



(51) International Patent Classification:

C12Q 1/68 (2006.01) *C07H 21/04* (2006.01)
C12N 1/20 (2006.01)

(21) International Application Number:

PCT/US2015/037455

(22) International Filing Date:

24 June 2015 (24.06.2015)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/016,292 24 June 2014 (24.06.2014) US

(71) Applicant: THE SCRIPPS RESEARCH INSTITUTE
[US/US]; 10550 North Torrey Pines Road, La Jolla, CA
92037 (US).(72) Inventor: SHEN, Ben; 213 Porgee Rock Place, Jupiter, FL
33458 (US).(74) Agent: ZACHARIADES, Nicholas, A.; McHale & Slavin,
P.A., 2855 PGA Boulevard, Palm Beach Gardens, FL
33410 (US).(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,

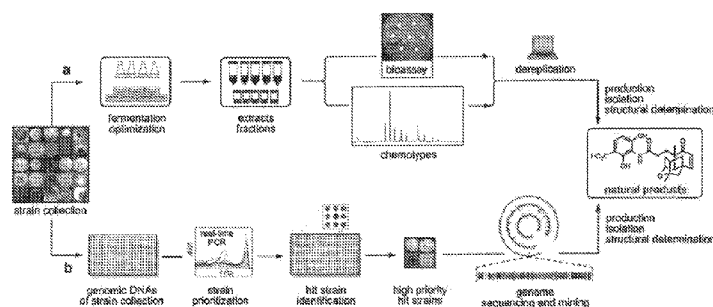
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,
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, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: STRAIN PRIORITIZATION FOR NATURAL PRODUCT DISCOVERY BY A HIGH THROUGHPUT REAL-TIME
PCR METHOD



FIGURES 1A, 1B

(57) Abstract: A high throughput method for identifying organisms which produce agents or classes of compounds. The method allows for the prioritizing or identifying those organisms having the most optimal success in producing the compound.

STRAIN PRIORITIZATION FOR NATURAL PRODUCT DISCOVERY BY A HIGH-THROUGHPUT REAL-TIME PCR METHOD

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

This invention was made with U.S. government support under grant
5 numbers GM086184, AI079070, CA078747 awarded by the National Institutes of Health.
The U.S. government may have certain rights in the invention.

FIELD OF THE INVENTION

Embodiments of the invention are directed to high-throughput methods for
screening and identification of compounds. In particular embodiments, some of the
10 methods utilize high-throughput real-time PCR.

BACKGROUND

Natural products offer unmatched chemical and structural diversity
compared to other small molecule libraries. Thus, natural products remain the best
source of drugs and drug leads and serve as excellent small molecule probes to
15 investigate biological processes (Newman, D.J. & Cragg, G.M., *J. Nat. Prod.* 75, 311-
335 (2012); Li, J.W.H. & Vederas, J.C., *Science* 325, 161-165 (2009)). The recently
discovered diterpenoids platensimycin (PTM) and platencin (PTN) represent two
promising natural product drug leads that target fatty acid biosynthesis. PTM and PTN
were discovered via the systematic screening of fermentation extracts from 83,000
20 strains, each of which was fermented in three different media (affording a total of
~250,000 extracts) (Wang, J. *et al.*, *Nature* 441, 358-361 (2006); Jayasuriya, H. *et al.*,
Angew. Chem., Int. Ed. 46, 4684-4688 (2007); Wang, J. *et al.*, *Proc. Natl. Acad. Sci.*
USA 104, 7612-7616 (2007); Singh, S.B. *et al.*, *J. Am. Chem. Soc.* 128, 11916-11920
(2006); Young, K. *et al.*, *Antimicrob. Agents Chemother.* 50, 519-526 (2006); Genilloud,
25 O. *et al.*, *J. Ind. Microbiol. Biotechnol.* 38:375-389 (2011)). The discovery of PTM and
PTN highlighted the exhaustiveness necessary for the successful isolation of novel
natural products by the traditional methods.

While effective, traditional natural product discovery programs are not
sustainable, demanding too much time, effort, and resources. As the sizes of microbial

strain collections continue to grow, innovations in strain prioritization are clearly needed. Resources could then be devoted preferentially to the strains that hold the highest promise in producing novel natural products, thereby accelerating detection and isolation of the targeted natural products and cutting the time and cost associated with traditional
5 natural product discovery programs. Rapid strain prioritization could fundamentally change how microbial natural products are discovered.

SUMMARY

Embodiments of the invention are directed to drug discovery. In some embodiments, the methods comprise the steps of real-time PCR and targeting genes
10 characteristic to the biosynthetic machinery of natural products with distinct scaffolds, in a high-throughput format. Embodiments are also directed to identification and screening for non-natural or synthetic products.

Other aspects are described *infra*.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A, 1B is a schematic representation showing the strategies for
15 discovering natural product from microorganisms. Figure 1A shows the “grind and find” approach of traditional natural product discovery relying on bioassays or chemotypes. Figure 1B shows the post-genomics approach of natural product discovery featuring a high-throughput real-time PCR method for strain prioritization, genome sequencing, and
20 genome mining of the high priority hit strains.

Figures 2A- 2D show an embodiment of strain prioritization for natural product discovery by a high-throughput real-time PCR method, showcasing the identification of platensimycin, platencin, viguiepinol and oxaloterpin, and other diterpenoid producers by targeting selected diterpene synthases from the actinomycete
25 collection. Figure 2A is a schematic representation showing the pathways for the biosynthesis of bacterial diterpenoid, highlighting the four diterpene synthases and the tailoring enzymes leading to platensimycin (PTM), platencin (PTN), viguiepinol and oxaloterpin, and other known diterpenoids in bacteria. Figure 2B is a schematic representation showing the genetic organization of the PTM-PTN dual biosynthetic gene

cluster in *Streptomyces platensis* MA7327 and the PTN biosynthetic gene cluster in *S. platensis* MA7339, and primer design targeting the genes encoding the four diterpene synthases ent-atiserene synthase (T1), ent-copalyl diphosphate synthase (T2), ent-kaurene synthase (T3), and geranylgeranyl diphosphate synthase (T4). Black arrows depicting the relative locations of the primers with the predicted sizes for each of the products indicated. Figure 2C shows a high-throughput real-time PCR method targeting (i) T4, (ii) T2, (iii) T1, and (iv) T3 and melting curve analysis of the resultant products to identify putative diterpenoid producers from a collection of 1,911 strains. Each panel depicting the melting curves with $\delta F/\delta T$ (y-axis) representing the rate of change in fluorescence as a function of temperature. Solid lines with open circles representing the positive controls with a normalized melting temperature (T_m). Solid lines representing hits found during the melting step of the real-time PCR experiment with T_m range at $T_m \pm 0.8$ °C. Dashed lines representing the negative controls with no template DNA. Insets showing PCR products of the hits that were analyzed by agarose gel electrophoresis and confirmed by DNA sequencing. Only the melting curves for one of the five plates were shown for T4, which yielded 71 putative hits. For T1, T2, and T3, the melting curves of the respective hits from all five plates were combined and depicted together. Figure 2D shows a Euler diagram depicting the 488 putative diterpenoid producers identified from the 1,911 strains by targeting T4, among which nine, six, and six were co-identified by targeting T2, T1, and T3, respectively, and confirmed by DNA sequencing.

Figures 3A-3C show that the new platensimycin and platencin dual-producers are distinct from *S. platensis* MA7327, showing superior genetic amenability. Figure 3A is a scan of a photograph showing the morphology of the six new platensimycin and platencin dual-producers of *Streptomyces platensis* CB00739, CB00765, CB00775, CB00789, CB02289, and CB02304 in comparison with *S. platensis* MA7327 and MA7339 on an ISP4 agar plate. Figure 3B is a phylogenetic tree generated from the alignment of concatenated partial sequences of the four housekeeping genes 16S rRNA, recA, rpoB, and trpB (2975 bp total) using the Tamura-Nei evolutionary distance method and the neighbor-joining algorithm. Bootstrap values $\geq 70\%$ (based on 100 resampled trials) are given at nodes. Bar, 0.02 substitutions per nucleotide position. Figure 3C is a scan of a photograph showing the morphology of the three new

platensimycin and platencin overproducers SB12026, SB12027, and SB12028 in comparison with SB12001 and SB12002 derived from *S. platensis* MA7327, and SB12600 derived from *S. platensis* MA7339 on an ISP4 agar plate. While SB12001 failed to sporulate and SB12002 sporulated poorly under all conditions examined, SB12026, SB12027, and SB12028 sporulated well on several media as exemplified with ISP4.

Figures 4A to 4D show the T1, T2, T3, and T4 sequence analysis for primer design. (Figure 4A; SEQ ID NOS: 39-46) T1, (Figure 4B; SEQ ID NOS: 47-66) T2, (Figure 4C; SEQ ID NOS: 67, 68) T3, and (Figure 4D; SEQ ID NOS: 69-92) T4. For T1, T2, and T4, amino acid sequences were aligned using Clustal 210 and conserved sequences were highlighted using Boxshade. Consensus sequences highlighted in black were chosen for primer design, which was based on the associated nucleotide sequences. Accession codes of proteins used in the alignment: PtmT1 (ACO31274), PtnT1 (ADD83014), Sko3988_Orf2 (BAD86797), Swt1.2 (AEV45183), Swt2.2 (AEW22921), PtmT2 (ACO31276), PtnT2 (ADD83015), PtmT3 (ACO31279), PtmT4 (ACO31283), PtnT4 (ADD83016), Kgris_Orf1 (BAB07816), PlaT4 (ABB69754), Bra2 (BAG16276), and Sko3988_Orf4 (BAD86799).

Figures 5A to 5D show the sequence comparisons of T1, T2, T3, and T4 PCR hit fragments. (Figure 5A; SEQ ID NOS: 93-103) T4, (Figure 5B; SEQ ID NOS: 104-113) T2, (Figure 5C; SEQ ID NOS 114-120) T3, and (Figure 5D; SEQ ID NOS: 121-127) T1. Relative to *ptmT1*, *ptmT2*, *ptmT3*, and *ptmT4* sequences, the percent identities of hit sequences was between 97-98, 63-98, 96-97, and 38-98, respectively.

Figure 6 is a schematic representation showing the genetic organization of the PTM-PTN biosynthetic gene clusters in *S. platensis* MA7327, *S. platensis* CB00739, and *S. platensis* CB00765. The PTM-PTN clusters in *S. platensis* CB00739 (KJ189771) and *S. platensis* CB00765 (KJ189772) share 99% identity in nucleotide sequence, and their nucleotide sequences are 97% identical to *S. platensis* MA7327 (FJ655920).

