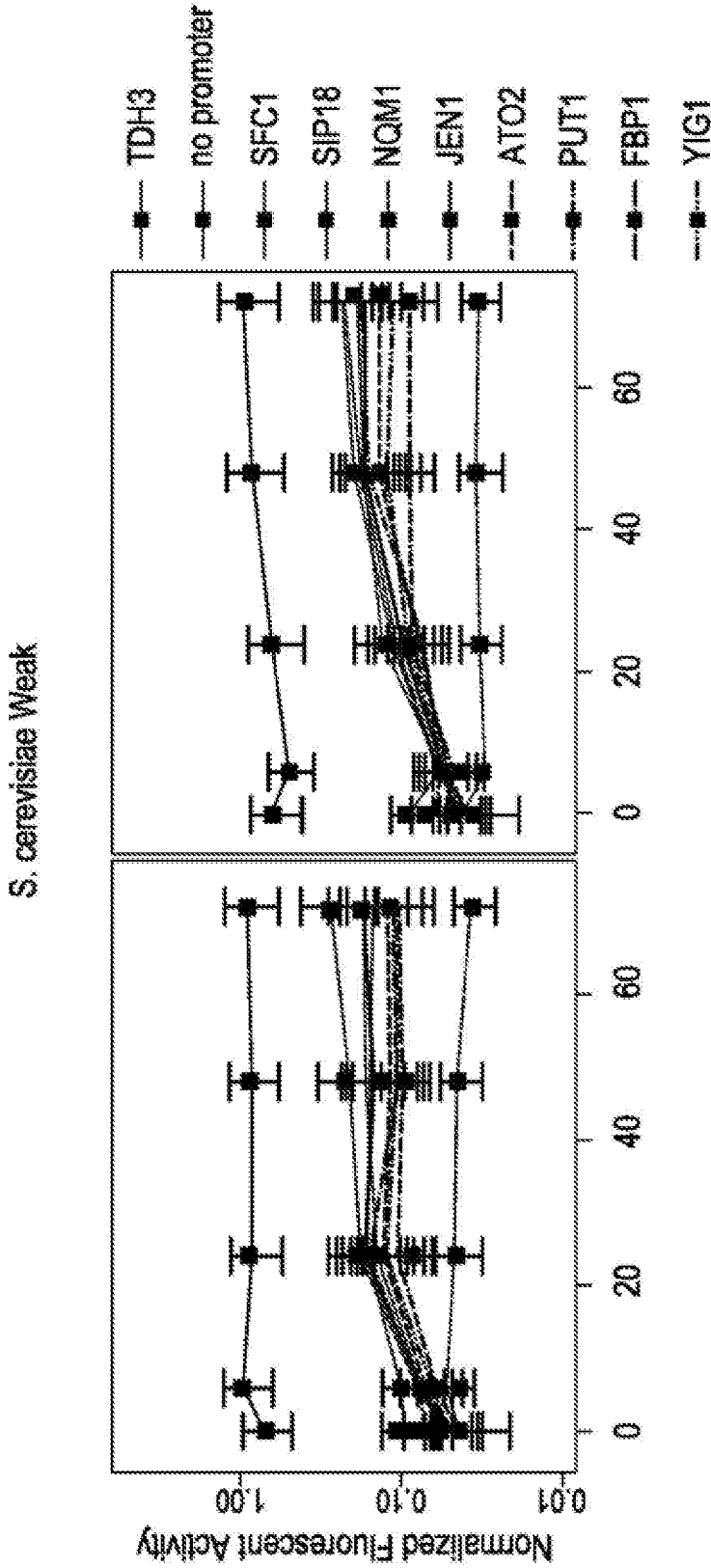


FIG. 17 (part 3/3)