Figures 7A to 7C show results from the production of platensimycin (●) and platencin (◆) by *Streptomyces platensis* spp. (Figure 7A). Structures of platensimycin (PTM) and platencin (PTN), (Figure 7B) HPLC chromatograms of crude extracts prepared from: left column, original and new PTM-PTN-producing *Streptomyces*

platensis spp.; and right column, previously reported and new $\Delta ptmR1$ deletion mutants. The y-axes are kept constant between the two panels for visualization of the significant increase in titers. Crude extracts for SB12001, SB12002, SB12026, and SB12028 were diluted 4-fold for better representation of PTM and PTN production, (Figure 7C)

5 Extracted ion (m/z at 442.1863 for the $[PTM + H]^+$ ion and m/z at 426.1914 for the $[PTN + H]^+$ ion) chromatograms from LC-MS analyses of: left column, original and new PTM-PTN-producing *Streptomyces platensis* spp.; and right column, previously reported and new $\Delta ptmR1$ deletion mutants.

Figures 8A, 8B show the results from the inactivation of *ptmR1* in three

10 new PTM-PTN producers affording *S. platensis* SB12026, SB12027, and SB12028. (Figure 8A) Schematic representation of the deletion of *S. platensis ptmR1* by insertion of an apramycin resistance-oriT cassette (*aac(3)IV + oriT*), (Figure 8B) PCR verification of wild-type and double crossover mutant genotypes, using the primers *ptmRidF* and *ptmRidR*. Lane 1, 1 Kb Plus DNA ladder (Invitrogen); lane 2, *S. platensis* CB00739;

15 lanes 3-4, two isolates of *S. platensis* SB12026; lane 5, *S. platensis* CB00775; lanes 6-7, *S. platensis* SB12028; lane 8, *S. platensis* CB00765; lanes 9-10, *S. platensis* SB12027.

Figures 9A-9J show results from the studies demonstrating the feasibility of the experiments. (Figure 9A) Alignment of the 11 known enediyne biosynthetic loci, highlighting (i) the enediyne PKS gene cassettes common to all enediynes (shown in red)

20 (note two different clusters, CYA and CYN, for the cyanosporasides and the nomenclature difference for 9- and 10-membered enediyne PKS gene cassettes), (ii) a subset of genes specific for 9-membered enediynes (shown in green), and (iii) pathway regulators (shown in blue). (Figure 9B) Design of PCR primers for enediyne PKS gene cassette taking into consideration of *E5/E/E10* all clustered together or *E5* or *E10*

25 separated from *E*. (Figure 9C) Representative melting curve analysis in real-time PCR in a 384-well plate format, as exemplified by using the *E/E5* primers, with each of the peaks indicating a specific PCR product. (Figure 9D) Confirmation of PCR products, identified on the basis of melting curve analysis, by gel electrophoresis. (Figure 9E) A unified model for enediyne core biosynthesis, featuring (i) enediyne PKS and TE common to

30 both 9- and 10-membered enediynes, (ii) production of heptaene by PKS and TE in the absence of associated enzymes (path a), and (iii) functional interactions between PKS-TE

and 9- or 10-membered specific associated enzymes differentiating 9- (path b) or 10-membered (path c) core biosynthesis. (Figure 9F) Detection of heptaene (●) by HPLC analysis from wild-type enediynes producers and selected *ΔpksE* mutant (i.e., *ΔsgcE*) strains, supporting heptaene as a phenotypic indicator for enediyne production. (Figure 5 9G) HPLC analysis of heptaene (●) production from selected new enediynes producers and their *ΔpksE* mutants (Figure 10) under the conditions examined. (Figure 9H) Correlation between heptaene production and functional expression of the enediynes PKS gene by RT-PCR transcriptional analysis of selected new enediynes producers (CB01580 produced heptaene, while CB02459 did not) with the C-1027 producer *S. globisporus* as 10 a positive control. (Figure 9I) Comparison of C-1027 (♦) and heptaene (●) titers by HPLC analysis between the wild-type and an engineered strain (SB1023), highlighting titer improvement by manipulating pathway regulation. (Figure 9J) Detection and isolation of UCM (◇) by comparison of the metabolite profiles between the UCM producing *S. sp.* wild-type and the *ΔucmE* mutant strains, exemplifying a metabolomics 15 approach for new enediynes discovery. The identity of compound 6 as UCM was confirmed by bioassay and MS analysis.

Figure 10 is a schematic representation showing genome sequencing of 20 of the 89 hits in comparison with 11 enediynes gene clusters whose enediynes structures are known, confirmed that 19 are distinct to all enediynes clusters known to date, hence 20 encoding novel enediynes natural products, and one, CB02366, is identical to the C-1027 cluster and has been confirmed as a new C-1027 producer. While all clusters featured the conserved enediynes PKS cassettes, they are rich in other open reading frames that are unprecedented in gene clusters that encode production of the known enediynes, indicative of novel structural features/functional groups for these new enediynes, and in gene 25 clusters that encode production of other natural products, promising the discovery of new chemistries and biosynthetic knowledge.

Figure 11 shows an embodiment of a high throughput method to survey enediynes biosynthetic machinery in an Actinomycetale collection and prioritize the most promising producers for new enediynes natural product discovery, isolation, structural 30 elucidation, production, and evaluation of the resultant new enediynes as novel anticancer

ADC payload leads. The new enediyne gene cluster or enediyne structures shown are hypothetical.

DETAILED DESCRIPTION

Several aspects of the invention are described below with reference to
5 example applications for illustration. It should be understood that numerous specific details, relationships, and methods are set forth to provide a full understanding of the invention. One having ordinary skill in the relevant art, however, will readily recognize that the invention can be practiced without one or more of the specific details or with other methods. The present invention is not limited by the illustrated ordering of acts or
10 events, as some acts may occur in different orders and/or concurrently with other acts or events. Furthermore, not all illustrated acts or events are required to implement a methodology in accordance with the present invention.

Embodiments of the invention may be practiced without the theoretical aspects presented. Moreover, the theoretical aspects are presented with the understanding
15 that Applicants do not seek to be bound by the theory presented.

All genes, gene names, and gene products disclosed herein are intended to correspond to homologs from any species for which the compositions and methods disclosed herein are applicable. It is understood that when a gene or gene product from a particular species is disclosed, this disclosure is intended to be exemplary only, and is not
20 to be interpreted as a limitation unless the context in which it appears clearly indicates. Thus, for example, for the genes disclosed herein, are intended to encompass homologous and/or orthologous genes and gene products from other organisms.

The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer
25 technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, phage display, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example herein below. However, other equivalent conventional
30 procedures can, of course, also be used. Such conventional techniques and descriptions

can be found in standard laboratory manuals such as Genome Analysis: A Laboratory Manual Series (Vols. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual (all from Cold Spring Harbor Laboratory Press), Stryer, L. (1995) Biochemistry (4th Ed.) Freeman, New York, Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London, Nelson and Cox (2000), Lehninger, Principles of Biochemistry 3rd Ed., W. H. Freeman Pub., New York, N.Y. and Berg *et al.* (2002) Biochemistry, 5th Ed., W. H. Freeman Pub., New York, N.Y., all of which are herein incorporated in their entirety by reference for all purposes.

10 *Definitions*

The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Furthermore, to the extent that the terms "including", "includes", "having", "has", "with", or variants thereof are used in either the detailed description and/or the claims, such terms are intended to be inclusive in a manner similar to the term "comprising."

As used herein, the terms "comprising," "comprise" or "comprised," and variations thereof, in reference to defined or described elements of an item, composition, apparatus, method, process, system, etc. are meant to be inclusive or open ended, permitting additional elements, thereby indicating that the defined or described item, composition, apparatus, method, process, system, etc. includes those specified elements--or, as appropriate, equivalents thereof--and that other elements can be included and still fall within the scope/definition of the defined item, composition, apparatus, method, process, system, etc.

As used in this specification and the appended claims, the term "or" is generally employed in its sense including "and/or" unless the content clearly dictates otherwise.

The term "about" or "approximately" means within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will

depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, “about” can mean within 1 or more than 1 standard deviation, per the practice in the art. Alternatively, “about” can mean a range of up to 20%, up to 10%, up to 5%, or up to 1% of a given value or range. Alternatively, particularly with respect to biological systems or processes, the term can mean within an order of magnitude within 5-fold, and also preferably within 2-fold, of a value. Where particular values are described in the application and claims, unless otherwise stated the term “about” meaning within an acceptable error range for the particular value should be assumed.

By “encoding” or “encoded”, “encodes”, with respect to a specified nucleic acid, is meant comprising the information for translation into the specified protein. A nucleic acid encoding a protein may comprise non-translated sequences (e.g., introns) within translated regions of the nucleic acid, or may lack such intervening non-translated sequences (e.g., as in cDNA). The information by which a protein is encoded is specified by the use of codons. Typically, the amino acid sequence is encoded by the nucleic acid using the “universal” genetic code.

An “isolated nucleic acid” refers to a nucleic acid segment or fragment which has been separated from sequences which flank it in a naturally occurring state, e.g., a DNA fragment which has been removed from the sequences which are normally adjacent to the fragment, e.g., the sequences adjacent to the fragment in a genome in which it naturally occurs, and refers to nucleic acid sequences in which one or more introns have been removed. The term applies to cDNA (complementary DNA) which is a piece of DNA lacking internal, non-coding segments (introns) and transcriptional regulatory sequences. cDNA also can contain untranslated regions (UTRs) that are responsible for translational control in the corresponding RNA molecule. cDNA can be synthesized in the laboratory by reverse transcription from RNA. The term also applies to nucleic acids which have been substantially purified from other components which naturally accompany the nucleic acid, e.g., RNA or DNA or proteins, which naturally accompany it in the cell. The term therefore includes, for example, a recombinant DNA which is incorporated into a vector, into an autonomously replicating plasmid or virus, or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate

molecule (e.g., as a cDNA or a genomic or cDNA fragment produced by PCR or restriction enzyme digestion) independent of other sequences. It also includes a recombinant DNA, for instance, DNA which is part of a hybrid gene encoding additional polypeptide sequences. The term "variant," when used in the context of a polynucleotide
5 sequence, may encompass a polynucleotide sequence related to a wild type gene. This definition may also include, for example, "allelic," "splice," "species," or "polymorphic" variants. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing of exons during mRNA processing. The corresponding polypeptide may possess additional
10 functional domains or an absence of domains. Species variants are polynucleotide sequences that vary from one species to another. Of particular utility in the invention are variants of wild type gene products. Variants may result from at least one mutation in the nucleic acid sequence and may result in altered mRNAs or in polypeptides whose structure or function may or may not be altered. Any given natural or recombinant gene
15 may have none, one, or many allelic forms. Common mutational changes that give rise to variants are generally ascribed to natural deletions, additions, or substitutions of nucleotides. Each of these types of changes may occur alone, or in combination with the others, one or more times in a given sequence.