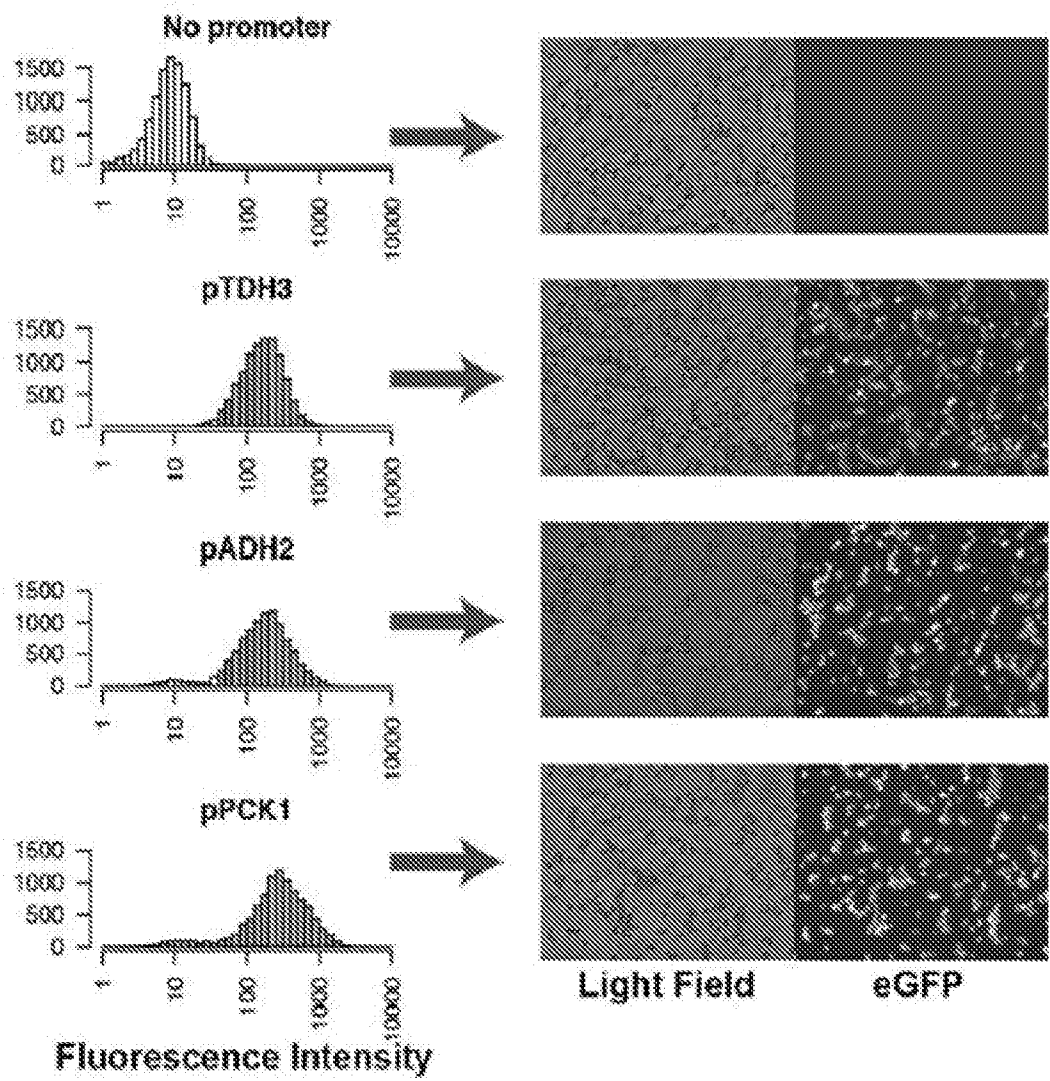


FIG. 18

39 / 87

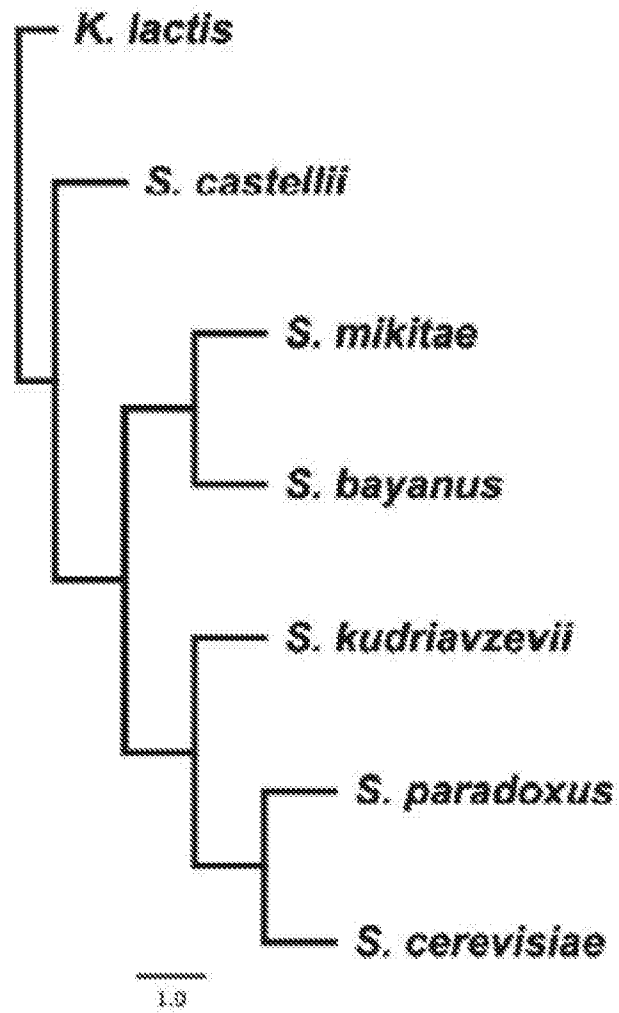


FIG. 19

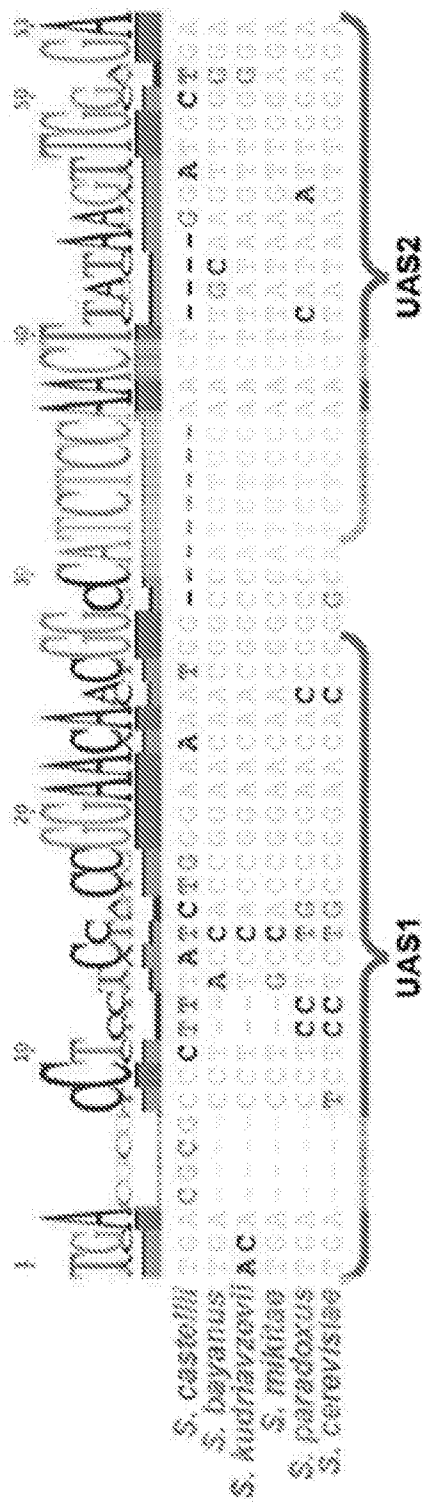


FIG. 20

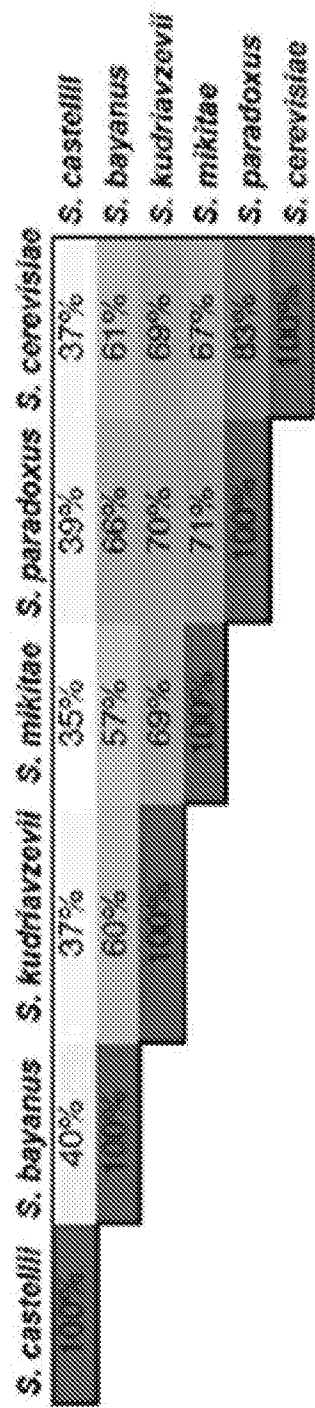


FIG. 21

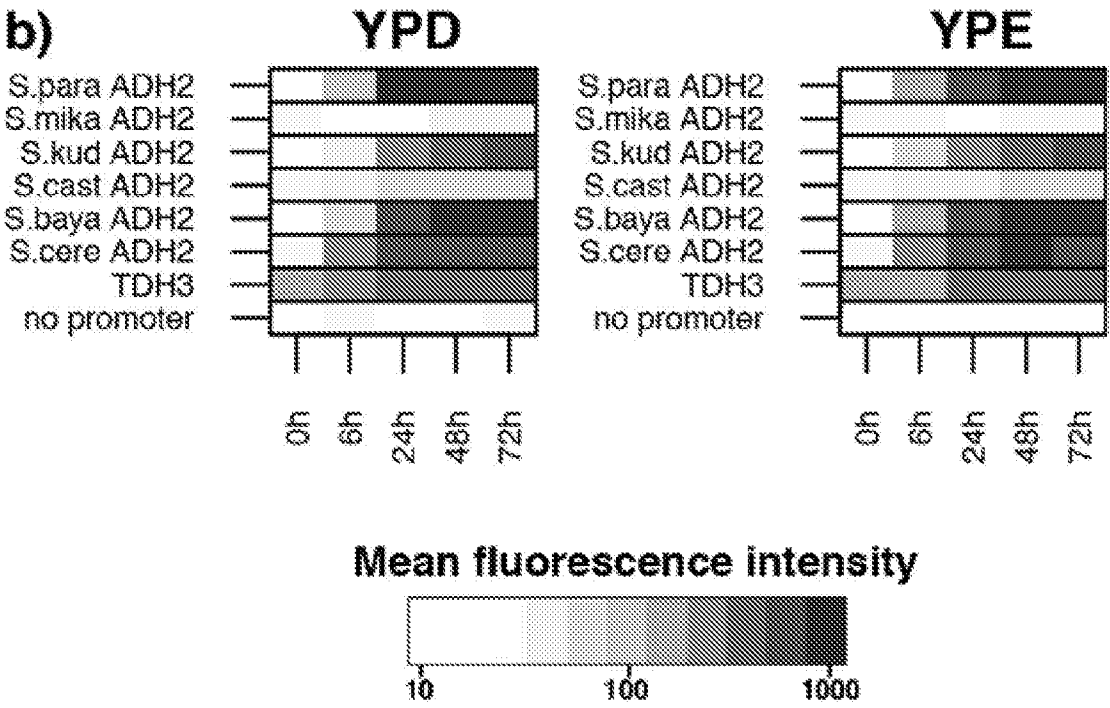


FIG. 22

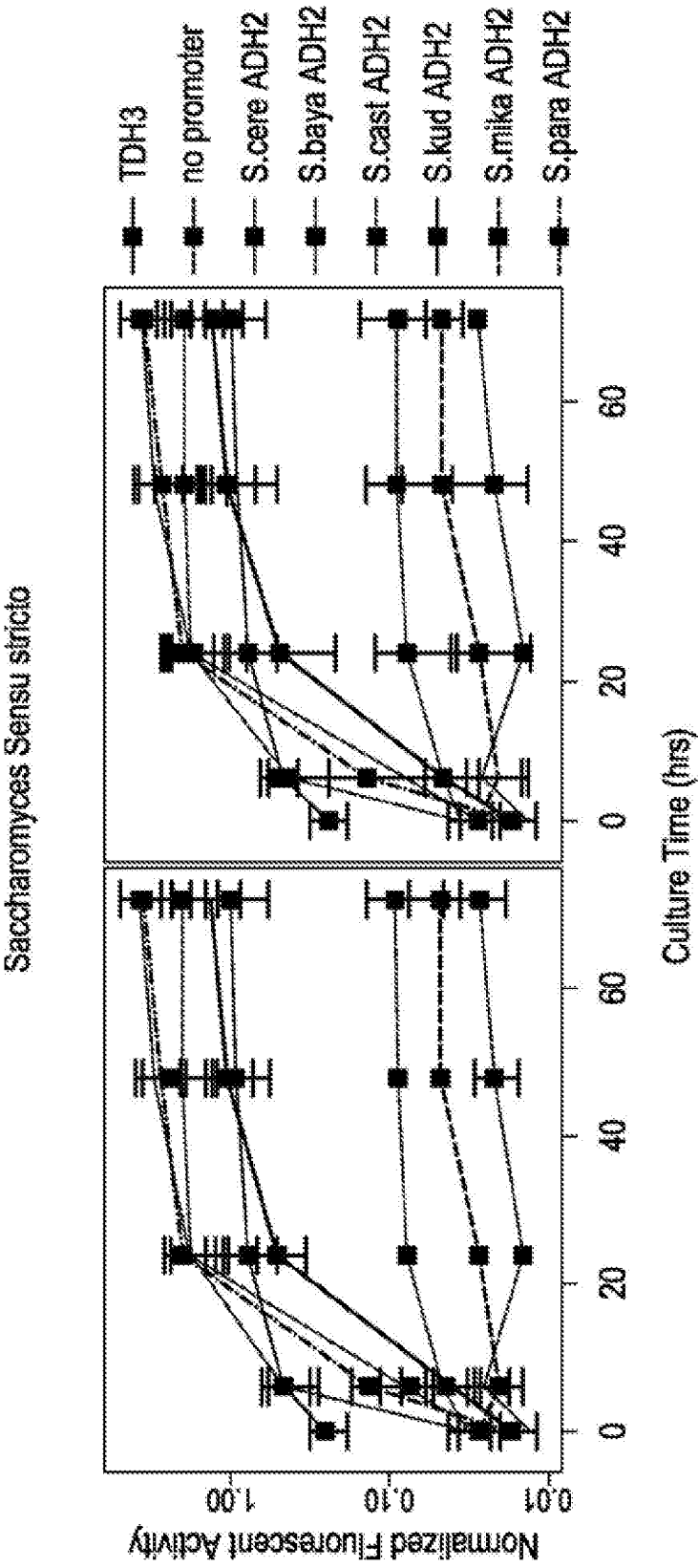


FIG. 23