As used herein, a "target nucleic acid molecule" is a nucleic acid molecule
20 whose detection, amplification, quantitation, qualitative detection, or a combination thereof, is intended. The nucleic acid molecule need not be in a purified form. Various other nucleic acid molecules can also be present with the target nucleic acid molecule. For example, the target nucleic acid molecule can be a specific nucleic acid molecule (which can include RNA such as viral RNA), the amplification of which is intended.
25 Purification or isolation of the target nucleic acid molecule, if needed, can be conducted by methods known to those in the art, such as by using a commercially available purification kit or the like. In embodiments, a target nucleic acid region or segment is a segment of gDNA which comprises conserved nucleic acid sequences. Targeting an oligonucleotide to a particular nucleic acid molecule, in the context of this invention, can
30 be a multistep process. The process usually begins with the identification of a target nucleic acid sequence. This target nucleic acid may be, for example, a gene associated

with biosynthetic machineries (or mRNA transcribed from the gene). The targeting process usually also includes determination of at least one target region, segment, or site within the target nucleic acid as described in detail in the examples section which follows. Within the context of the present invention, the term “region” is defined as a portion of the target nucleic acid having at least one identifiable structure, function, or characteristic. Within regions of target nucleic acids are segments. “Segments” are defined as smaller or sub-portions of regions within a target nucleic acid. “Sites,” as used in the present invention, are defined as positions within a target nucleic acid.

As used herein, “amplification” refers to increase in the number of copies of a nucleic acid molecule. The resulting amplification products are called “amplicons.” Amplification of a nucleic acid molecule (such as a DNA or RNA molecule) refers to use of a technique that increases the number of copies of a nucleic acid molecule in a sample. An example of amplification is the polymerase chain reaction (PCR), in which a sample is contacted with a pair of oligonucleotide primers under conditions that allow for the hybridization of the primers to a nucleic acid template in the sample. The primers are extended under suitable conditions, dissociated from the template, re-annealed, extended, and dissociated to amplify the number of copies of the nucleic acid. This cycle can be repeated. The product of amplification can be characterized by such techniques as electrophoresis, restriction endonuclease cleavage patterns, oligonucleotide hybridization or ligation, and/or nucleic acid sequencing. Other examples of *in vitro* amplification techniques include quantitative real-time PCR; reverse transcriptase PCR; real-time reverse transcriptase PCR (rt RT-PCR); nested PCR; strand displacement amplification (see U.S. Pat. No. 5,744,311); transcription-free isothermal amplification (see U.S. Pat. No. 6,033,881, repair chain reaction amplification (see WO 90/01069); ligase chain reaction amplification (see EP-A-320 308); gap filling ligase chain reaction amplification (see U.S. Pat. No. 5,427,930); coupled ligase detection and PCR (see U.S. Pat. No. 6,027,889); and NASBATM RNA transcription-free amplification (see U.S. Pat. No. 6,025,134) amongst others.

As used herein, the term “complementary” refers to two complementary strands of base pairs of a double-stranded DNA or RNA. Complementary binding occurs when the base of one nucleic acid molecule forms a hydrogen bond to the base of another

nucleic acid molecule. Normally, the base adenine (A) is complementary to thymidine (T) and uracil (U), while cytosine (C) is complementary to guanine (G). For example, the sequence 5'-ATCG-3' of one ssDNA molecule can bond to 3'-TAGC-5' of another ssDNA to form a dsDNA. In this example, the sequence 5'-ATCG-3' is the reverse complement of 3'-TAGC-5'. Nucleic acid molecules can be complementary to each other even without complete hydrogen-bonding of all bases of each molecule. For example, hybridization with a complementary nucleic acid sequence can occur under conditions of differing stringency in which a complement will bind at some but not all nucleotide positions.

Hybridization conditions resulting in particular degrees of stringency will vary depending upon the nature of the hybridization method and the composition and length of the hybridizing nucleic acid sequences. Generally, the temperature of hybridization and the ionic strength (such as the Na^+ concentration) of the hybridization buffer will determine the stringency of hybridization. Calculations regarding hybridization conditions for attaining particular degrees of stringency are discussed in Sambrook *et al.*, (1989) Molecular Cloning, second edition, Cold Spring Harbor Laboratory, Plainview, N.Y. (chapters 9 and 11). The stringency conditions can be varied depending on the sequence identity desired for detection. For example, the very high stringency can detect sequences that share at least 90% identity; high stringency detects stringency can detect sequences that share at least 80% identity; low stringency can detect sequences that share at least 50% identity.

As used herein, "primers" are short nucleic acid molecules, such as a DNA oligonucleotide, for example sequences of at least 5 nucleotides, which can be annealed to a complementary target nucleic acid molecule by nucleic acid hybridization to form a hybrid between the primer and the target nucleic acid strand. A primer can be extended along the target nucleic acid molecule by a polymerase enzyme. Therefore, primers can be used to amplify a target nucleic acid molecule wherein the sequence of the primer is specific for the target nucleic acid molecule, for example so that the primer will hybridize to the target nucleic acid molecule under very high stringency hybridization conditions. The specificity of a primer increases with its length. Thus, for example, a primer that includes 30 consecutive nucleotides will anneal to a target sequence with a higher

specificity than a corresponding primer of only 15 nucleotides. Thus, to obtain greater specificity, probes and primers can be selected that include at least 15, 20, 25, 30, 35, 40, 45, 50 or more consecutive nucleotides.

Primer pairs can be used for amplification of a nucleic acid sequence, for example, by PCR, real-time PCR, or other nucleic-acid amplification methods known in the art. An “upstream” or “forward” primer is a primer 5' to a reference point on a nucleic acid sequence. A “downstream” or “reverse” primer is a primer 3' to a reference point on a nucleic acid sequence. In general, at least one forward and one reverse primer are included in an amplification reaction. PCR primer pairs can be derived from a known sequence, for example, by using computer programs intended for that purpose such as PRIMER® (Version 0.5, 1991, Whitehead Institute for Biomedical Research, Cambridge, Mass.). Methods for preparing and using primers are described in, for example, Sambrook *et al.* (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor, N.Y.; Ausubel *et al.* (1987) *Current Protocols in Molecular Biology*, Greene Publ. Assoc. & Wiley-Intersciences. In one example, a primer includes a label.

The terms “determining”, “measuring”, “evaluating”, “detecting”, “assessing” and “assaying” are used interchangeably herein to refer to any form of measurement, and include determining if an element is present or not. These terms include both quantitative and/or qualitative determinations. Assessing may be relative or absolute. “Assessing the presence of” includes determining the amount of something present, as well as determining whether it is present or absent.

The term “assay” used herein, whether in the singular or plural shall not be misconstrued or limited as being directed to only one assay with specific steps but shall also include, without limitation any further steps, materials, various iterations, alternatives etc., that can also be used. Thus, if the term “assay” is used in the singular, it is merely for illustrative purposes.

A “label” or a “detectable label” is a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radio labeled molecules fluorophores, luminescent compounds, electron-dense reagents, enzymes (e.g., as commonly used in an ELISA),

biotin, digoxigenin, or haptens and proteins which can be made detectable, e.g., by incorporating a label into the peptide.

The term “high-throughput screening” or “HTS” refers to a method drawing on different technologies and disciplines, for example, optics, chemistry, biology
5 or image analysis to permit rapid, highly parallel biological research and drug discovery. HTS methods are known in the art and they are generally performed in multiwell plates with automated liquid handling and detection equipment; however it is envisioned that the methods of the invention may be practiced on a microarray or in a microfluidic system.

10 *Strain Prioritization*

The connection between natural products and the genes encoding their biosynthesis has now been well recognized, and genes, as well as chemistry, are being increasingly exploited to categorize known natural products and discover new ones. Targeting biosynthetic genes using regular PCR methods has proven useful for natural
15 product discovery; however, the lack of high-throughput capabilities has limited their utility in natural product discovery. Real-time PCR, an advancement in PCR technology, has not been fully utilized for natural product discovery efforts. In real-time PCR, each cycle of amplification can be directly monitored based on the detection of fluorescence, hence eliminating gel electrophoresis and staining for visualization as the necessary post-
20 PCR analysis steps. Real-time PCR also offers improved sensitivity for target detection and melting curve analysis for target specificity.

Traditional natural product discovery programs are not sustainable, demanding too much time, effort, and resources.

Accordingly, embodiments of the invention are directed to, *inter alia*, high
25 throughput screening assays for identifying and prioritizing organisms which produce a potential agent having utility as a diagnostic, a therapeutic agent, research tool, etc. The organisms which are targeted can produce a natural product, and/or one which has been transformed with a plasmid and encodes any number of agents. Once the organisms producing certain products are identified these organisms can be classed as a priority

strain in which further studies can be conducted and products isolated and produced for further development.

Briefly, the studies herein provide for a strain prioritization method(s) for natural product discovery, details of which are discussed in the examples section which follows. Central to the method is the application of real-time PCR, targeting genes characteristic to the biosynthetic machinery of natural products with distinct scaffolds in a high-throughput format. The practicality and effectiveness of the method were showcased by prioritizing 1911 actinomycete strains for diterpenoid discovery. A total of 488 potential diterpenoid producers were identified, among which six were confirmed as platensimycin and platencin dual producers and one as a viguiepinol and oxaloterpin producer. While the method as described is most appropriate to prioritize strains for discovering specific natural products, variations of this method is applicable to the discovery of other classes of natural products. Applications of genome sequencing and genome mining to the high-priority strains would essentially eliminate the chance elements from traditional discovery programs and fundamentally change how natural products are discovered.

Accordingly, in some embodiments, a method of high throughput screening for identifying compounds encoded by an organism comprises isolating genomic DNA (gDNA) from the organism; incubating the gDNA with a detectable label; targeting and amplifying gDNA regions or segments comprising conserved nucleic acid sequences; determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences; and, comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying a nucleic acid sequence encoding a compound by an organism. Probes or primers can be used for targeting the regions comprising conserved nucleic acid sequences. The probes or primers can be labeled with a detectable label.

In another embodiment, a method of high throughput screening for identifying organisms encoding a compound comprises isolating genomic DNA (gDNA)

from the organism; targeting and amplifying gDNA regions or segments comprising conserved nucleic acid sequences; determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences; and, comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control
5 and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying a organisms encoding a compound. Probes or primers can be used for targeting the regions comprising conserved nucleic acid sequences. The probes or primers can be labeled with a detectable label. In other embodiments the gDNA is labeled with a detectable label.

10 In another embodiment, a method of high throughput screening for prioritizing organisms encoding compounds comprising isolating genomic DNA (gDNA) from the organism; targeting and amplifying gDNA regions or segments comprising conserved nucleic acid sequences; determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences; comparing the melting temperatures of the
15 amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby prioritizing organisms encoding compounds. In some embodiments the gDNA is labeled with a detectable label.

The melting curve analysis is described in detail in the “Examples” section
20 which follows. The melting curve analysis may generally comprise the following steps: after nucleic acid amplification, a probe bound to a target sequence dissociates from the target sequence and results in changes of fluorescence intensity as the temperature increases; obtaining a melting curve by plotting the rate of fluorescence intensity change (as y axis) as a function of changing temperature (x axis); by detecting during this
25 process, in real-time, the changes of fluorescence intensity with the change of the temperature, variations of the target sequences may be detected using this melting curve. The melting curve mentioned above may also be obtained in a manner of decreasing the temperature, namely from high temperature to low temperature, to detect fluorescence changes. Melting curve analysis is then performed by processing the data obtained.

In some embodiments, a target nucleic acid sequence comprises conserved regions. A conserved region is when two or more nucleic acid sequences (or amino acid sequences) have a high sequence identity or similarity. One can arbitrarily set the percent sequence identity as a parameter when identifying target sequences, for example, a 90%
5 sequence identity between two or more nucleic acid sequences. The conserved regions can be particular to one species or across species, e.g. homologs, paralogs, orthologs. Target nucleic acid sequences can be identified by any known method including databases.

The sequence identity/similarity between two or more nucleic acid
10 sequences, or two or more amino acid sequences, is expressed in terms of the identity or similarity between the sequences. Sequence identity can be measured in terms of percentage identity; the higher the percentage, the more identical the sequences are. Homologs or orthologs of nucleic acid or amino acid sequences possess a relatively high degree of sequence identity/similarity when aligned using standard methods.

15 Methods of alignment of sequences for comparison are well known in the art. Various programs and alignment algorithms are described in: Smith & Waterman, *Adv. Appl. Math.* 2:482, 1981; Needleman & Wunsch, *J. Mol. Biol.* 48:443, 1970; Pearson & Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444, 1988; Higgins & Sharp, *Gene*, 73:237-44, 1988; Higgins & Sharp, *CABIOS* 5:151-3, 1989; Corpet *et al.*, *Nuc. Acids Res.* 16:10881-90, 1988; Huang *et al. Computer Appls. in the Biosciences* 8, 155-65, 1992; and Pearson *et al.*, *Meth. Mol. Bio.* 24:307-31, 1994. Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990, presents a detailed consideration of sequence alignment methods and homology calculations. The NCBI Basic Local Alignment Search Tool (BLAST)
20 (Altschul *et al.*, *J. Mol. Biol.* 215:403-10, 1990) is available from several sources, including the National Center for Biological Information (NCBI, National Library of Medicine, Bethesda, Md. 20894) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn, and tblastx. Blastn is used to compare nucleic acid sequences, while blastp is used to compare amino acid sequences. Additional information can be found at the NCBI web site.

Once aligned, the number of matches is determined by counting the number of positions where an identical nucleotide or amino acid residue is present in both sequences. The percent sequence identity is determined by dividing the number of matches either by the length of the sequence set forth in the identified sequence, or by an articulated length (such as 100 consecutive nucleotides or amino acid residues from a sequence set forth in an identified sequence), followed by multiplying the resulting value by 100. For example, a nucleic acid sequence that has 1166 matches when aligned with a test sequence having 1554 nucleotides is 75.0 percent identical to the test sequence ($1166/1554 \times 100 = 75.0$). The percent sequence identity value is rounded to the nearest tenth. For example, 75.11, 75.12, 75.13, and 75.14 are rounded down to 75.1, while 75.15, 75.16, 75.17, 75.18, and 75.19 are rounded up to 75.2. The length value will always be an integer.

One indication that two nucleic acid molecules are closely related is that the two molecules hybridize to each other under stringent conditions. Stringent conditions are sequence-dependent and are different under different environmental parameters.

Accordingly, in some embodiments, the target nucleic acid sequences are conserved nucleic acid sequences encoding for one or more biosynthetic molecules. However, it is to be understood, that the agents can be any type of molecule produced by an organism. For example, the organism can be transformed with a vector encoding one or more compounds. In other embodiments, the organism comprises one or more mutants, e.g. the organism can be subjected to one or more mutagens. In other embodiments, the organism is cultured under varying conditions and culture media so that the organism produces compounds produced only under certain conditions, e.g. acidic, low oxygen, etc. In some embodiments, the organisms are contacted with regulators which suppress or enhance the transcription of one molecule over another. In some embodiments, a mammalian cell is used as the "organism." In some embodiments, the conserved nucleic acid sequences encoding for the biosynthetic molecules are spliced into an expression vector. The expression vector encoding the biosynthetic molecule is transfected into an appropriate host cell to express the biosynthetic molecule. In some embodiments the nucleic acid sequences are mutated or are synthetic, the synthetic nucleic acid sequences comprising one or more segments of sequences having at least

about 50% sequence identity to an isolated nucleic acid sequence encoding for an identified product. The organisms may be cultured under varying culture conditions and culture media.

A wide variety of host/expression vector combinations may be employed in expressing the DNA sequences of this invention. Useful expression vectors, for example, may consist of segments of chromosomal, non-chromosomal and synthetic DNA sequences. Suitable vectors include derivatives of SV40 and known bacterial plasmids, e.g., *E. coli* plasmids col E1, pCR1, pBR322, pMal-C2, pET, pGEX (Smith et al., *Gene* 67:31-40, 1988), pMB9 and their derivatives, plasmids such as RP4; phage DNAs, e.g., the numerous derivatives of phage 1, e.g., NM989, and other phage DNA, e.g., M13 and filamentous single stranded phage DNA; yeast plasmids such as the 2 μ plasmid or derivatives thereof, vectors useful in eukaryotic cells, such as vectors useful in insect or mammalian cells; vectors derived from combinations of plasmids and phage DNAs, such as plasmids that have been modified to employ phage DNA or other expression control sequences; and the like.

Yeast expression systems can also be used according to the invention to express the biosynthetic molecules. For example, the non-fusion pYES2 vector (XbaI, SphI, ShoI, NotI, GstXI, EcoRI, BstXI, BamHI, SacI, KpnI, and HindIII cloning sites; Invitrogen) or the fusion pYESHisA, B, C (XbaI, SphI, ShoI, NotI, BstXI, EcoRI, BamHI, SacI, KpnI, and HindIII cloning sites, N-terminal peptide purified with ProBond resin and cleaved with enterokinase; Invitrogen), to mention just two, can be employed according to the invention. A yeast two-hybrid expression system can be prepared in accordance with the invention.

In embodiments, once the target nucleic acid sequences have been identified, and the primers synthesized, these regions comprising the conserved nucleic acid sequences are amplified by real time polymerase chain reaction (RT-PCR). Real-time PCR is a method for detecting and measuring products generated during each cycle of a PCR, which are proportionate to the amount of template nucleic acid prior to the start of PCR. The information obtained, such as an amplification curve, melting temperatures can be used to determine the presence of a target nucleic acid and/or quantitate the initial amounts of a target nucleic acid sequence. In some examples, real time PCR is real time

reverse transcriptase PCR (rt RT-PCR). In some embodiments, the real time PCR is a multiplex real time PCR targeting multiple genes in a sample. Details of the RT-PCR are provided in the "Examples" section which follows.

In some embodiments, the primers for targeting a particular conserved
5 nucleic acid sequence optionally comprise one or more modifications. The primers may be formed as composite structures of two or more oligonucleotides, modified oligonucleotides, oligonucleosides and/or oligonucleotide mimetics. Such compounds have also been referred to in the art as hybrids or gapmers. Representative United States patents that teach the preparation of such hybrid structures comprise, but are not limited
10 to, U.S. Pat. Nos. 5,013,830; 5,149,797; 5,220,007; 5,256,775; 5,366,878; 5,403,711; 5,491,133; 5,565,350; 5,623,065; 5,652,355; 5,652,356; and 5,700,922, each of which is herein incorporated by reference.

Specific examples of some modified primers include those comprising modified backbones, for example, phosphorothioates, phosphotriesters, methyl
15 phosphonates, short chain alkyl or cycloalkyl intersugar linkages or short chain heteroatomic or heterocyclic intersugar linkages. Examples of oligonucleotides with phosphorothioate backbones and those with heteroatom backbones, include without limitation: $\text{CH}_2\text{-NH--O--CH}_2$, $\text{CH}_2\text{--N(CH}_3\text{)--O--CH}_2$ [known as a methylene(methylimino) or MMI backbone], $\text{CH}_2\text{-O--N(CH}_3\text{)-CH}_2$, $\text{CH}_2\text{-N(CH}_3\text{)-N}$
20 $\text{(CH}_3\text{)-CH}_2$ and $\text{O--N(CH}_3\text{)-CH}_2\text{-CH}_2$ backbones, wherein the native phosphodiester backbone is represented as O--P--O--CH_2 . The amide backbones disclosed by De Mesmacker *et al.* (1995) *Acc. Chem. Res.* 28:366-374 are also one example. In other embodiments, a primer comprises morpholino backbone structures (Summerton and Weller, U.S. Pat. No. 5,034,506). In other embodiments, such as the peptide nucleic acid
25 (PNA) backbone, the phosphodiester backbone of the nucleic acid sequence is replaced with a polyamide backbone, the nucleotides being bound directly or indirectly to the nitrogen atoms of the polyamide backbone (Nielsen *et al.* (1991) *Science* 254, 1497). Nucleic acid sequences may also comprise one or more substituted sugar moieties.

The primers may also include, additionally or alternatively, nucleobase
30 (often referred to in the art simply as "base") modifications or substitutions. As used

herein, “unmodified” or “natural” nucleotides include adenine (A), guanine (G), thymine (T), cytosine (C) and uracil (U). Modified nucleotides include nucleotides found only infrequently or transiently in natural nucleic acids, e.g., hypoxanthine, 6-methyladenine, 5-Me pyrimidines, particularly 5-methylcytosine (also referred to as 5-methyl-2'
5 deoxycytosine and often referred to in the art as 5-Me-C), 5-hydroxymethylcytosine (HMC), glycosyl HMC and gentobiosyl HMC, as well as synthetic nucleotides, e.g., 2-aminoadenine, 2-(methylamino)adenine, 2-(imidazolylalkyl)adenine, 2-(aminoalkylamino)adenine or other heterosubstituted alkyladenines, 2-thiouracil, 2-thiothymine, 5-bromouracil, 5-hydroxymethyluracil, 8-azaguanine, 7-deazaguanine, N6
10 (6-aminohexyl)adenine and 2,6-diaminopurine. (Kornberg, A., DNA Replication, W.H. Freeman & Co., San Francisco, 1980, pp 75-77; Gebeyehu, G., (1987) *et al. Nucl. Acids Res.* 15:4513). A “universal” base known in the art, e.g., inosine, may be included.

In embodiments, the nucleic acid sequences, gDNA, primers, probes etc., comprise a detectable label. The detectable label allows for the detection and melting
15 temperature analysis of the amplified products. In some embodiments, the detectable label comprises: nucleic acid intercalating dyes, fluorophores, radio labeled molecules, fluorescent agents, molecular beacons, chemiluminescent agents, luminescent agents, electron-dense reagents, enzymes, biotin, digoxigenin, haptens or peptides.

In another embodiment, a method of high throughput screening for
20 prioritizing bacterial strains producing diterpenoids comprises isolating genomic DNA (gDNA) from the organism; targeting and amplifying gDNA regions comprising conserved nucleic acid sequences; determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences; comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and
25 selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying a nucleic acid sequence encoding a diterpenoid by a bacterial strain.

In some embodiments, the bacterial strain is an Actinomycetale strain.