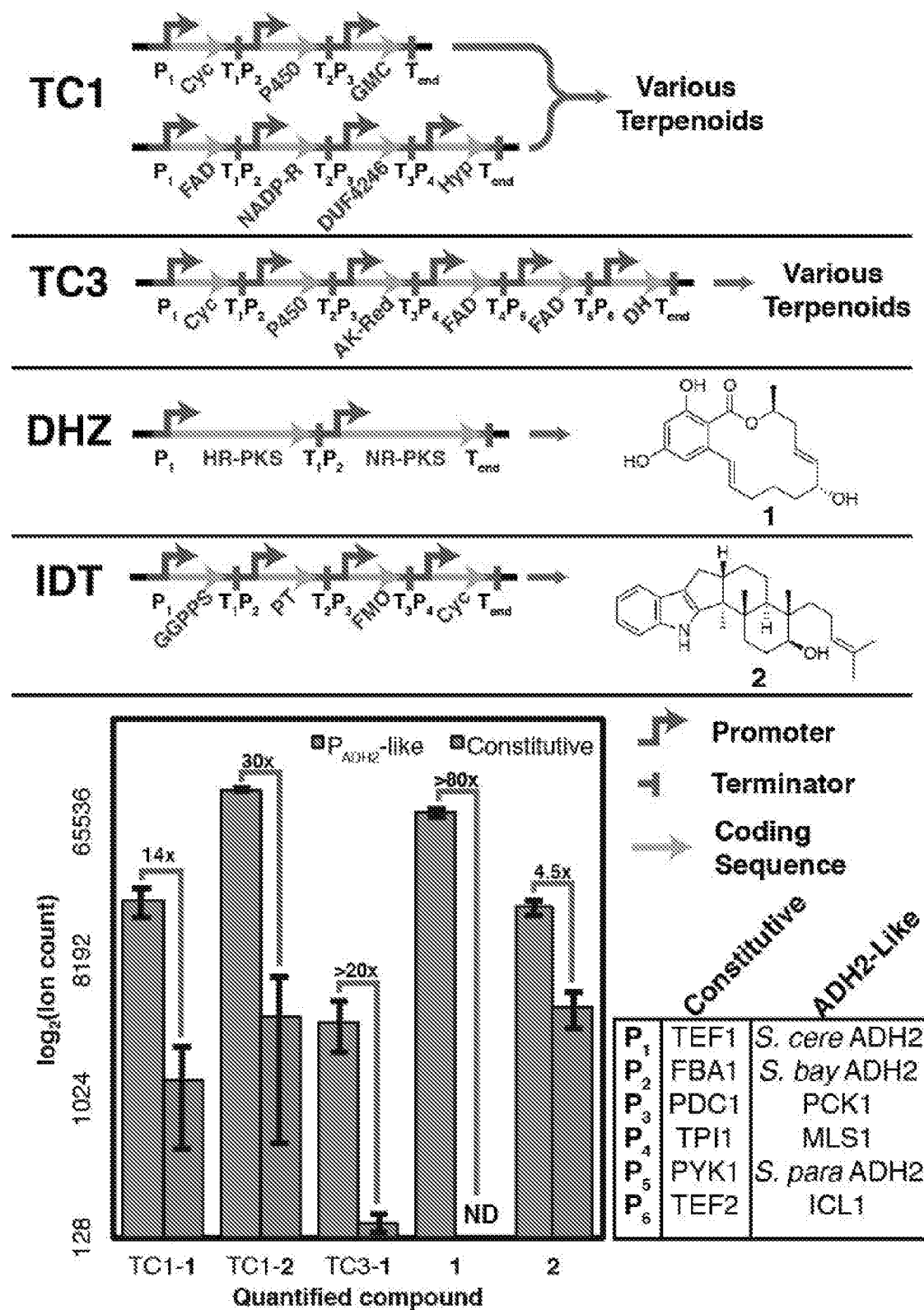


FIG. 24

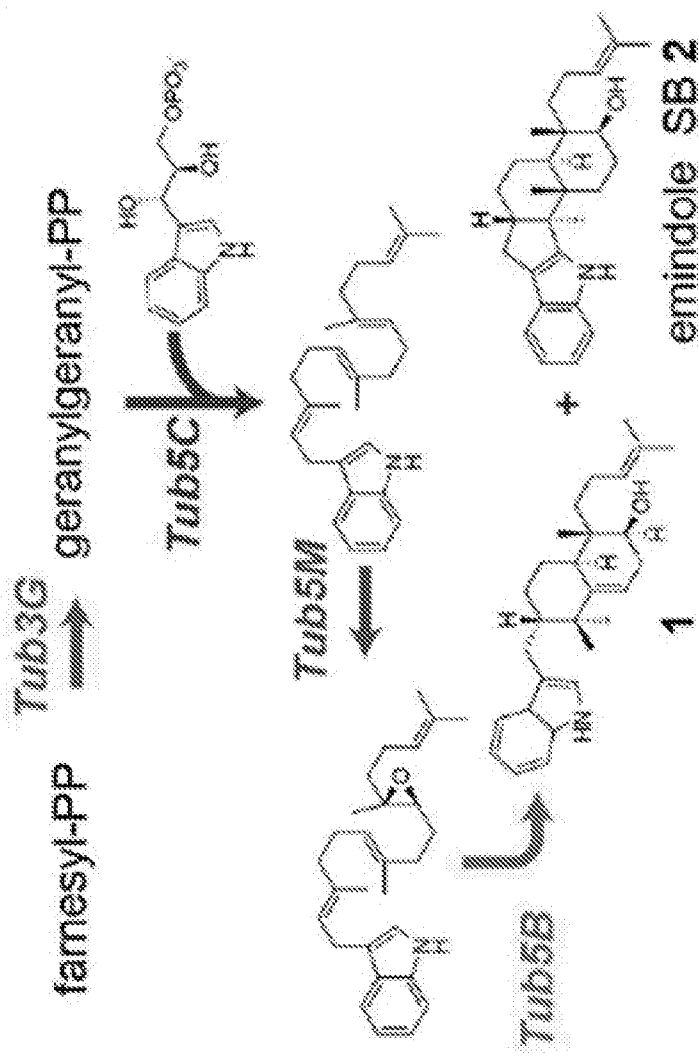


FIG. 25

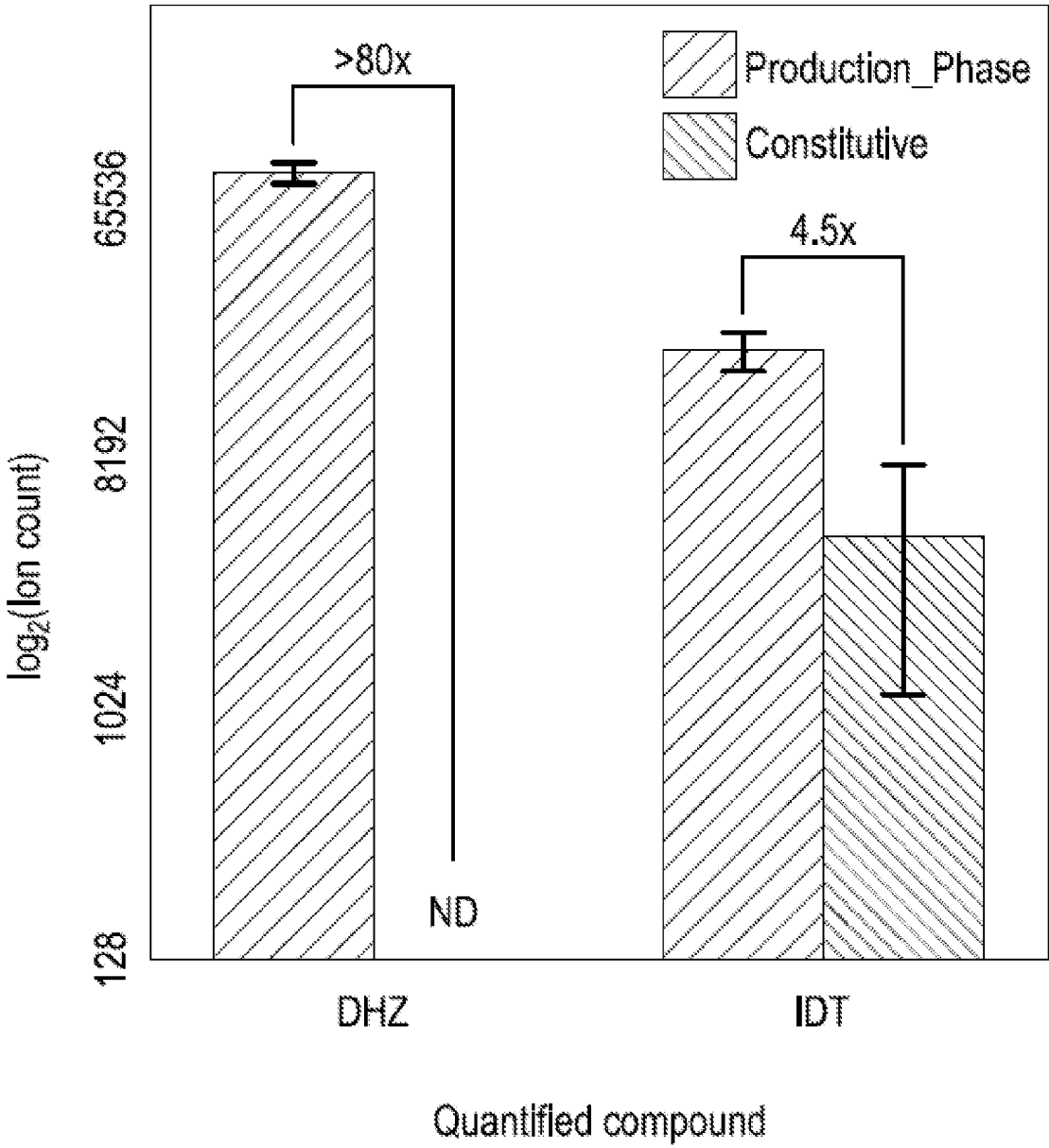


FIG. 26

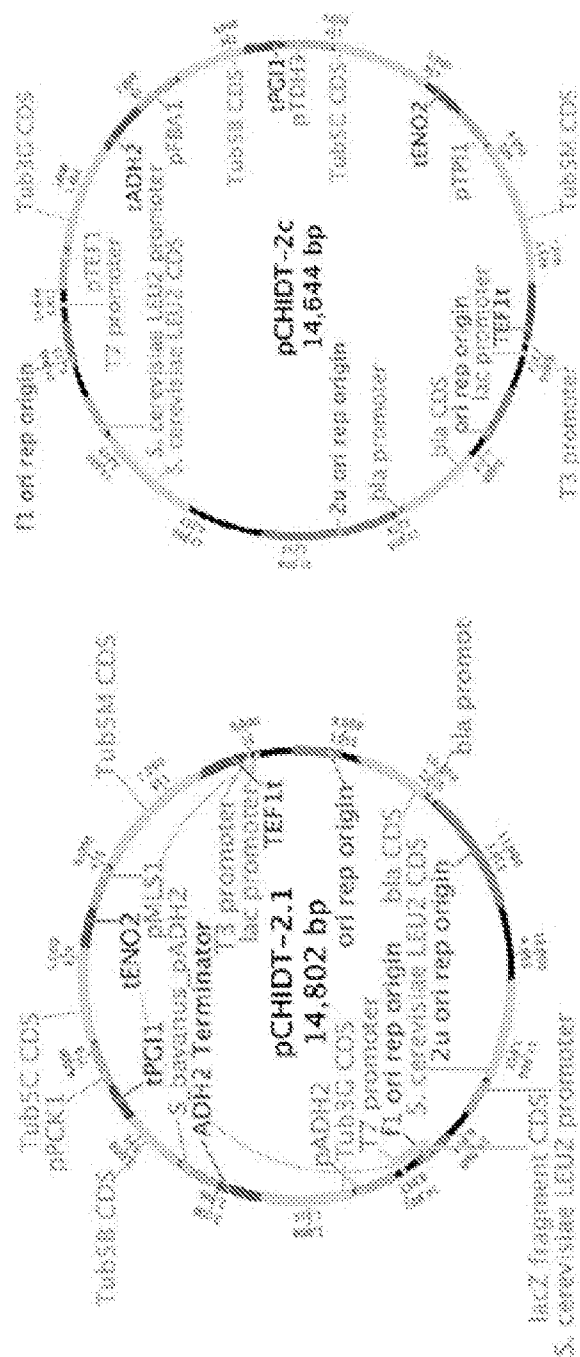


FIG. 27A

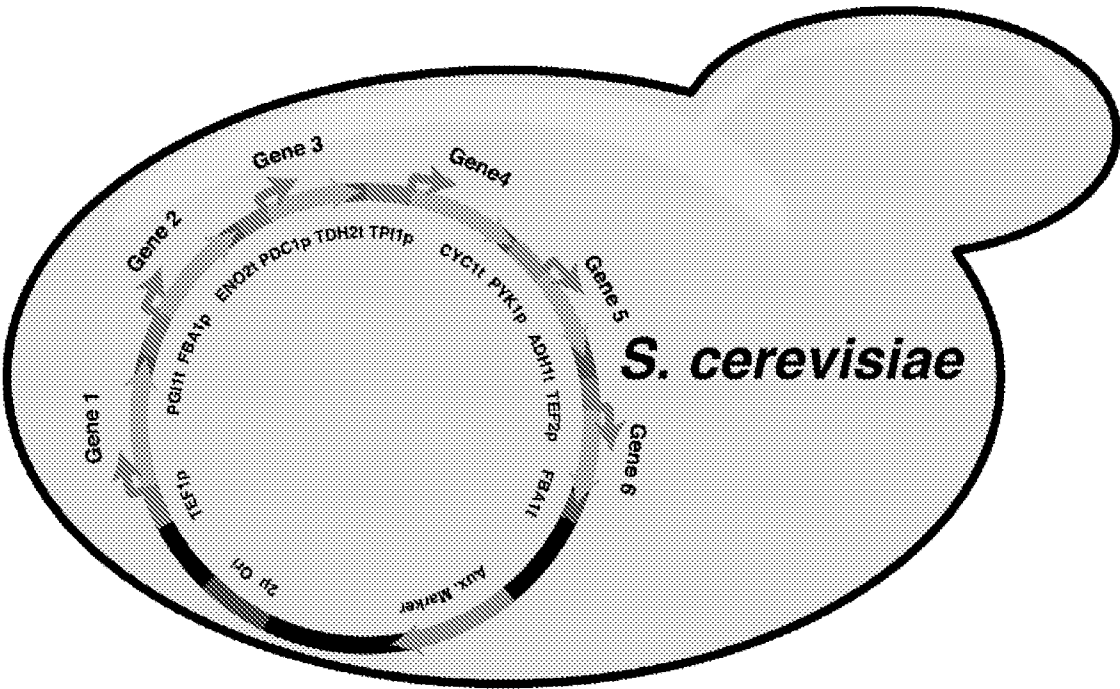


FIG. 27B

49 / 87

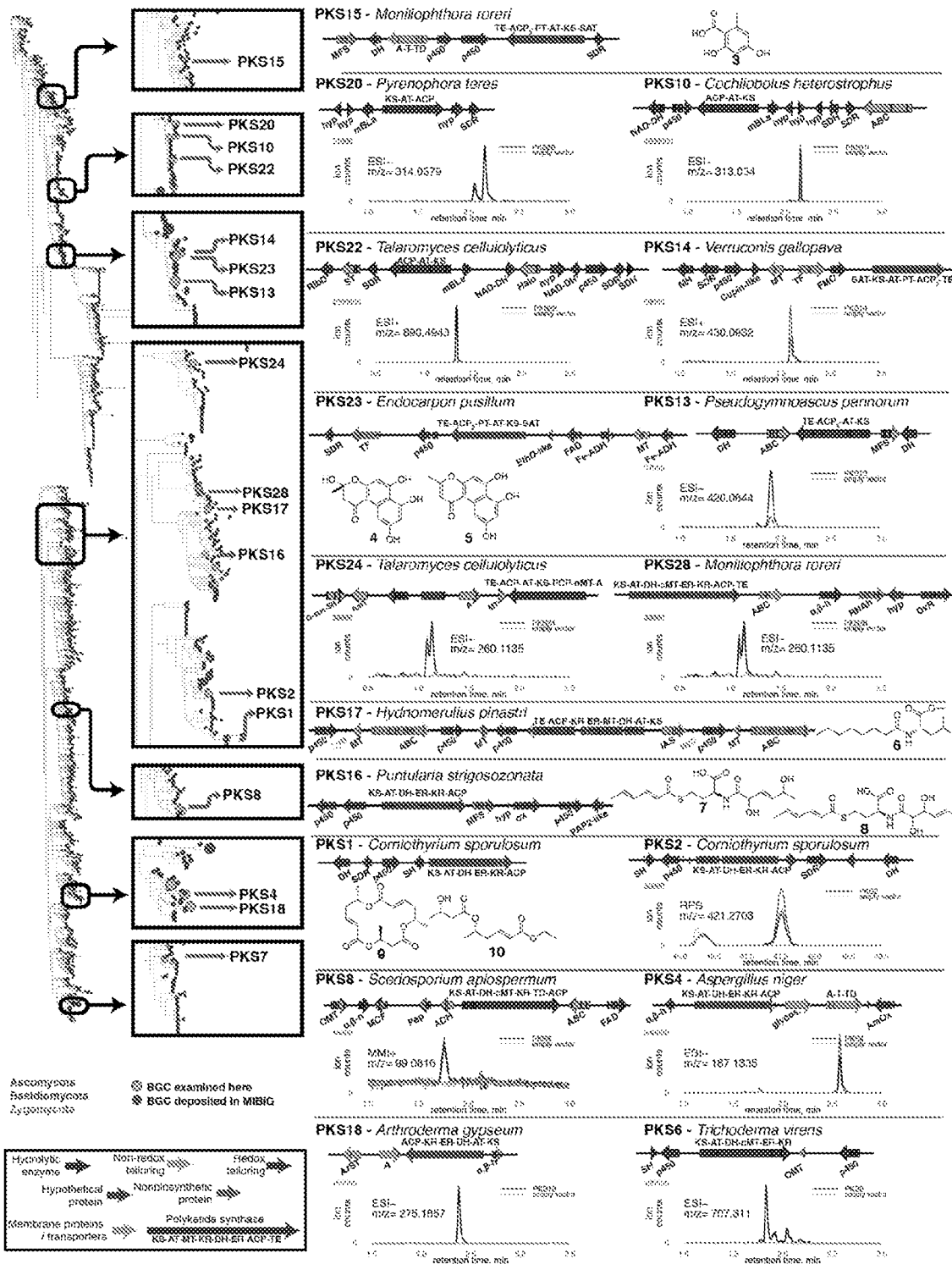


FIG. 28A

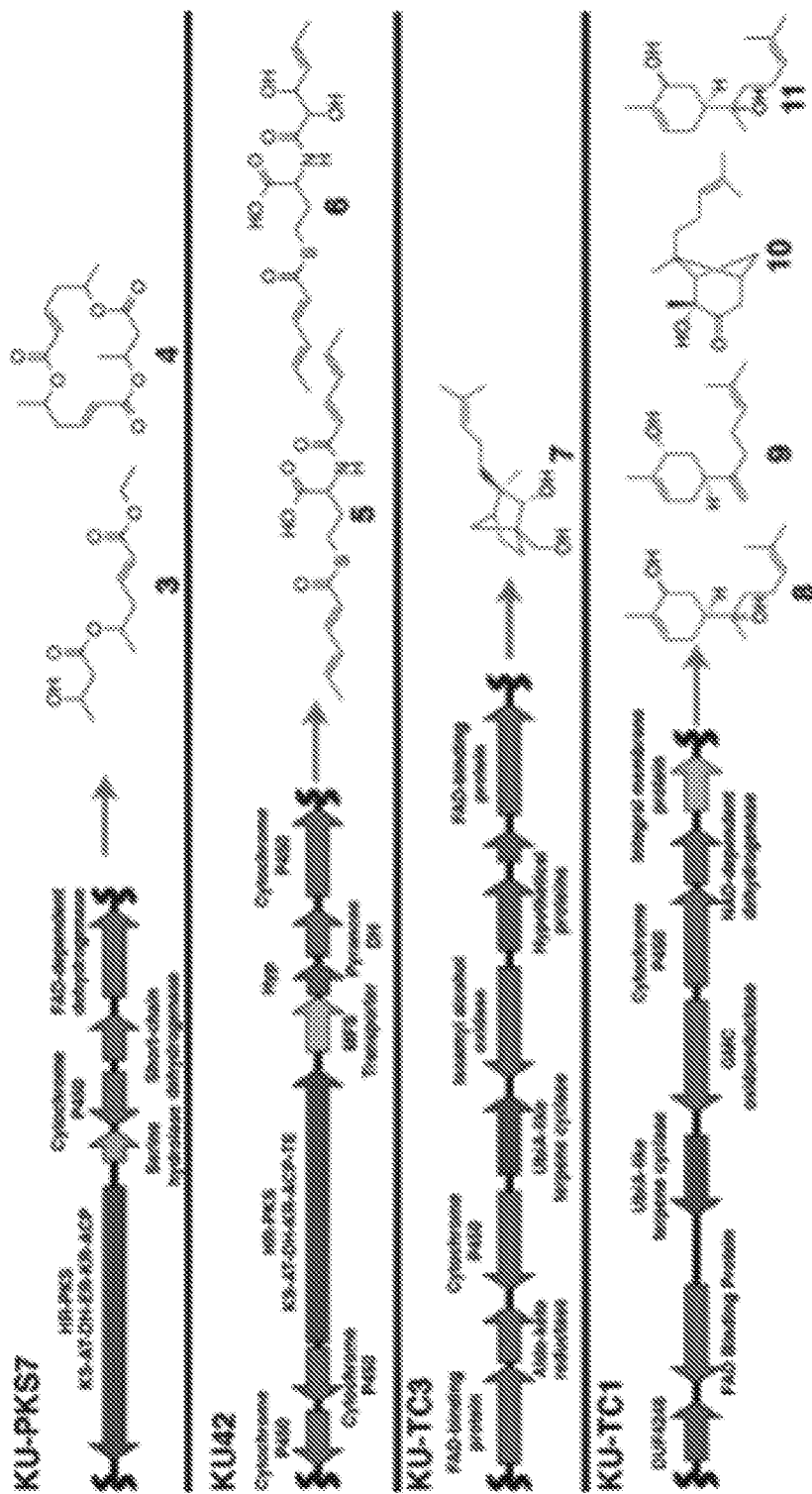


FIG. 28B

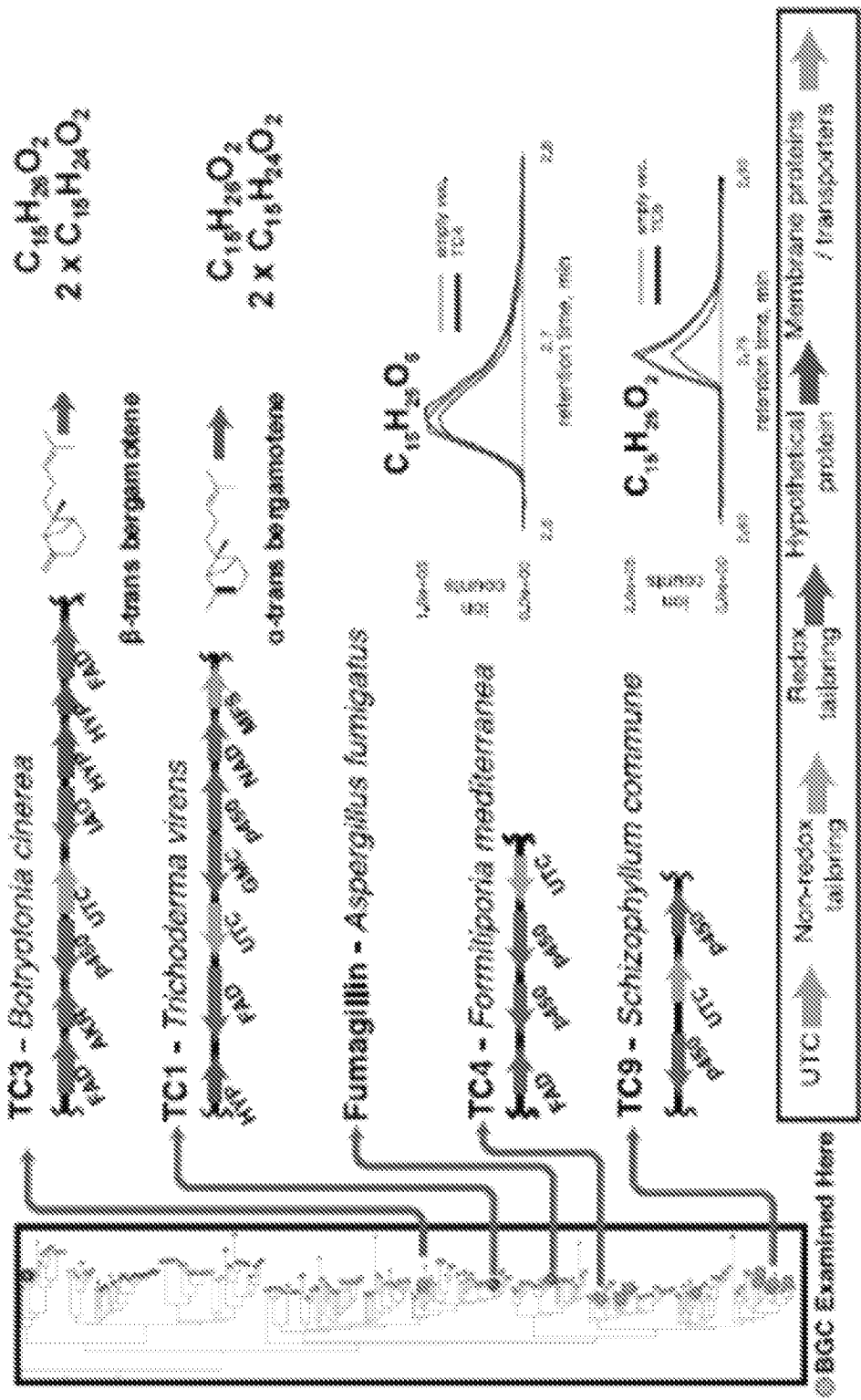


FIG. 28C

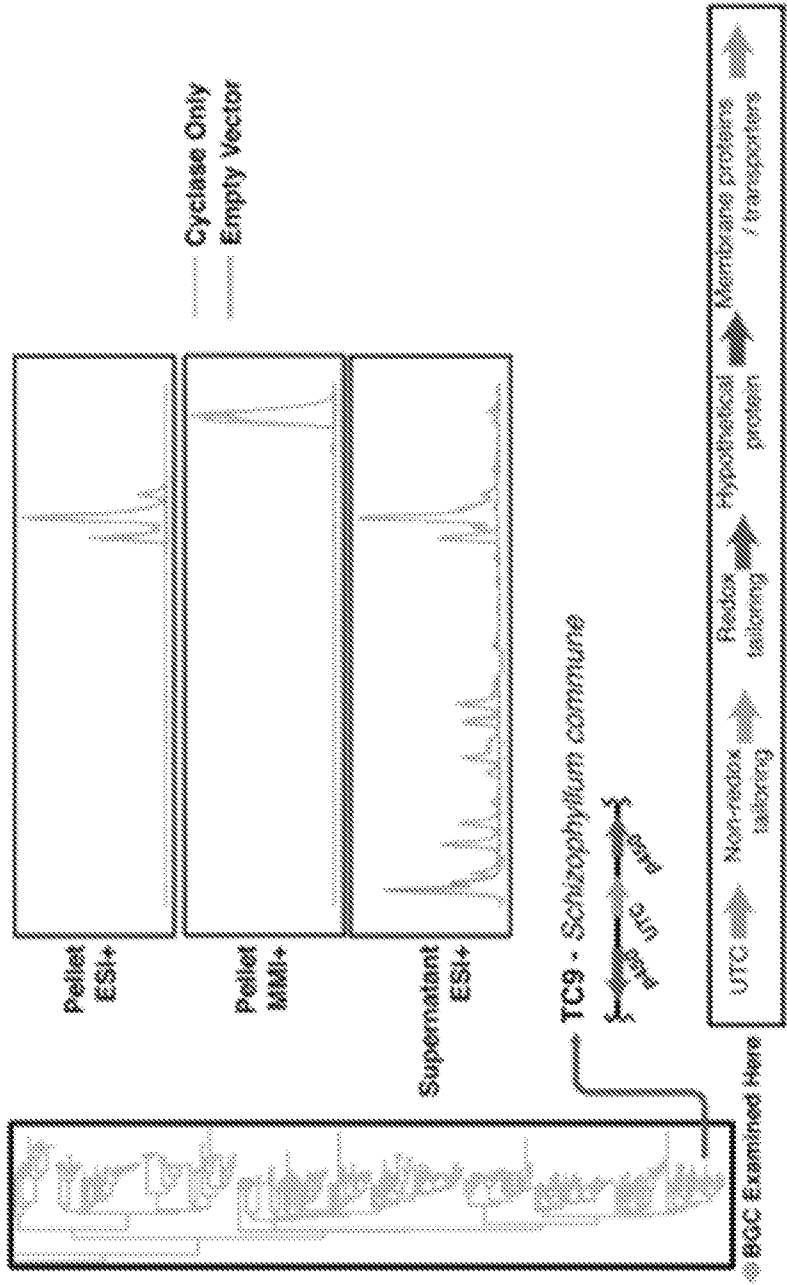


FIG. 28D

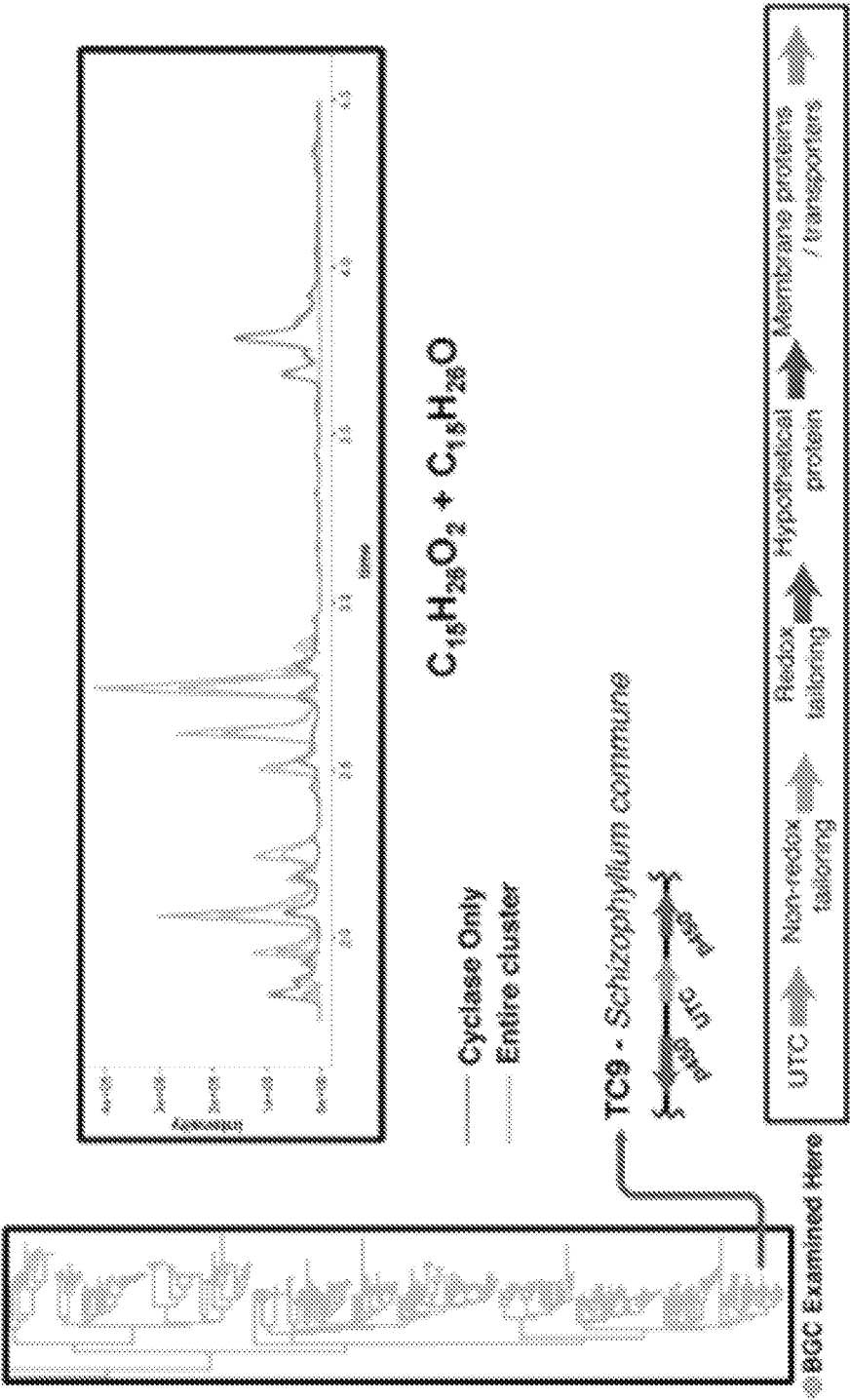


FIG. 28E

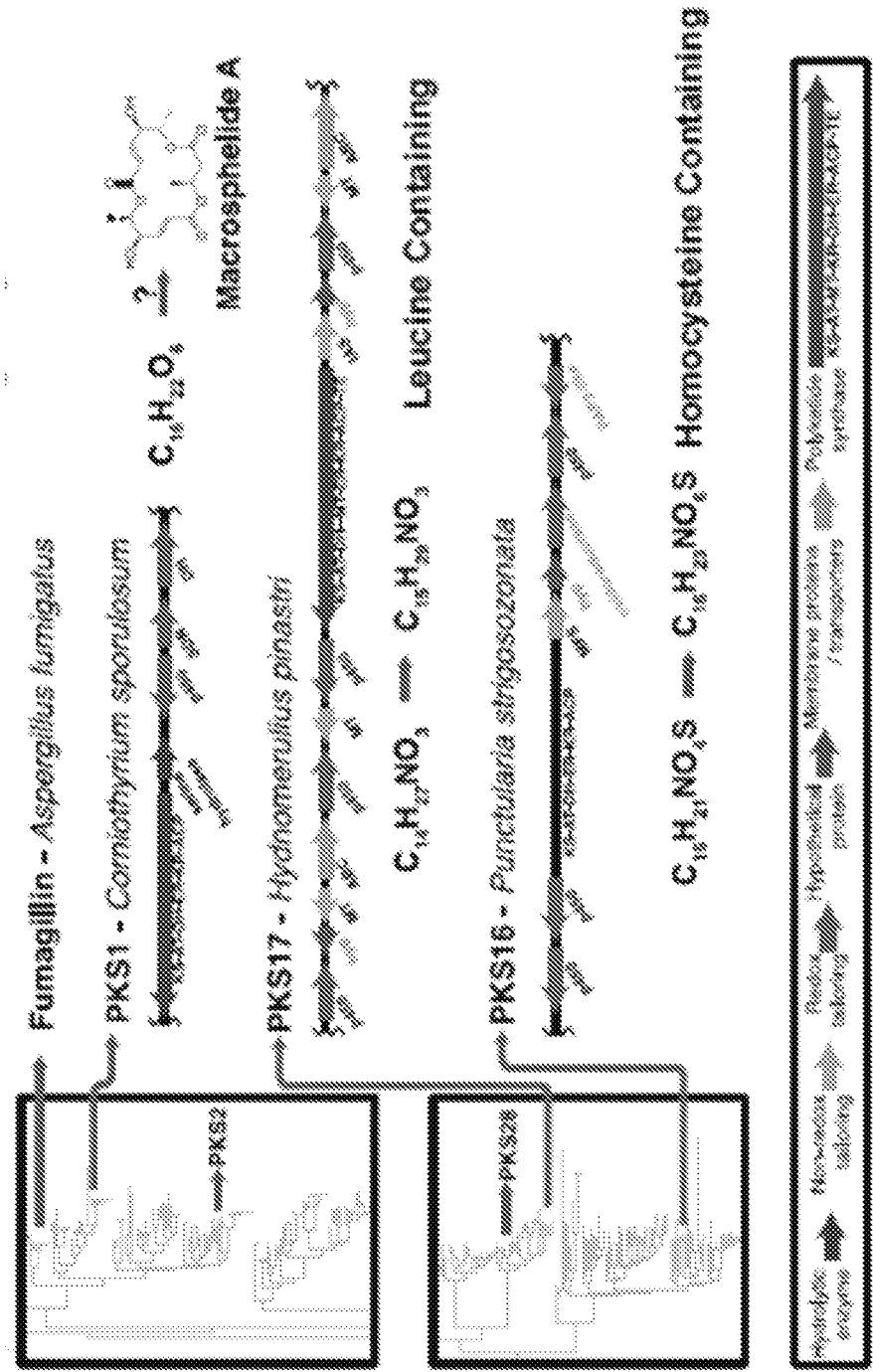


FIG. 28F

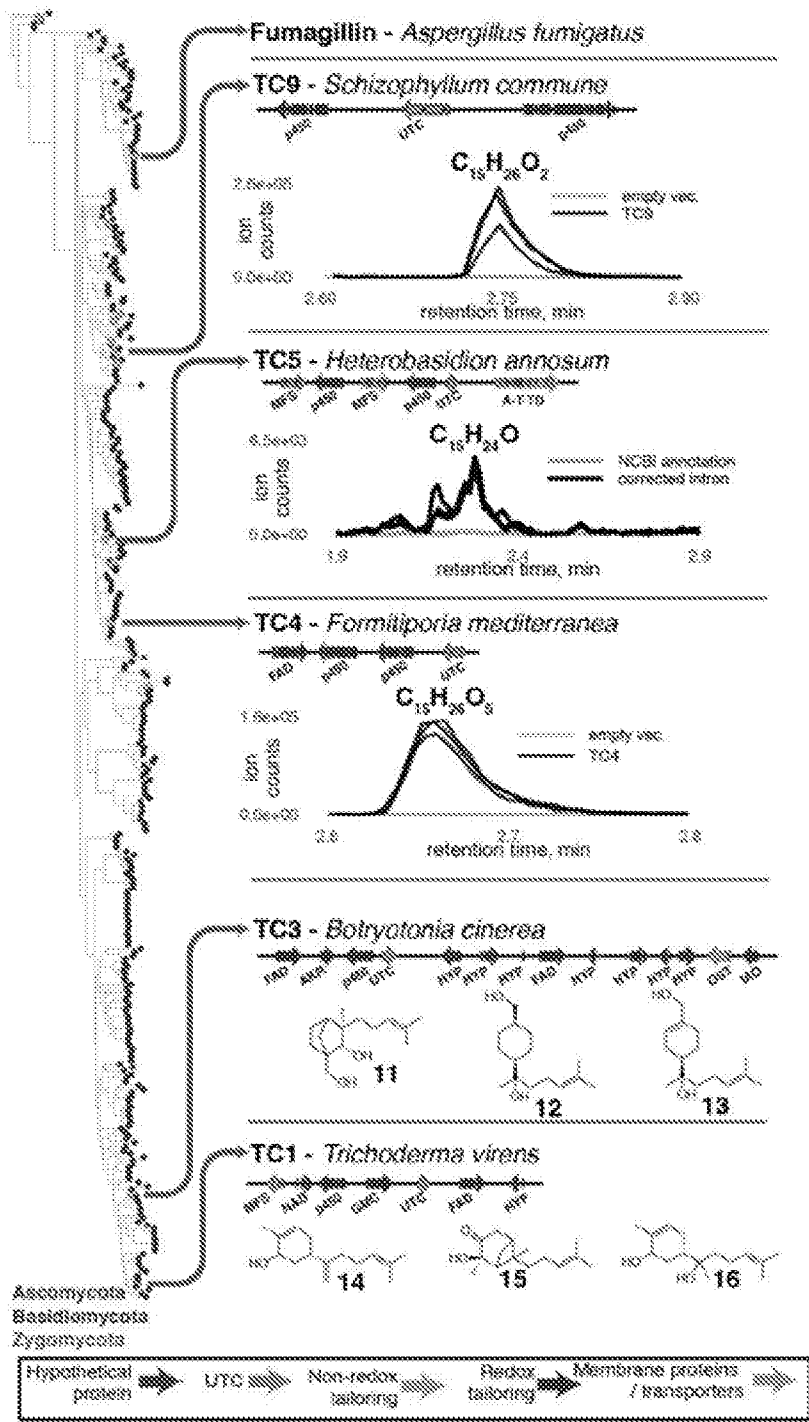


FIG. 29A