In some embodiments, the conserved nucleic acid sequences encode for
30 one or more diterpene synthase molecules. In some embodiments, the conserved

diterpene synthase sequences are targeted by primers having a sequence identity of at least about 50% with at least five consecutive nucleic acid bases of a 5' to 3' or 3' to 5' strand of the conserved diterpene synthase sequences. In other embodiments, the primers comprise any one or combinations of primers set forth in SEQ ID NOS: 1-127. In other
5 embodiments, the primers comprise any one or combinations of primers set forth in SEQ ID NOS: 1-31.

In other embodiments, a compound is identified by any of the methods embodied herein. In some embodiments, a pharmaceutical composition comprises a compound identified by the methods embodied herein.

10 In embodiments, an oligonucleotide comprises the nucleic acid sequences set forth as SEQ ID NOS: 1-127, mutants, variants or complementary sequences thereof. In other embodiments, an oligonucleotide comprises a nucleic acid sequence having at least about 50% sequence identity to one or more oligonucleotides set forth as SEQ ID NOS: 1-127 or complementary sequences thereof.

15 In other embodiments, a method of high throughput screening for prioritizing bacterial strains producing enediynes comprises; isolating genomic DNA (gDNA) from the organism; targeting and amplifying regions in the gDNA comprising conserved nucleic acid sequences; determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences; comparing the melting temperatures of the
20 amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying a nucleic acid sequence encoding an enediyne by a bacterial strain. In some embodiments the gDNA is labeled with a detectable label.

25 In embodiments, the conserved nucleic acid sequences encode for one or more enediyne molecules.

In other embodiments, a conserved nucleic acid sequence comprises polyketide synthase (PKS) nucleic acid sequences, mutants or variants thereof.

In other embodiments, a conserved enediyne nucleic acid producer sequence is targeted by primers having a sequence identity of at least about 50% with at least five consecutive nucleic acid bases of a 5' to 3' or 3' to 5' strand of the conserved enediyne nucleic acid producer sequences. In some embodiments, the primers comprise a detectable label.

In other embodiments, the primers comprise any one or combinations of primers set forth in SEQ ID NOS: 33-38 or 1-127.

The invention has been described in detail with reference to preferred embodiments thereof. However, it will be appreciated that those skilled in the art, upon consideration of this disclosure, may make modifications and improvements within the spirit and scope of the invention.

All documents mentioned herein are incorporated herein by reference. All publications and patent documents cited in this application are incorporated by reference for all purposes to the same extent as if each individual publication or patent document were so individually denoted. By their citation of various references in this document, Applicants do not admit any particular reference is "prior art" to their invention.

EXAMPLES

While various embodiments of the present invention have been described above, it should be understood that they have been presented by way of example only, and not limitation. The following non-limiting examples are illustrative of the invention.

Example 1: Strain Prioritization for Natural Product Discovery by A High-Throughput Real-Time PCR Method – Discovery of Ditepenoids in General and Platensimycin and Platencin in Particular

A high-throughput method is described for strain prioritization by real-time PCR to identify the most promising strains from a microbial strain collection for natural product discovery (Figure 1B). Considering the tremendous effort spent in discovering the original PTM and PTN producers, *Streptomyces platensis* MA7327 and MA7339, respectively, as well as the limited genetic amenabilities of the two producers for metabolic pathway engineering, new PTM and PTN producers from a collection of

1,911 actinomycete strains were chosen for identification to demonstrate the utility and effectiveness of this method. Six new PTM and PTN producers were identified – (i) all six were confirmed to contain the targeted diterpene synthase genes and verified as true PTM and PTN producers, (ii) three of them were demonstrated to overproduce PTM and PTN upon inactivation of the pathway-specific negative regulator, and (iii) two of them were confirmed to harbor the complete PTM-PTN dual biosynthetic gene clusters. The six new PTM and PTN producers, all classified as *Streptomyces platensis* species on the basis of their housekeeping genes, show distinct morphology, differing from *S. platensis* MA7327 and MA7339, and three of them showed superior genetic amenability. While the method as described is most appropriate to prioritize strains for discovering specific natural products, variations of this method should be applicable to the discovery of other classes of products, natural or otherwise.

Materials and Methods

Accession codes: The following partial gene sequences were deposited to GenBank for the *S. platensis* strains MA7327, MA7339, CB00739, CB00765, CB00775, CB00789, CB02289 and CB02304, respectively: 16S rRNA, KJ469279-KJ469286; *recA*, KJ469287-KJ469294; *rpoB*, KJ469295-KJ469302; and *trpB*, KJ469303-KJ469310. The following complete gene sequences were deposited to GenBank for the *S. platensis* strains CB00739, CB00765, CB00775, CB00789, CB02289 and CB02304, respectively: *ptmT1*, KJ469311-KJ469316; *ptmT2*, KJ469317-KJ469322; *ptmT3*, KJ469323-KJ469328; *ptmT4*, KJ469329-KJ469334; and *ptmR1*, KJ469335-KJ469340. The PTM-PTN gene clusters for *S. platensis* CB00739 and CB00765 were deposited as KJ189771 and KJ189772, respectively.

Preparation of genomic DNA library. The actinomycete collection at The Scripps Research Institute consists of strains isolated from various unexplored and underexplored ecological niches (Xie, P. *et al.*, *J. Nat. Prod.* 77, 377-387 (2014)). The cultivation of actinomycete strains and genomic DNA (gDNA) preparation followed standard protocols (Kieser, T. *et al.*, *Practical Streptomyces Genetics* (The John Innes Foundation, Norwich, UK, 2000)). The concentrations of genomic DNA samples were estimated using a microplate fluorescence assay with minor modifications (Leggate, J. *et*

al., *Biotechnol. Lett.* 28, 1587-1594 (2006)). Briefly, gDNA samples were diluted to several dilution factors in 10 mM Tris-HCl, pH 8.0. SYBR Green I dye (Sigma-Aldrich) was added to each dilution of DNA, mixed, and incubated at room temperature for 10 min in the dark. Fluorescence levels were measured using a SpectraMax M5 Multi-Mode
5 Microplate Reader (Molecular Devices) with excitation and emission wavelengths of 450 and 520 nm, respectively. After the normalization of gDNA concentration, samples were arrayed in 384-well masterblock deep-well plates as working stocks. gDNA samples were transferred (from working stocks) to microplates using Biomek FX workstation (Beckman Coulter).

10 *Real-time PCR.* Real-time PCR was performed with an Applied Biosystems 7900HT Fast Real-Time PCR system. Primer design was based on the *ptmT3* sequence and the conserved sequences of *ptmT1*, *ptmT2*, and *ptmT4* homologues (Figure 4A to 4D). Genomic DNA (1-100 ng), 0.1 μ L of each primer (10 μ M stock), 0.5 μ L of DMSO, 0.5 μ L of 10X SYBR Green I dye, 5 μ L of Taq 2X Master Mix, and water were
15 mixed to give a reaction volume of 10 μ L per well. A “no template” negative control and one positive control with the genomic DNA of *S. platensis* MA7327 were included in each 384-well plate. The reaction conditions consisted of a background check at 50 °C for 2 min; initial denaturation at 95 °C for 7 min; 37-40 cycles of denaturation at 95 °C for 30 s, primer annealing at 64-68 °C for 15 s, extension at 68 °C for 30-60 s, and
20 melting at 95 °C for 15 s with a ramp rate of 2% from 68 °C to 95 °C. For T4 amplification, three additional cycles of denaturation at 95 °C for 30 s, temperature ramping to 30 °C, and extension at 68 °C for 30 s were included after the initial denaturation step mentioned above. The melting step subjected amplicons to a range of temperatures for determining melting temperatures (T_m). Each T_m was normalized to a
25 theoretical T_m calculated using the nearest neighbor model (Kibbe W.A. *et al.*, *Nucleic Acids Res.* 35 (suppl 2), W43-W46 (2007); Breslauer K.J. *et al.*, *Proc. Natl. Acad. Sci. USA* 83, 3746-3750 (1986); Sugimoto N. *et al.*, *Nucleic Acids Res.* 24, 4501-4505 (1996)). For determining each theoretical T_m , the sequence of the target gene was used assuming an initial DNA amount of 100 ng and salt concentration of 50 mM. Samples
30 with $T_m \pm 0.8$ °C compared to the positive control were considered hits. To confirm hits,

regular PCR was performed, and the resulting products were analyzed by 1% agarose gel electrophoresis, purified by gel extraction, and sequenced.

DNA sequencing. To amplify and sequence the T1, T2, T3, T4 and R1 genes of six new PTM-PTN dual producers (CB00739, CB00765, CB00775, CB00789, 5 CB02289, and CB02304), primers were designed (Table 2) according to the PTM-PTN dual biosynthetic gene cluster from *S. platensis* MA7327. The amplification and sequencing of four housekeeping gene (16S rRNA, *recA*, *rpoB*, *trpB*) regions was performed using primers listed in Table 2. The draft genome sequences of *S. platensis* CB00739 and CB00765 was obtained using the Ion 316™ chips and the Ion PGM™ 10 Sequencing 300 kit (Life Technologies) following the manufacturer's instructions. Fragment libraries (~300 bp inserts) and matepair libraries (~5 kb inserts) were sequenced. For *S. platensis* CB00739, a final sequence assembly of 5,955,880 reads was completed using GS De Novo Assembler (454 Life Sciences), including 3,018,552 reads from fragment ends and 2,937,328 reads from matepair ends, resulting in a 9,441,068 bp- 15 draft genome consisting of 23 scaffolds. For *S. platensis* CB00765, a final sequence assembly of 5,293,288 reads was completed, including 2,963,442 reads from fragment ends and 2,329,846 reads from matepair ends, resulting in a 9,404,443 bp-draft genome consisting of 20 scaffolds. Gaps found within the PTM-PTN gene clusters were manually filled by PCR. Putative protein-coding sequences were predicted using 20 Glimmer 3.02 (Salzberg S.L. *et al.*, *Nucleic Acids Res.* 26, 544-548 (1998); Delcher A.L. *et al.*, *Nucleic Acids Res.* 27, 4636-4641 (1999)) and the open reading frames in the PTM-PTN gene clusters were analyzed by FramePlot (Ishikawa J. & Hotta, K. *FEMS Microbiol. Lett.* 174, 251-253 (1999)). All DNA sequences were deposited in the NCBI database (see Accession codes).

25 *Phylogenetic analysis.* Alignment of the concatenated housekeeping genes (partial sequences, 2975 bp total) was generated using MegAlign 10.1 (Ikeda C. *et al.*, *J. Biochem.* 141, 37-45 (2007); Guo Y. *et al.*, *Int. J. Syst. Evol. Microbiol.* 58, 149-159 (2008)). Phylogenetic analysis was performed using the MEGA 5.2.1 software (Tamura, K. *et al.*, *Mol. Biol. Evol.* 28, 2731-2739 (2011)). The phylogenetic tree was 30 constructed using the Tamura-Nei evolutionary distance method and 1000 bootstrap replications. Nucleotide sequences of the housekeeping genes for *S. coelicolor* A3(2)

(AL645882), *S. avermitilis* MA-4680 (BA000030), *S. rimosus* ATCC 10970 (ANSJ000000000), *S. griseus* NBRC 13350 (AP009493), *S. scabiei* 87.22 (NC_013929), *S. venezuelae* ATCC 10712 (FR845719), and *S. clavuligerus* ATCC 27064 (ADWJ000000000) were obtained from the NCBI database and used as relevant

5 *Streptomyces* spp. for phylogenetic analysis. *Streptosporangium roseum* DSM43021 (CP001814) and *Mycobacterium tuberculosis* H37Rv (AL123456) were used as outgroups.

Gene disruption. Strains and plasmids/cosmids used in this study are summarized in Table 3. Gigapack III XL packaging extracts and *E. coli* XL1-Blue MRF' (Agilent) were used for cosmid library construction. Primers used in the gene disruption

10 experiments are listed in Table 3. Gene disruption of *ptmR1* was performed in *E. coli* BW25113/pIJ790 (Gust B. *et al.*, *Proc. Natl. Acad. Sci. USA* 100, 1541-1546 (2003)) carrying appropriate cosmids. The *ptmR1* genes were replaced with the *aac(3)IV* + *oriT* resistance cassette from pIJ773. The mutant cosmids were introduced into the *S.*

15 *platensis* CB00739, CB00765, and CB00775 by intergeneric conjugation (Bierman M. *et al.*, *Gene* 116, 43-49 (1992)) after passaging the mutant cosmids through the non-methylating *E. coli* ET12567/pUZ8002 (Kieser T. *et al.*, *Practical Streptomyces Genetics* (The John Innes Foundation, Norwich, UK, 2000)). Apramycin selection, and kanamycin sensitivity, on ISP4 medium were used to determine double crossover mutants of the

20 *ptmR1* genes. The mutations were confirmed by PCR analysis.

Production of PTM and PTN. PTM and PTN were produced from *Streptomyces* spp. following previously published procedures (Smanski M.J. *et al.*, *Proc. Natl. Acad. Sci. USA* 108, 13498-13503 (2011); Smanski M.J. *et al.*, *Antimicrob. Agents Chemother.* 53, 1299-1304 (2009); Yu Z. *et al.*, *Org. Lett.* 12, 1744-1747 (2010)).

25 Briefly, *Streptomyces* spp. spores were inoculated into seed medium, incubated for 48 h, and 2 mL of seed culture were used to inoculate 50 mL of PTM medium (REF) supplemented with 10 mL L⁻¹ trace elements and 1.5 g of Amberlite XAD-16 resin (Sigma-Aldrich). After incubation for 7 days at 28 °C and 250 rpm, the resin was harvested by centrifugation, washed three times with water, and extracted with actone (3

30 × 10 mL). Acetone was removed under reduced pressure and the resulting oil was resuspended in 1.5 mL of methanol.

HPLC and LC-MS analysis. HPLC was carried out on a Varian Liquid Chromatography System consisting of Varian ProStar 210 pumps and a ProStar 330 photodiode array detector equipped with an Apollo C₁₈ column. HPLC analysis of PTM and PTN was performed using a 20 min solvent gradient (1 mL min⁻¹) from 15% acetonitrile in H₂O containing 0.1% formic acid to 90% acetonitrile in H₂O containing 0.1% formic acid. The peak area at 240 nm was used to quantify PTM and PTN production.

Liquid chromatography-mass spectrometry (LCMS) was carried out on an Agilent 1260 Infinity LC coupled to a 6230 TOF equipped with an Agilent Extend C₁₈ column (50 mm × 2.1 mm, 1.8 μm). Liquid chromatography was performed using a 17 min solvent gradient (0.4 mL min⁻¹) from 10% acetonitrile (CH₃CN) in H₂O containing 0.1% formic acid to 100% acetonitrile in H₂O containing 0.1% formic acid, and PTM and PTN were eluted with retention times of 7.4 and 9.7 min, respectively. Extracted ion (*m/z* at 442.1863 for the [PTM + H]⁺ ion and *m/z* at 426.1914 for the [PTN + H]⁺ ion) chromatograms (*m/z* at [M + H]⁺ ± 0.5) verified PTM and PTN production.

Results

Development of a strain prioritization method by real-time PCR for natural product discovery. The continued growth of microbial strain collections used for natural product discovery demands a high-throughput method for rapid strain prioritization. Sensitive and reliable survey of microbial genomes in a strain collection for genes characteristic for biosynthesis of the targeted class of natural products would identify the most promising ones for strain prioritization, thereby significantly increasing the likelihood of discovering the targeted class of natural products. The methods described herein, exploit the capability of real-time PCR, in a high-throughput manner, to detect DNA regions, encoding biosynthesis of the targeted class of natural products, from the genomic DNA samples of a microbial strain collection for strain prioritization. Subsequent genome sequencing and genome mining of the high priority hit strains promise to accelerate natural product discovery (Figure 1B).

In real-time PCR applications, fluorescence levels, which correspond to the amount of accumulating products, are generated by fluorophores that intercalate

double-stranded (ds) DNA or by fluorophore-conjugated oligonucleotides that prime to DNA targets. SYBR Green I was chosen as the DNA intercalating fluorophore as it enabled us to apply melting curve analysis of the PCR products. Melting curve analysis reveals the apparent melting temperature (T_m) of the PCR products, a value dependent on product length, base composition, and nucleotide mismatch (Wetmur, J.G., *Crit. Rev. Biochem. Mol. Biol.* 26, 227-259 (1991)). The T_m values therefore can be exploited as an indicator of PCR product specificity without the need for gel electrophoresis.

To demonstrate the utility and effectiveness of the strain prioritization method by real-time PCR, the diterpene synthase genes were targeted to search for diterpenoid producers in general and PTM and PTN producers in particular from the actinomycete strain collection at The Scripps Research Institute. Diterpenoids of bacterial origin are underrepresented among known natural products (Smanski, M.J., Peterson, R.M., Huang, S.-X. & Shen, B. *Curr. Opinion Chem. Biol.* 16, 132-141 (2012)). All diterpenoids are derived from the common precursor geranylgeranyl diphosphate (GGDP). Diterpene synthases catalyze the critical steps in diterpenoid biosynthesis by morphing geranylgeranyl diphosphate into one of the many diterpenoid scaffolds, further transformations of which by pathway-specific tailoring enzymes afford the vast structural diversity known for diterpenoid natural products (Figure 2A).

The biosynthetic gene clusters for PTM and PTN production in *S. platensis* MA7327 and MA7339 (Figure 2B) (Smanski, M.J. *et al.*, *Proc. Natl. Acad. Sci. USA* 108, 13498-13503 (2011)) were previously cloned and characterized in the inventor's laboratory. Among the four diterpene synthases from the PTM and PTN biosynthetic machineries, geranylgeranyl diphosphate synthase (PtmT4/PtnT4) is common for all bacterial diterpenoids, *ent*-copalyl diphosphate synthase (PtmT2/PtnT2) is shared by both PTM and PTN, but *ent*-kaurene synthase (PtmT3) or *ent*-atiserene synthase (PtmT1/PtnT1) is specific for PTM or PTN, respectively (Figure 2A). It was reasoned that targeting T1 or T3 should specifically narrow the search to producers of *ent*-atiserene-derived or *ent*-kaurene-derived diterpenoids, such as PTM or PTN, while targeting T2 or T4 could broaden the search to include producers that biosynthesize other *ent*-copalyl diphosphate-derived or all geranylgeranyl diphosphate-derived diterpenoids, respectively (Figure 2A).

The primers, targeting the genes that encode T1, T2, T3 and T4, were designed based on the conserved nucleotide sequences of *ptmT1/ptnT1*, *ptmT2/ptnT2*, *ptmT3*, and *ptmT/ptnT4* within the PTM and PTN gene clusters, as well as known homologues of actinomycete origin, and one of each primers for T1 and T3 was selected
 5 based on the atypical amino acid motifs (DXXXD; SEQ ID NO: 32) found in bacterial type I diterpene synthases (Smanski, M.J. *et al.*, *Curr. Opinion Chem. Biol.* 16, 132-141 (2012); Smanski, M.J. *et al.*, *Proc. Natl. Acad. Sci. USA* 108, 13498-13503 (2011)) (Figures 4A-4D and Table 2). The sizes for each of the PCR products were predicted to be 696 bp, 559 bp, 461 bp, or 411 bp, with the calculated T_m of 92.1 °C, 91.6 °C, 91.6 °C,
 10 or 91.8 °C for T1, T2, T3, or T4, respectively, according the *ptmT1/ptnT1*, *ptmT2/ptnT2*, *ptmT3*, and *ptmT/ptnT4* sequences (Figure 2B). The specificity cutoff for each of the targeted PCR products was arbitrarily set at $T_m \pm 0.8^\circ\text{C}$, where T_m is the melting temperature of the positive control, to identify putative hits. The deviation from T_m is inversely proportional to PCR product specificity, and smaller deviations were preferred
 15 to minimize false positives.

Strain prioritization by real-time PCR for four diterpene synthase genes affording six PTM and PTN producers. 1,911 strains were selected from the actinomycete collection (Xie, P. *et al.*, *J. Nat. Prod.* 77, 377-387 (2014)) and applied the method described above to prioritize the strains for diterpenoid producers in general and
 20 PTM and PTN producers in particular. The genomic DNAs for each of the 1,911 strains were individually prepared, normalized, and arrayed in 384-well plates. Each of the plates also contained one blank (i.e., no template DNA) as a negative control and one with the genomic DNA of *S. platensis* MA7327 as a positive control. Real-time PCRs were carried out with each of the four sets of primers (Figure 2B and Table 2), and the
 25 resultant PCR products were subjected to melting curve analysis to identify putative hits (Figure 2C).

As summarized in Figure 2D: 488, 17, 6, and 6 out of the 1,911 strains were identified as putative hits using the primers, targeting T4, T2, T3, and T1, respectively. Although the 488 hits identified using the T4 primers represented a hit rate
 30 that was too high to justify detailed follow up analysis of each hit, confirmatory PCR on selected hits, using the same T4 primers followed by agarose gel electrophoresis of the

resultant products, yielded specific bands with the predicted size (Figure 2C, panel i). However, the resultant PCR products from each of the T4 hits were not all sequenced (Figure 5A), and the possibility of false positives due to nonspecific PCR amplification therefore cannot be excluded. For the 17 hits identified using the T2 primers, follow up
5 analysis by PCR using the same T2 primers all afforded specific products with the predicted size (Figure 2C, panel ii). DNA sequencing of the resultant products revealed that nine out of the 17 hits were true hits that contain genes encoding putative *ent*-copalyl diphosphate synthase (Figure 5B), with the remaining eight hits resulting from nonspecific PCR amplification. Finally, each set of the six hits identified using T3 or T1
10 primers, respectively, were similarly validated by PCR using the same T3 and T1 primers, and remarkably both set of six hits afforded specific PCR products with the predicted size (Figure 2C, panels iii and iv). DNA sequencing of the resultant products confirmed that both set of six hits were true diterpenoid producers that contain genes encoding the putative *ent*-kaurene synthase (Figure 5C) or *ent*-atiserene synthase (Figure
15 5D), respectively.