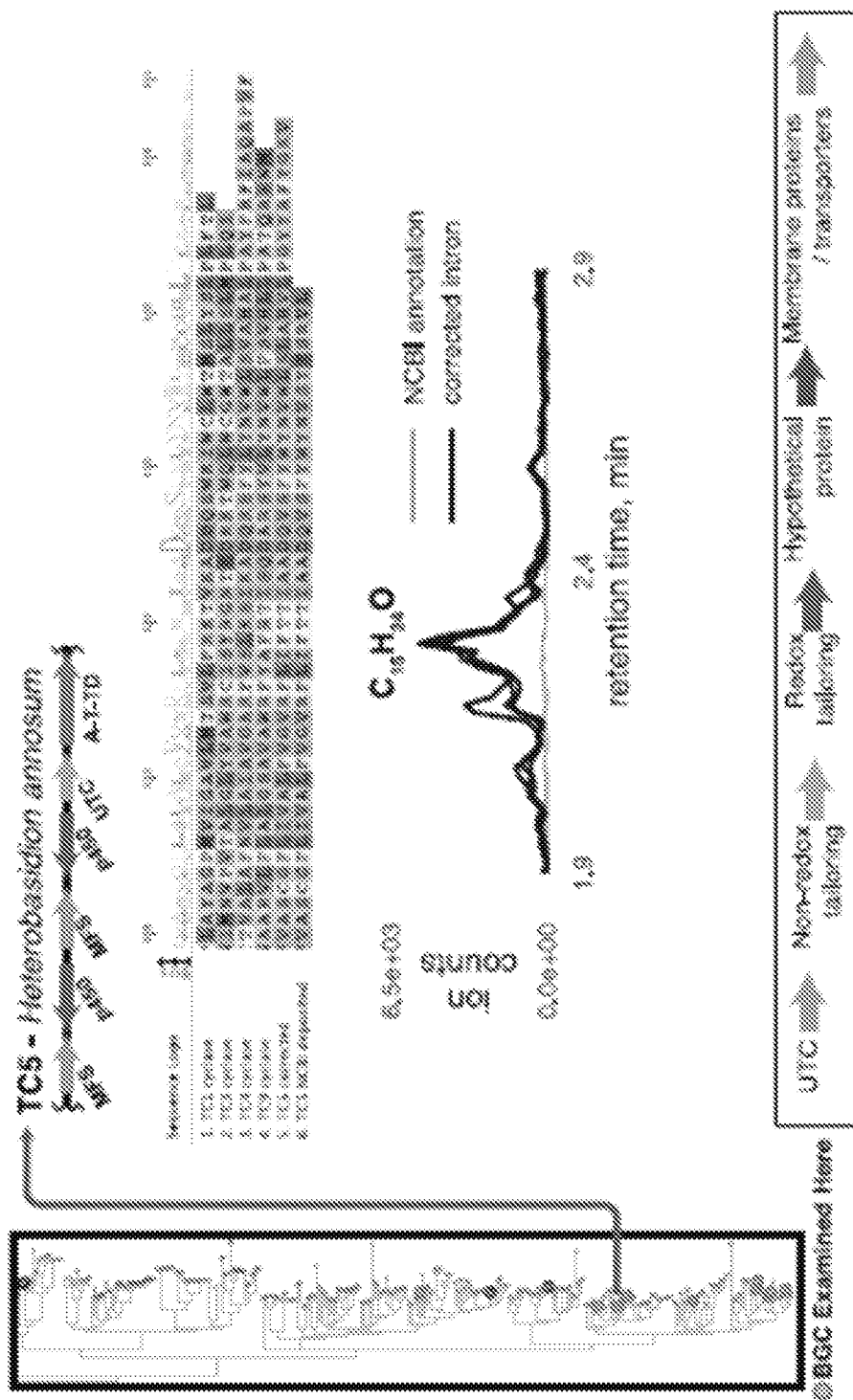


FIG. 29B

57 / 87



FIG. 30A

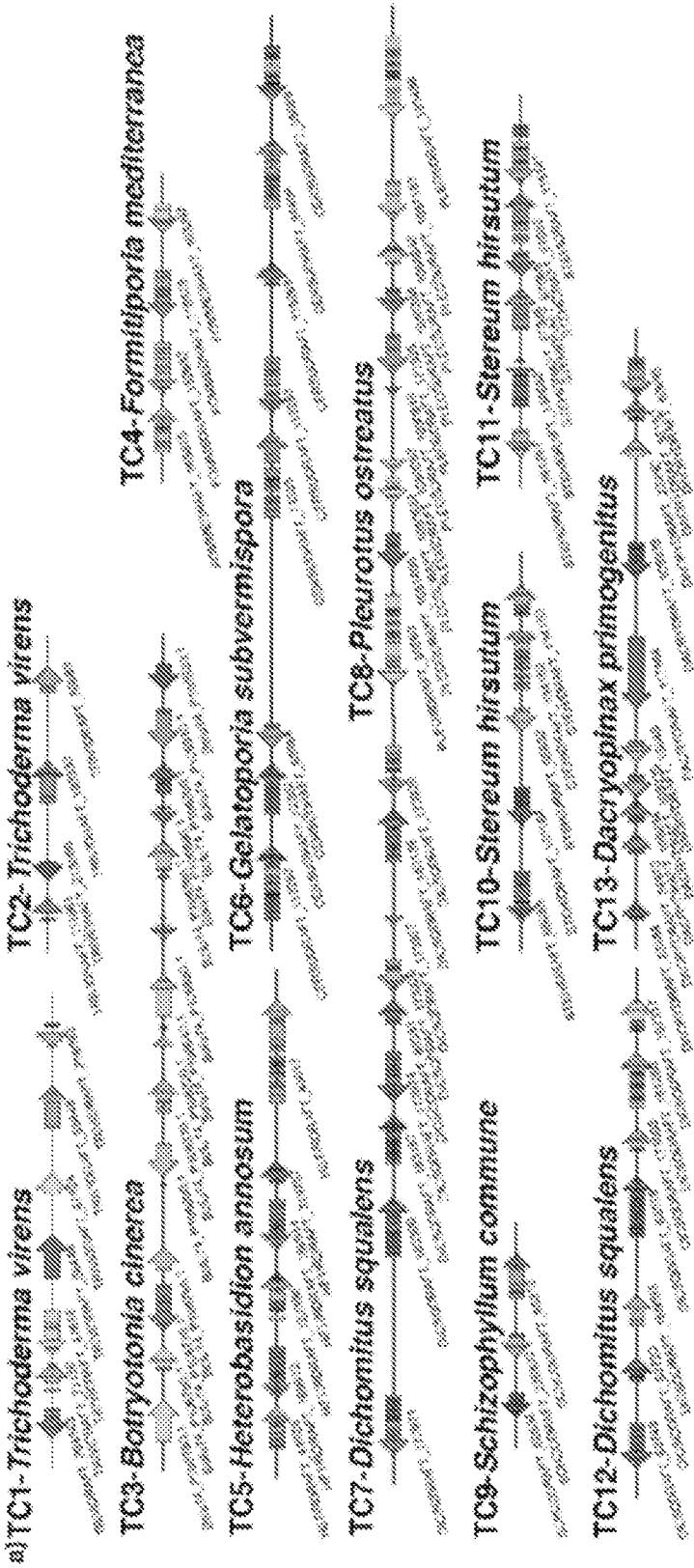


FIG. 30B

59 / 87

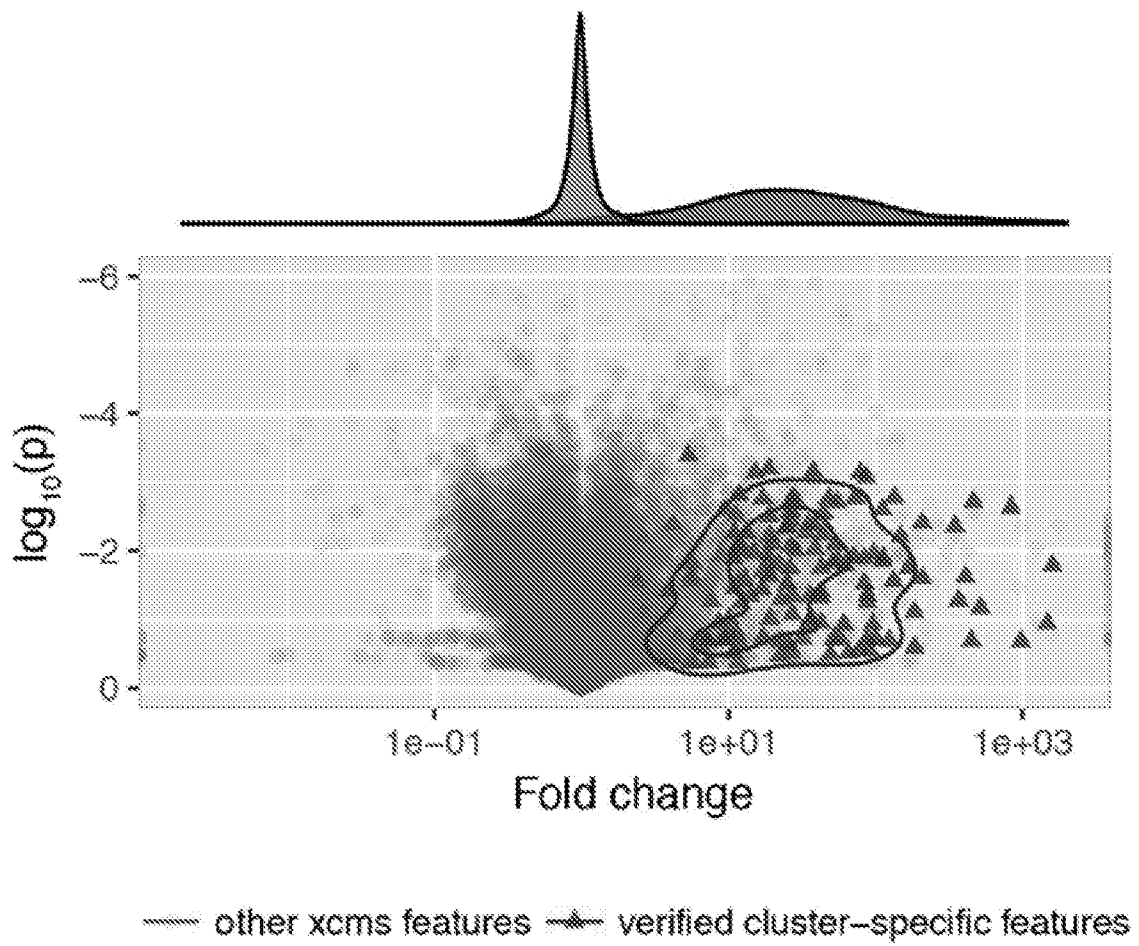


FIG. 31

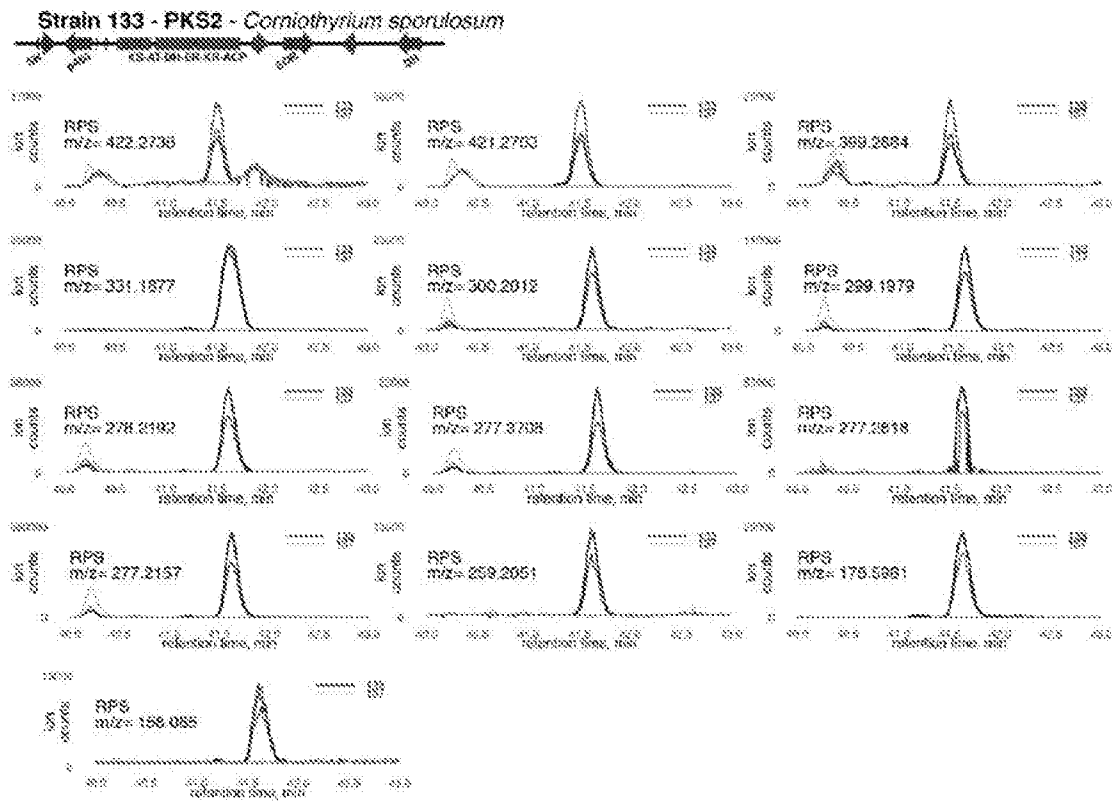


FIG. 33

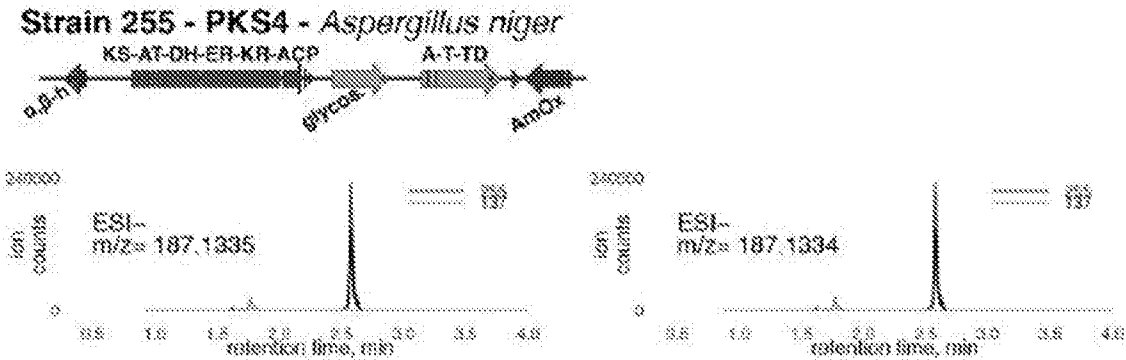


FIG. 34

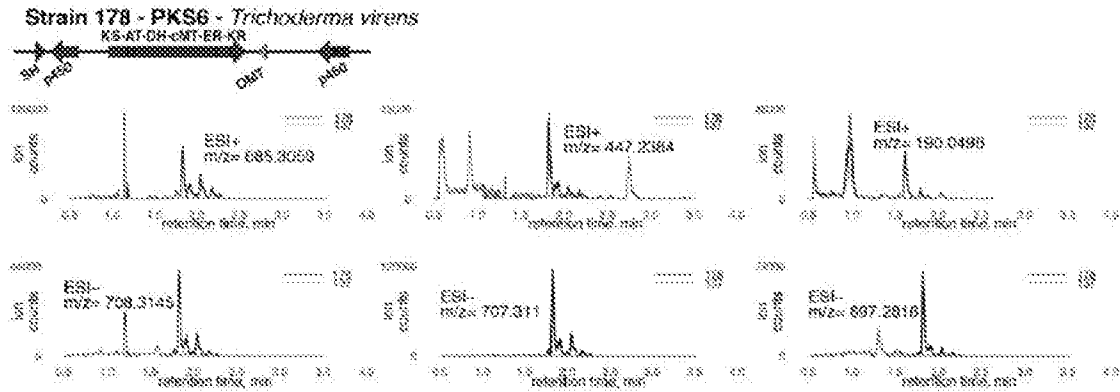


FIG. 35

Strain 164 - PKS8 - *Scedosporium apiospermum*

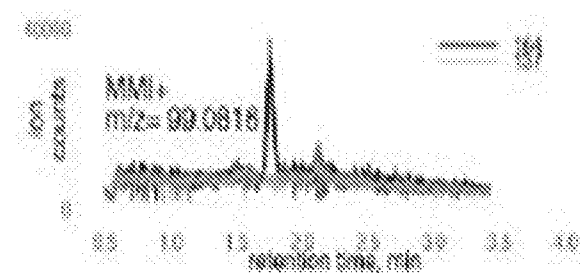
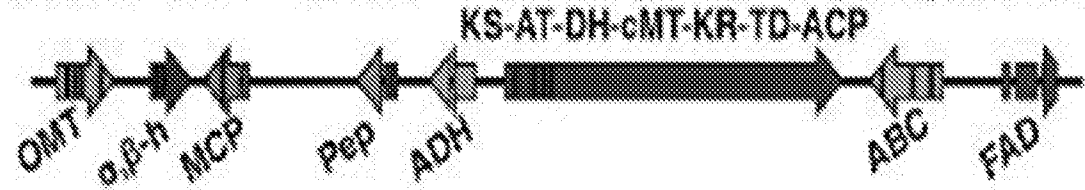


FIG. 36

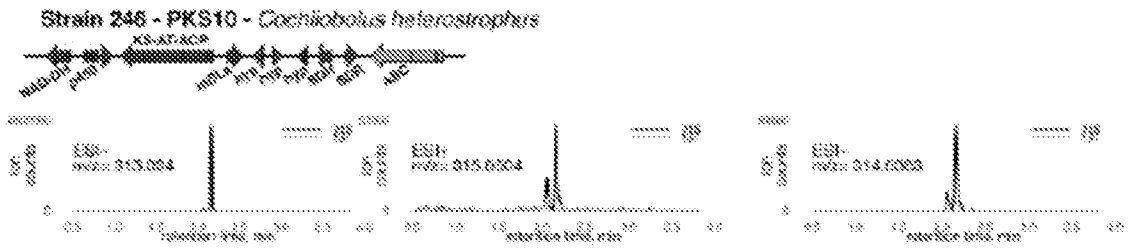


FIG. 37

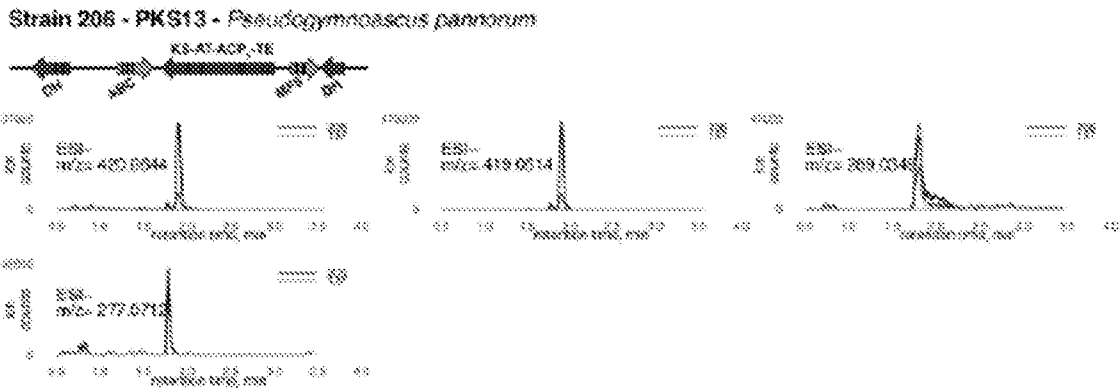


FIG. 38

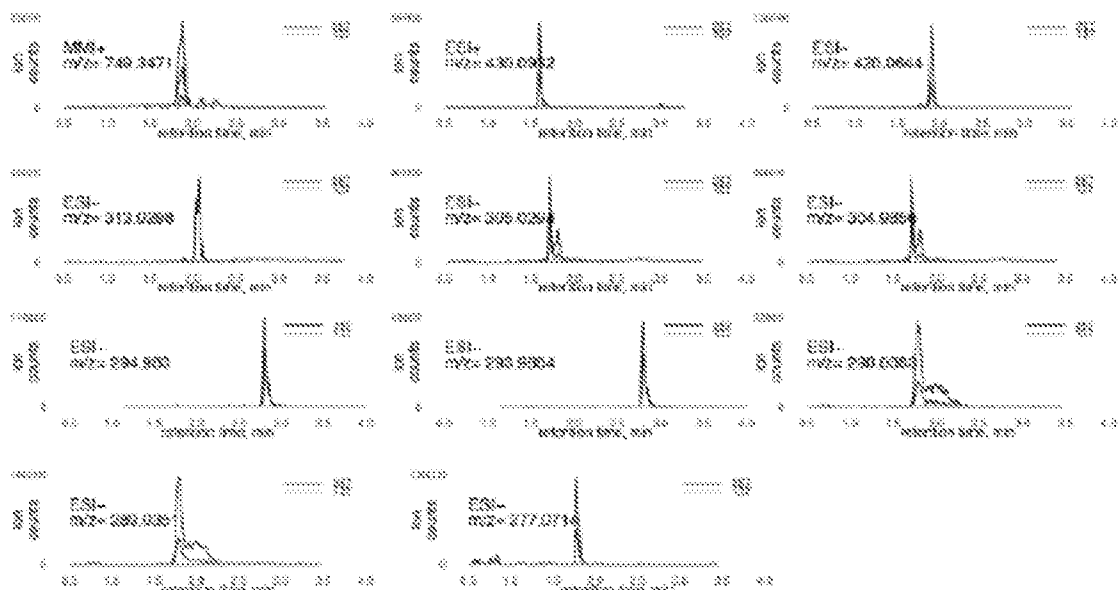
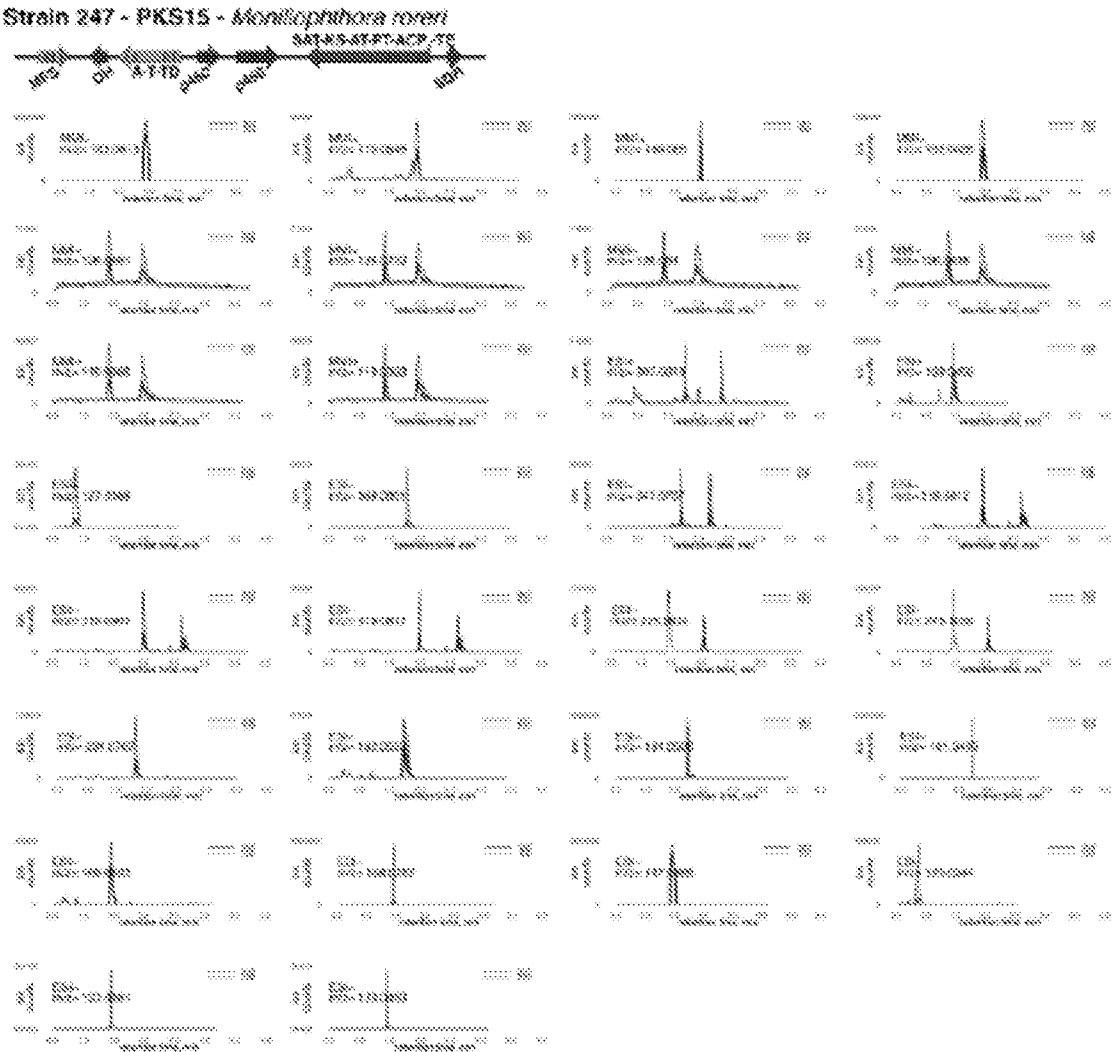