Cross examination of all hits revealed that six strains, CB00739, CB00765, CB00775, CB00789, CB02289, and CB02304, harbored all four targeted genes (Figure 2D). The fact that they possessed both T1 and T3, in addition to T2 and T4, would predict all six hits as PTM-PTN dual producers, a genetic disposition that
20 resembled the PTM-PTN dual producer *S. platensis* MA7327 but contrasted to the PTN-only producer *S. platensis* MA7339 (Figures 2A and 2B). Of the remaining three hits identified using T2, they (CB00028, CB00830, and CB01059) all also harbored T4 as would be expected for diterpenoid producers (Figure 2A). In fact, this laboratory previously isolated from CB00830 viguiepinol, oxaloterpin E and C, diterpenoids whose
25 biosynthesis was confirmed to involve T2 and T4 (Xie, P. *et al.*, *J. Nat. Prod.* 77, 377-387 (2014); Ikeda, C. *et al.*, *J. Biochem.* 141, 37-45 (2007)). Therefore, CB00028 and CB01059 are promising producers of new diterpenoid natural products featuring *ent*-copalyl diphosphate-derived scaffolds (Figure 2A).

DNA sequencing confirming the six hits to harbor PTM-PTN dual biosynthetic gene clusters. To verify each of the six hits possessing a functional PTM-PTN dual biosynthetic cluster gene, selected intact genes, encoding the four diterpene
30

synthases T1, T2, T3, and T4, as well as the pathway-specific negative transcriptional regulator PtmR1, were first amplified by PCR using primers that were designed according to the PTM-PTN dual biosynthetic gene cluster from *S. platensis* MA7327 (Table 2). DNA sequencing of the resultant products confirmed that each of the targeted
5 genes were highly homologous (>95%) to those of the known PTM-PTN producers and each other (Figure 6). The hit strains were next subjected to polyphasic taxonomy studies. Phylogenetic analysis based on four selected housekeeping genes (16S rRNA, *recA*, *rpoB*, and *trpB*) (Table 2) assigned all six hits as *Streptomyce platensis* species (Figure 3A). Close examination of the six hit strains, in comparison with *S. platensis*
10 MA7327 and MA7339, however, revealed that the new strains were distinct as judged by the morphology and appearance of pigmented spores (Figure 3B). Two of the hit strains, CB00739 and CB00765, were selected and subjected to genome sequencing, confirming that these strains contained complete PTM-PTN gene clusters. The PTM-PTN dual biosynthetic gene clusters from CB00739 and CB00765 have the same number of open
15 reading frames and same genetic organization as the one from *S. platensis* MA7327, and the overall nucleotide sequence identities among the three clusters are >97% (Figure 6).

The six hits produce and three of their engineered strains overproduce PTM and PTN. To validate that the identified PTM-PTN dual biosynthetic gene clusters are functional, the six hit strains were first subjected to fermentation optimization for
20 PTM and PTN production and confirmation. Under the same conditions used for *S. platensis* MA7327 and MA7339, (Smanski, M.J. *et al.*, *Proc. Natl. Acad. Sci. USA* 108, 13498-13503 (2011); Smanski, M.J. *et al.*, *Antimicrob. Agents Chemother.* 53, 1299-1304 (2009); Yu, Z. *et al.*, *Org. Lett.* 12, 1744-1747 (2010)) PTM and PTN production was confirmed for all six hit strains (Figures 7A-7C), with PTM and PTN titers varying from
25 0.66 – 10 mg L⁻¹ and <0.1 – 11 mg L⁻¹, respectively (Table 1). These titers correlated well with those reported previously from the *S. platensis* MA7327 and MA7339 wild-type strains.

Removal of the pathway specific, negatively acting transcriptional regulator PtmR1 or PtnR1 from *S. platensis* MA7327 and MA7339, afforded
30 recombinant strains that dramatically overproduced PTM, PTN or both. However, the recombinant strains sporulated poorly under all conditions examined (Figure 3C). This

has prevented further engineering PTM and PTN biosynthesis in these overproducers. PtmR1 was similarly inactivated in three of the six hits, CB00739, CB00765, and CB00775 (Table 3), and the genotypes of the resultant recombinant strains SB12026, SB12027, and SB12028, respectively, were confirmed by PCR (Figures 8A, 8B). Under the same conditions for PTM and PTN production with the wild-type strains as controls, SB12026, SB12027, and SB12028 indeed overproduced PTM and PTN up to 310-fold, consistent with what has been observed for the *S. platensis* MA7327 and MA7339 strains (Table 1 and Figures 4A-4C). Gratifyingly, SB12026, SB12027, and SB12028 sporulate well (Figure 3C), opening up the opportunity to further manipulate PTM and PTN biosynthesis in these overproducers for pathway characterization and structural diversity.

Discussion

In spite of the indisputable track record of natural products as drugs and drug leads and small molecule probes, the “grind and find” approach of traditional natural product discovery, with its elements of serendipity, demands too much time, effort, and resources (Newman, D.J. & Cragg, G.M., *J. Nat. Prod.* 75, 311-335 (2012); Li, J.W.H. & Vederas, J.C. *Science* 325, 161-165 (2009)). It is reported herein, a high-throughput method using real-time PCR for strain prioritization to identify the most promising strains from a microbial strain collection for natural product discovery. Resources could then be devoted preferentially to the strains that hold the highest promise in producing novel natural products, thereby accelerating detection and isolation of the targeted natural products and cutting the time and cost associated with traditional natural product discovery programs. Applications of genome sequencing and genome mining to the high priority strains could essentially eliminate the chance elements from traditional natural product discovery programs and fundamentally change how microbial natural products are to be discovered (Figures 1A, 1B).

Central to the method is the application of real-time PCR, targeting genes characteristic to the biosynthetic machinery of natural products with distinct scaffolds, in a high throughput format and with superior sensitivity and specificity. The practicality and effectiveness of this method was showcased by targeting four diterpene synthase genes, *T1*, *T2*, *T3*, and *T4*, to prioritize 1,911 strains, selected from the actinomycete

strain collection at The Scripps Research Institute, for diterpenoid producers in general and PTM and PTN producers in particular (Figures 2A-2D). Diterpenoid producers were chosen because they are underrepresented among microbial natural products and for new PTM and PTN producers because of the heroic effort in the discovery of the original *S. platensis* MA7327 and MA7339 strains, as well as the need for alternative producers with better genetic amenability. A total of 488 potential diterpenoid producers were rapidly identified, among which six were confirmed as PTM-PTN dual producers and one as a viguiepinol and oxaloterpin producer. Although the new PTM-PTN dual producers are all *S. platensis* species on the basis of polyphasic taxonomy, they exhibit distinct morphology, three of which showed superior genetic amenability to the original *S. platensis* MA7327 and MA7339 strains. Genetic amenability of the producing organisms is of paramount importance in applying microbial genomics to natural product discovery, and new strains, alternatives to genetically recalcitrant original producers for the same natural products or families of natural products with similar scaffolds, represent an innovative solution to this critical challenge. The 25% (i.e., 488 out of 1911) hit rate of diterpenoid producers is higher than expected, given the small number of bacterial diterpenoids known to date, but is consistent with an early survey of this strain collection. These findings support that the biosynthetic potential of diterpenoids in bacterial is significantly underestimated, and further prioritization of the hits identified in this study promises the discovery of novel diterpenoid natural products.

Natural products occupy tremendous chemical structural space that is unmatched by any other small molecule libraries. While the rich functionality of natural products is, without doubt, one of their great strengths, providing potency and selectivity, the biosynthetic machineries for each of the major molecular scaffolds are highly conserved across the entire family of natural products, as exemplified by the diterpene synthases T1, T2, T3, and T4, for diterpenoids. Variations among the myriad of tailoring enzymes associated with each of the biosynthetic machineries further imbue the remarkable structural diversity within each of the natural product families (Figure 2A). Although the method has been showcased by targeting diterpenoids, variation of the method is readily applicable to the discovery of other classes of natural products by targeting conserved genes within each of the major biosynthetic machineries. Deviations

to the targeted T_m value would be correlated to the product specificity. The T_m range used in this study was 0.8 °C, and this range allowed for the identification of hits with 97-98%, 63-98%, 96-97%, and 38-98% identity to *ptmT1*, *ptmT2*, *ptmT3*, and *ptmT4*, respectively (Figures 5A-5D). The range of T_m can be exploited to address the needs of each
5 experiment, and a larger range of T_m would increase the likelihood of finding diverse hits. Finally, a DNA intercalating fluorophore was used in the current method, allowing detection and subsequent melting temperature analysis of the amplified products. One could envision variations of the method with fluorophore-labeled oligonucleotides or molecular beacons in a multiplex approach of real-time PCR to target multiple genes
10 simultaneously within a single reaction, further expanding its utility in strain prioritization, thereby accelerating natural product discovery.

Table 1. Titers of platensimycin and platencin produced by *Streptomyces platensis* spp.

Compound	Titer ^a (mg L ⁻¹) for the following <i>S. platensis</i> strain:											
	MA7327	SB12001 ^b	SB12002 ^b	CB00739	SB12026	CB00765	SB12027	CB00775	SB12028	CB00789	CB02289	CB02304
PTM	4.1 ± 1.0	110 ± 4	220 ± 9	1.0 ± 0.3	310 ± 12	5.2 ± 1.0	11 ± 0.9	4.7 ± 1.3	230 ± 20	6.5 ± 1.7	0.66 ± 0.17	10 ± 4
PTN	0.77 ± 0.21	150 ± 14	74 ± 3	1.6 ± 0.2	170 ± 6	2.4 ± 0.2	12 ± 1	0.65 ± 0.05	200 ± 25	2.2 ± 0.6	<0.1 ^c	11 ± 4

^aUnless otherwise noted, values are the averages of at least three independent trials and are reported with standard deviations.

^bStrains originally reported in reference 24.

^cThe PTN peak at λ_{240} was too small to calculate a reliable titer; PTN was detected by EIC (m/z at 426.20 for the [PTN + H]⁺ ion) and is shown in Figure 6C.

Table 2. Primers used in this study.