FIG. 39



Strain 177 - PKS16 - *Puntularia strigosozonata*

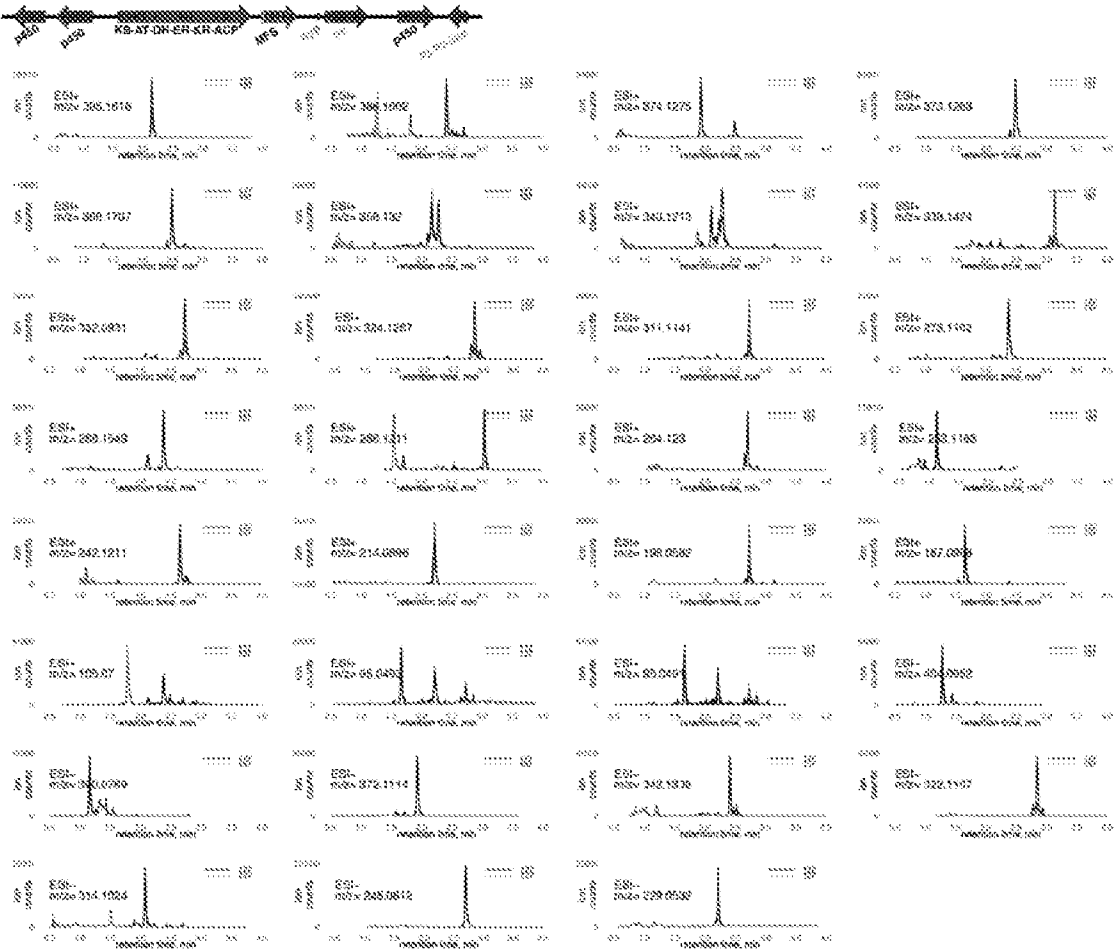


FIG. 41

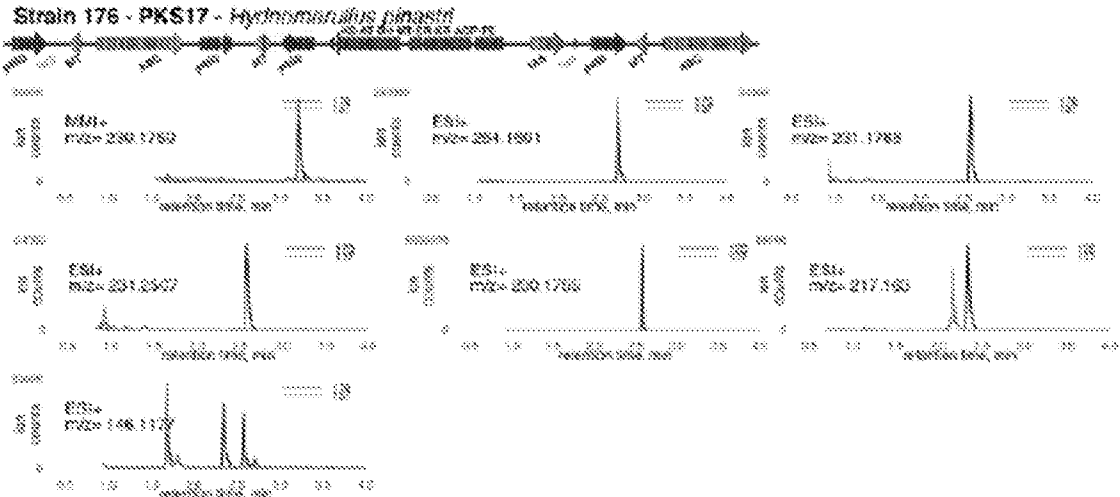


FIG. 42

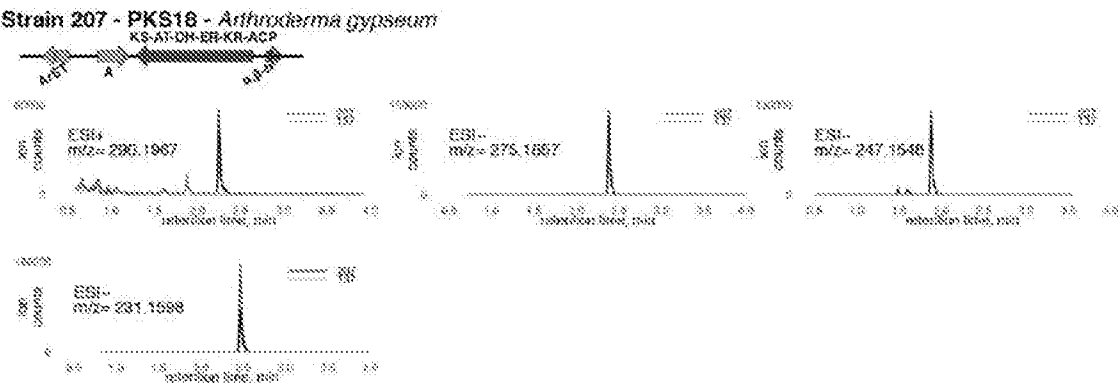


FIG. 43

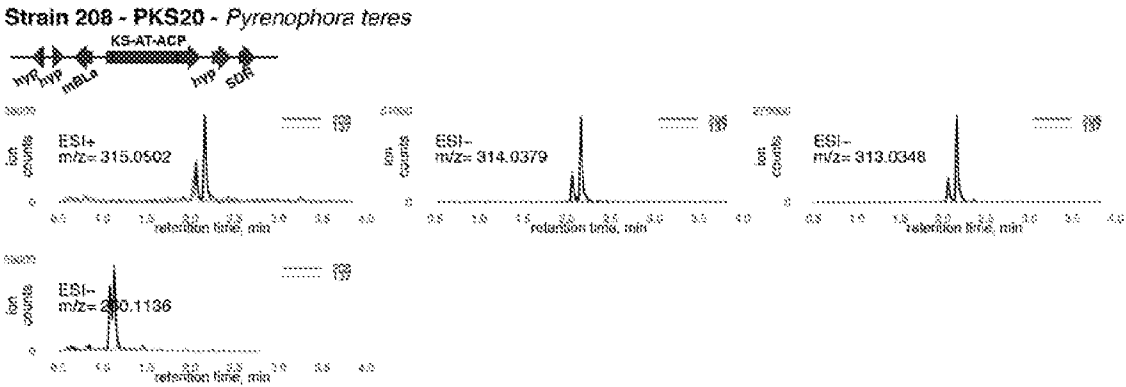


FIG. 44

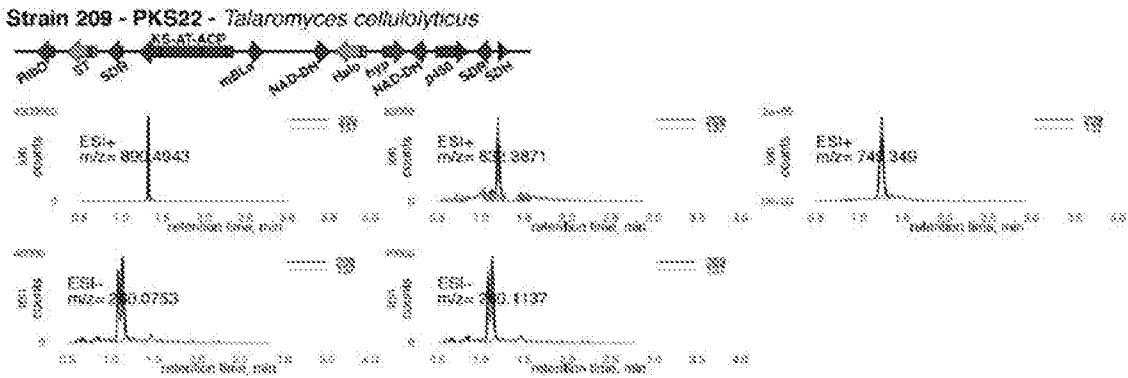


FIG. 45

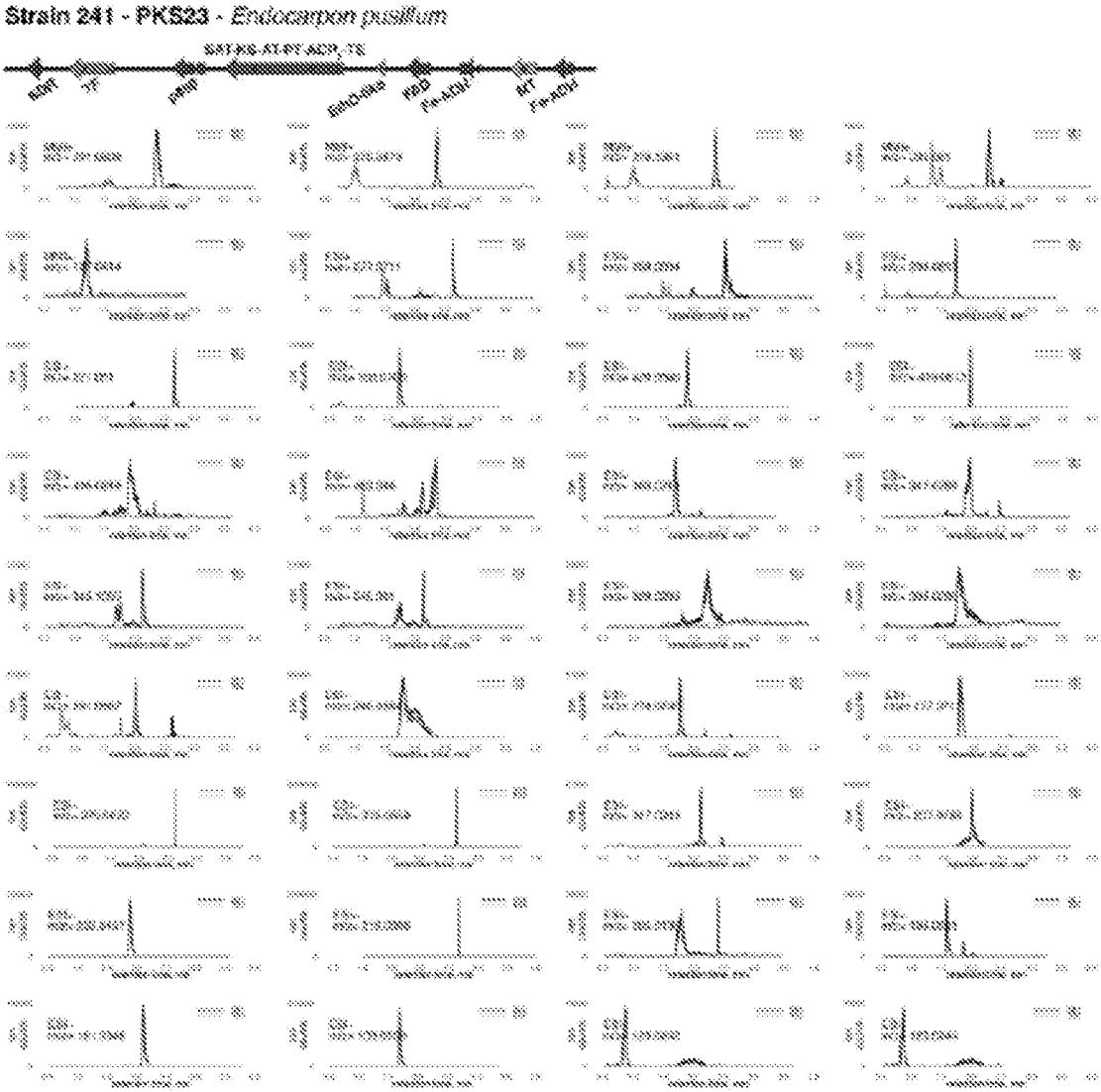


FIG. 46

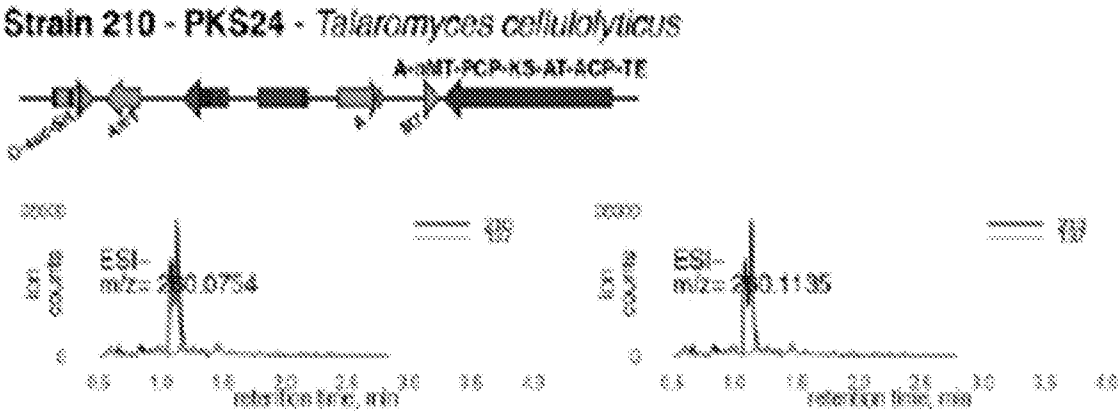


FIG. 47

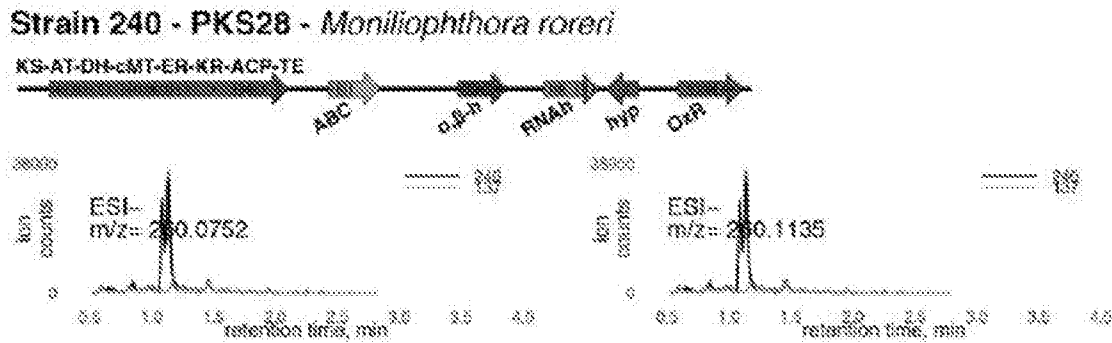


FIG. 48

77 / 87

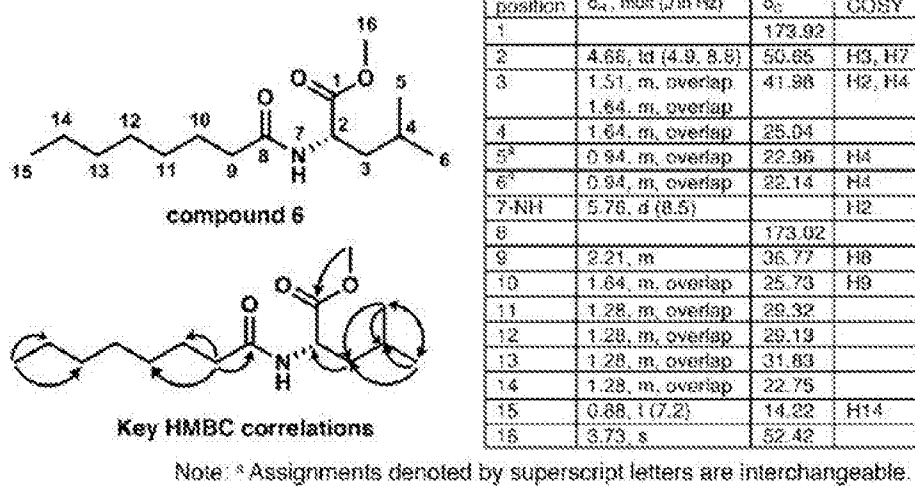
 ^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), CDCl_3 

FIG. 49

^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), Acetone- d_6

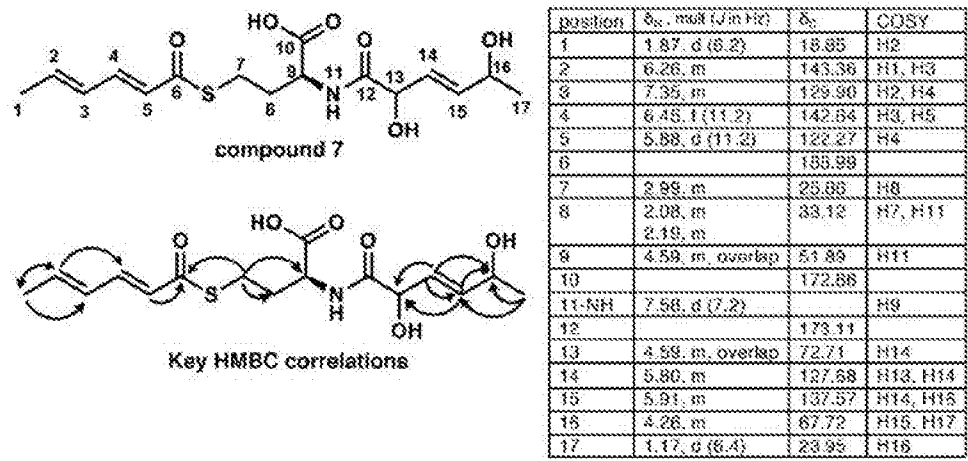


FIG. 50

¹H NMR spectrum (500 MHz), ¹³C NMR spectrum (125 MHz), Acetone-*d*₆

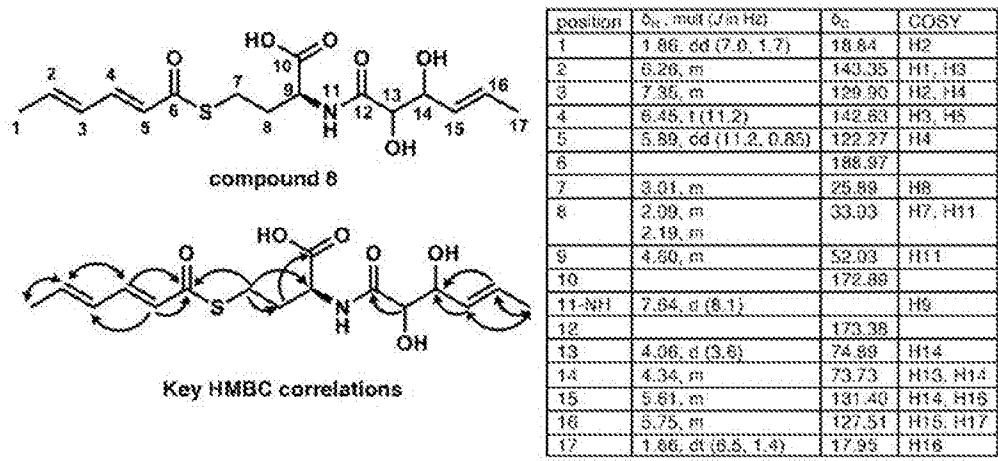
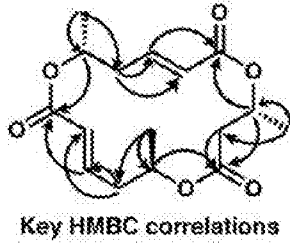
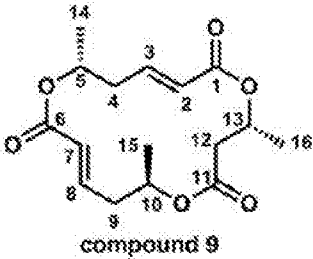


FIG. 51

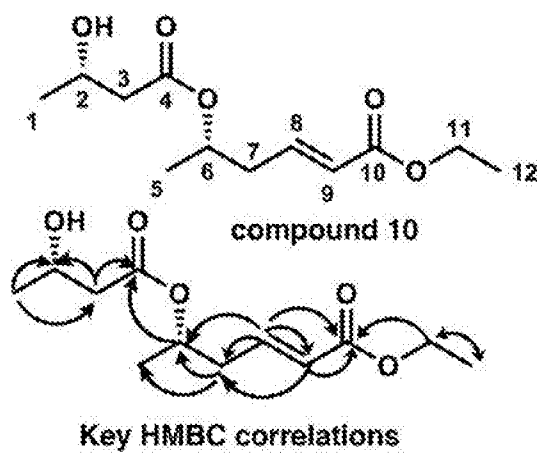
¹H NMR spectrum (500 MHz), ¹³C NMR spectrum (125 MHz), CDCl₃



position	δ_H , mult (J in Hz)	δ_C	COSY
1		165.07	
2	5.80, m	125.02	H3
3	5.84, m, overlap	143.61	H2, H4
4	2.31, m	38.99	H3, H5
	2.49, m		
5	5.05, m	68.72	H4, H14
6		165.46	
7	5.77, m	123.84	H8
8	5.80, m, overlap	144.69	H7, H9
9	2.39, m	59.50	H8, H10
10	5.11, m	70.21	H9, H15
11		170.30	
12	2.47, dd (15.3, 2.3)	41.03	H13
	2.57, dd (15.3, 10.6)		
13	5.35, m	67.61	H12, H16
14	1.34, d (6.3)	21.16	H5
15	1.27, d (6.3)	20.68	H10
16	1.30, d (6.3)	20.21	H13

FIG. 52

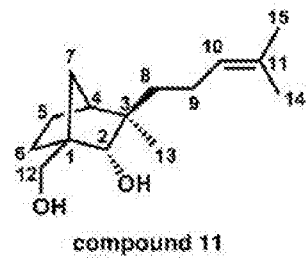
81 / 87

 ^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), CDCl_3 

position	δ_{H} , mult (J in Hz)	δ_{C}	COSY
1	1.22, d (6.3)	22.56	H2
2	4.19, m, overlap	64.41	H1, H3
3	2.45, m, overlap	43.70	H2
4		172.48	
5	1.28, m, overlap	19.81	H6
6	5.07, m	69.71	H5, H7
7	2.45, m, overlap	38.48	H6, H8
8	6.86, dt (15.5, 7.4)	143.31	H7, H9
9	5.87, dt (15.5, 1.4)	124.51	H8
10		166.27	
11	4.19, m, overlap	60.55	H12
12	1.28, m, overlap	14.38	H11

FIG. 53

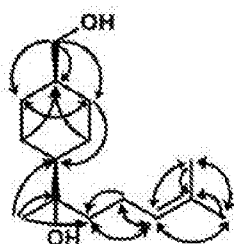
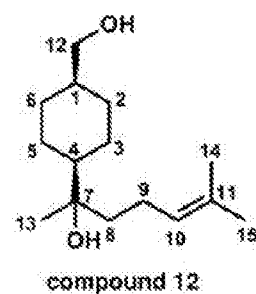
¹H NMR spectrum (500 MHz), ¹³C NMR spectrum (125 MHz), CDCl₃



position	δ_c , mult (J in Hz)	δ_c
1		53.82
2	3.63, brs	82.58
3		42.57
4	1.88, m	46.26
5	1.46, m	25.83
	1.74, m	
6	1.20, m	22.16
	1.88, m	
7	1.09, d (10.1)	36.98
	1.45, m	
8	1.31, m	42.61
9	1.96, m	22.95
10	5.09, m	124.95
11		131.47
12	3.75, d (10.5)	68.54
	3.84, d (10.5)	
13	0.87, s	16.47
14	1.61, s	17.79
15	1.68, s	25.65

FIG. 54

83 / 87

 ^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), CDCl_3 

Key HMBC correlations

position	δ_{H} , mult (J in Hz)	δ_{C}	COSY
1	1.43, m	40.83	H2, H6, H12
2 ^a	0.95, m, overlap 1.87, m, overlap	29.71	H1, H3
3 ^b	1.08, m, overlap 1.81, m	26.85	H2, H4
4	1.32, m	47.52	H3, H5
5 ^c	1.12, m, overlap 1.89, m, overlap	26.22	H4, H6
6 ^d	0.95, m, overlap 1.87, m, overlap	29.76	H1, H5
7		74.51	
8	1.49, dd (9.9, 6.8)	33.90	H9
9	2.05, m	22.27	H8, H10
10	5.15, t (7.3)	124.70	H9
11		131.92	
12	3.45, d (8.3)	66.78	
13	1.11, s	24.03	
14	1.62, s	17.83	
15	1.69, s	25.88	

Note: ^a Assignments denoted by superscript letters are interchangeable.

FIG. 55

84 / 87

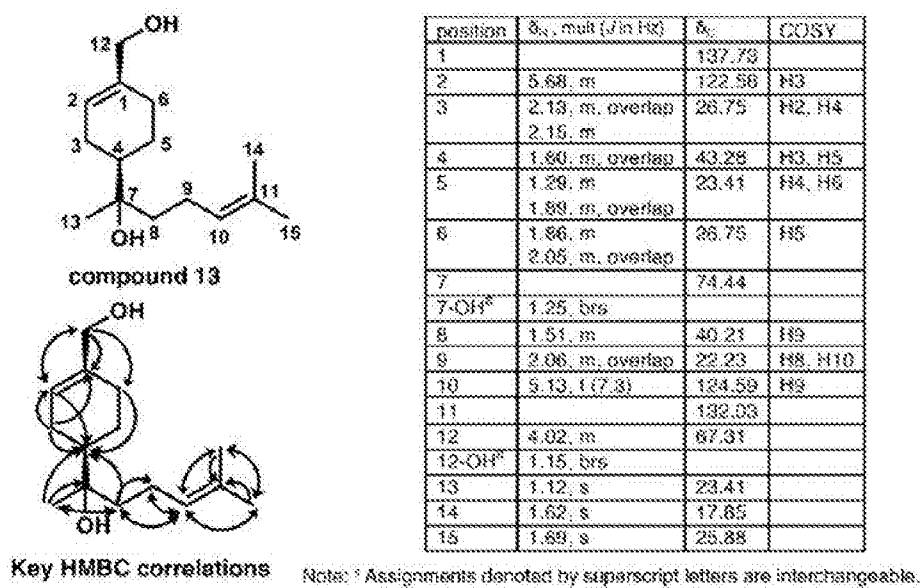
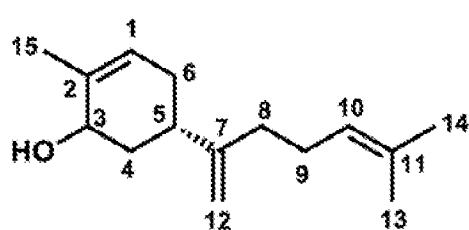
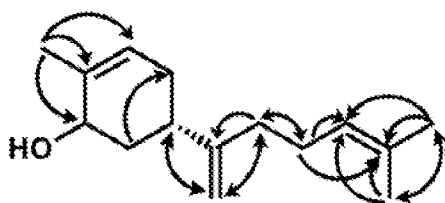
 ^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), CDCl_3 

FIG. 56

85 / 87

 ^1H NMR spectrum (500 MHz), ^{13}C NMR spectrum (125 MHz), CDCl_3 

compound 14



Key HMBC correlations

position	δ_{H} , mult (J in Hz)	δ_{C}	COSY
1	5.59	125.69	H6
2		134.41	
3	4.02, brs	68.81	H4
4	1.59, m, overlap 1.95, m	37.34	H3, H5
5	2.33, m	34.14	H4, H6
6	1.83, m 2.16, m, overlap	31.74	H1, H5
7		153.42	
8	2.08, m	35.04	H9
9	2.13, m, overlap	26.90	H8, H10
10	5.12	124.30	H9
11		131.79	
12	4.78, s	107.88	
13	1.61, s	17.90	
14	1.66, s	25.85	
15	1.80, s	20.99	

FIG. 57

¹H NMR spectrum (500 MHz), ¹³C NMR spectrum (125 MHz), CDCl₃

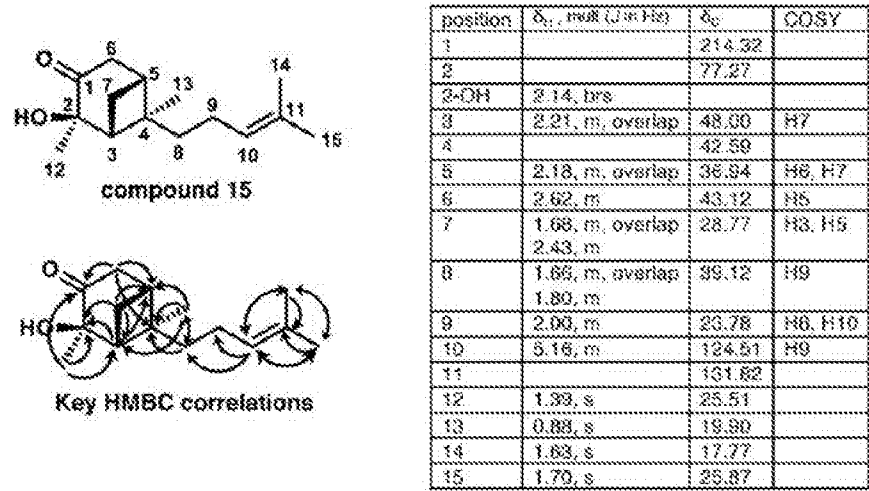


FIG. 58

¹H NMR spectrum (500 MHz), ¹³C NMR spectrum (125 MHz), CDCl₃

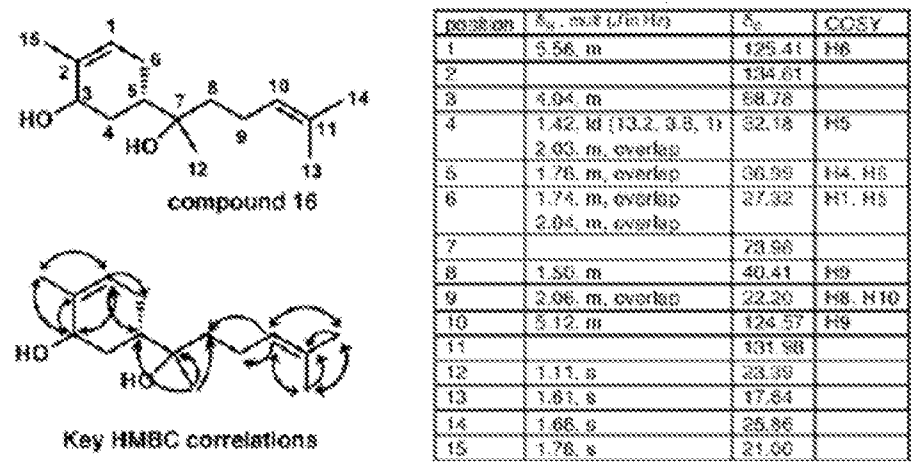


FIG. 59