Primer	Nucleotide Sequence (5'-3')	Function (Reference)
ptmn1-S	CCGGGCTGGACATCCGGGCGGAC (SEQ ID NO: 1)	PCR targeting <i>ptm/ptnT1</i>
ptmn1-AS	GGATGGCGCAGAGCAGGAGGTC (SEQ ID NO: 2)	PCR targeting <i>ptm/ptnT1</i>
ptmn2-S	CTGCTCCCCGCCGCCACC (SEQ ID NO: 3)	PCR targeting <i>ptm/ptnT2</i>
ptmn2-AS	CGTAGTACGGCGAGGCGTGC (SEQ ID NO: 4)	PCR targeting <i>ptm/ptnT2</i>
ptm3-S	TATCTGCTCGACGGCAGGCTCGAC (SEQ ID NO: 5)	PCR targeting <i>ptmT3</i>
ptm3-AS	TTGGCCCAGGTCCGCAGATCGTT (SEQ ID NO: 6)	PCR targeting <i>ptmT3</i>
ptmnT4mix-S	CTGBTSCACGACGAYVTSATGGAC (SEQ ID NO: 7)	PCR targeting <i>ptm/ptnT4</i>
ptmnT4mix-AS	GCCSAKBABGTCGTCSTRYVWDYTGGAA (SEQ ID NO: 8)	PCR targeting <i>ptm/ptnT4</i>
16SrRNA_for	AGAGTTTGATCCTGGCTCAG (SEQ ID NO: 9)	Phylogenetic Analysis
16SrRNA_rev	ACGGCTACCTTGTACGACTT (SEQ ID NO: 10)	Phylogenetic Analysis
recAfor	TAATACGACTCACTATAGGGCCGCRCTCGCACARATTG AACG (SEQ ID NO: 11)	Phylogenetic Analysis
recArev	GCTAGTTATTGCTCAGCGGCGTCGGGGTTGTCCTTSAG GAAG (SEQ ID NO: 12)	Phylogenetic Analysis
rpoB-2	CATCGACCACTTCGGCAAC (SEQ ID NO: 13)	Phylogenetic Analysis
ActRpoB3303R	GAANCGCTGDCCRCCGAAGT (SEQ ID NO: 14)	Phylogenetic Analysis (1)
trpBfor	TAATACGACTCACTATAGGGGCGCGAGGACCTGAACC ACAC (SEQ ID NO: 15)	Phylogenetic Analysis
trpBrev	GCTAGTTATTGCTCAGCGGCATGGCCGGGATGATGCCC (SEQ ID NO: 16)	Phylogenetic Analysis
ptmT1for	GACATCGAGGGGCATGGGAAGG (SEQ ID NO: 17)	Sequencing
ptmT1rev	GTCAATCCGGAGACCGGGGTAC (SEQ ID NO: 18)	Sequencing
ptmT2for	CTTGTGGCGTCCAGGAAGGAG (SEQ ID NO: 19)	Sequencing
ptmT2rev	CCACGAACGACAACAACAGTTGCG (SEQ ID NO: 20)	Sequencing
ptmT2mid	GATGTGTGCGTTGGTGCTGG (SEQ ID NO: 21)	Sequencing
ptmT3for	CATCTGGCGCGACAACCGCATTG (SEQ ID NO: 22)	Sequencing
ptmT3rev	CATCGCGTCCTTTGATCGGGAGG (SEQ ID NO: 23)	Sequencing
ptmT4for2	CGGAACACCGCCGCGTAG (SEQ ID NO: 24)	Sequencing
ptmT4rev	GAGCACCACGTCGCCGTG (SEQ ID NO: 25)	Sequencing
ptmR1for	GTCTTCGGCAGCCGGCTCTC (SEQ ID NO: 26)	Sequencing
ptmR1rev	CGACTCAACAGGGCGTAAAGGTGC (SEQ ID NO: 27)	Sequencing
ptmRtgtF	CCCGCCGGAATCGGCCCTGATGGAGCAGTTCGGCATT CCATTCCGGGGATCCGTCGACC (SEQ ID NO: 28)	λ RED-mediated PCR targeting replacement of <i>ptmR1</i>
ptmRtgtR	TCGAGGAGTTCCAGACGGGTATTGGCGCCGCTCGCATT CAATGTAGGCTGGAGCTGCTTC (SEQ ID NO: 29)	λ RED-mediated PCR targeting replacement of <i>ptmR1</i>
ptmRidF	CCTGATGGAGCAGTTTCGG (SEQ ID NO: 30)	Δ <i>ptmR1</i> PCR confirmation
ptmRidR	GGAGTTCCAGACGGGTATTG (SEQ ID NO: 31)	Δ <i>ptmR1</i> PCR confirmation

Table 3. Strains and plasmids used in this study.

Strain/Plasmid	Genotype, Description	Source (Reference)
<i>E. coli</i> DH5α	<i>E. coli</i> host for cloning	Life Technologies
<i>E. coli</i> XL1-Blue MRF'	<i>E. coli</i> host for library construction	Agilent
<i>E. coli</i> ET12567/pUZ8002	Methylation-deficient <i>E. coli</i> host for intergeneric conjugation; contains pUZ8002, a nontransmissible <i>oriT</i> mobilizing plasmid	(2)
<i>E. coli</i> BW25113/pIJ790	<i>E. coli</i> host for PCR targeting	(3)
<i>S. platensis</i> MA7327	Wildtype PTM/PTN producer	Merck (4,5)
<i>S. platensis</i> MA7339	Wildtype PTN producer	Merck (6,7)
<i>S. platensis</i> SB12001	PTM/PTN overproducing strain	(8)
<i>S. platensis</i> SB12002	PTM/PTN overproducing strain	(8)
<i>S. platensis</i> SB12600	PTN overproducing strain	(9)
<i>S. platensis</i> CB00739	PTM-PTN strain hit	This study
<i>S. platensis</i> CB00765	PTM-PTN strain hit	This study
<i>S. platensis</i> CB00775	PTM-PTN strain hit	This study
<i>S. platensis</i> CB00789	PTM-PTN strain hit	This study
<i>S. platensis</i> CB02289	PTM-PTN strain hit	This study
<i>S. platensis</i> CB02304	PTM-PTN strain hit	This study
<i>S. platensis</i> SB12026	CB00739 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> cassette	This study
<i>S. platensis</i> SB12027	CB00765 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> cassette	This study
<i>S. platensis</i> SB12028	CB00775 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> cassette	This study
SuperCos1	Vector for the construction of cosmid libraries	Agilent
pIJ773	Plasmid containing the apramycin resistance cassette (<i>aac(3)IV+oriT</i>)	(3)
pBS12031	Cosmid 18H9 from CB00739 cosmid library, containing partial <i>ptm</i> gene cluster	This study
pBS12032	Cosmid 18H10 from CB00765 cosmid library, containing partial <i>ptm</i> gene cluster	This study
pBS12033	Cosmid 10F8 from CB00775 cosmid library, containing partial <i>ptm</i> gene cluster	This study
pBS12034	pBS12031 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> by PCR targeting	This study
pBS12035	pBS12032 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> by PCR targeting	This study
pBS12036	pBS12033 with <i>ptmR1</i> replaced with <i>aac(3)IV+oriT</i> by PCR targeting	This study

Examples 2: Strain Prioritization for Natural Product Discovery by A High-Throughput Real-Time PCR Method – Eneidyne Natural product Discovery

The following have been achieved herein: (i) completed the genome survey of 3,200 strains, identifying 89 hits as new enediyne producers, (ii) established the genetic amenability for 66 hits, (iii) constructed the $\Delta pksE$ mutants for 9 hits, (iv) detected heptaene production for 15 hits, (v) sequenced the genomes for 20 hits, revealing 20 enediyne gene clusters with distinct organization and architecture, and (v) confirmed one hit CB02366 as a new C-1027 producer (Figure 10). Taken together, these results: (i) validate the genome survey strategy to identify new enediyne producers, (ii) unveil Actinomycetales as the most prolific enediyne producers, and (iii) demonstrate the feasibility to mine the genomes of the Actinomycetale collection for the discovery, production, and isolation of new enediyne natural products.

Materials and Methods

Figure 11 depicts the overall research design and methods to mine genomes of the Actinomycetale collection for new enediyne natural products. Central to the approach are: (i) the recognition of Actinomycetales as the most prolific enediyne producers, (ii) the development of a genome survey method to rapidly identify the most promising enediyne producers from our Actinomycetale collection, and (iii) the combination of genomics, bioinformatics, metabolic pathway engineering, medium and fermentation optimization, and metabolomics to activate enediyne biosynthesis in the new producers, thereby production, isolation, and structural characterization, of the novel enediynes.

General methods for microbial natural product discovery. (i) While the “grind and find” paradigm of traditional natural product discovery program is laborious, it remains effective, variations of which still characterize new natural product discovery today. Microbial cultures (wild-type and engineered strains) are typically grown in 250-mL flasks (containing 50 mL culture) in a shaker for analytical scale production and could be scaled up in fermentors (10 L in a 14-L vessel or 30 L in a 40-L vessel) for preparation scale. It was found that metabolite profiles are readily amenable to HPLC and

LC-MS analyses by a metabolomics approach, in particular if knockout mutant strains for the targeted natural products are available as negative controls. The inventor has an inventory of >40 different fermentation media, collected from both pharmaceutical industry and academic labs all over the world, to support natural product production in Actinomycetales. As summarized in Figure 9G and Figure 10, by fermenting the hits in five different media, the production of heptaene has been realized, and by association the novel enediyne natural products, for 15 of the 66 new producers examined, as exemplified by the discovery and confirmation of CB02366 as a new C-1027 producer.

(ii) Complementary to traditional natural product discovery, connecting natural products to the genes that encode their biosynthesis has fundamentally changed the landscape of natural products research and sparked the emergence of a suite of contemporary approaches to natural product discovery. Critical to these contemporary approaches is to have a genetic system in place for the producing organisms so that the biosynthetic machinery of the targeted natural products can be manipulated *in vivo*. Four general methods are available for introduction of plasmid DNAs into *Streptomyces* and closely related species: protoplast transformation, electroporation, conjugation, and phage transduction. The inventor has extensive experience in the application of all four methods and has developed genetic systems for more than 30 antibiotic-producing *Streptomyces* species over the years, as exemplified by the engineered production of novel C-1027 analogues in *S. globisporus* and generation of the $\Delta ucmE$ mutant in *S. sp.* for fermentation optimization and UCM detection by a metabolomics approach (Figure 9J), as well as development of genetic systems for the 69 new enediyne producers and generation of the $\Delta pksE$ mutants for nine of them in preliminary studies (Figure 9G and Figure 10). The fact that most of the new enediyne producers identified in this application are *Streptomyces* ensures that the expedient recombinant DNA technologies and genetic tools in *Streptomyces* can be readily adopted. (iii) Complementary to manipulating natural product biosynthetic gene clusters in their native producers, expression of biosynthetic gene clusters in heterologous hosts for natural product production has emerged as an attractive alternative. The inventor has extensive experience and has made several important contributions to heterologous expression of complex natural product biosynthetic pathways in selected model *Streptomyces* hosts,

producing some of the most complex natural products known to the field, including fredericamycin, iso-migrastatin, platensimycin and platencin, and bleomycin.

(iv) The new enediyne natural products can be isolated by common chromatographic techniques, and their structures determined by standard spectroscopic methods, such as MS, 1D and 2D ¹H and ¹³C NMR (COSY, TOCSY, HMQC, and HMBC).

Strain prioritization for enediyne natural product discovery by a real-time PCR method. (i) The Actinomycetale strain collection at The Scripps Research Institute consists of strains collected from various unexplored and underexplored ecological niches. (ii) A genome survey of 3,200 strains was completed, identifying 89 hits. Thus, genomic DNAs for each of the strains were prepared according to standard procedures (Kieser, T. et al., Practical Streptomyces genetics, The John Innes Foundation: Norwich, UK, 2000), arrayed in 384-well plates, and their concentrations normalized (0.1 - 1 ng/well). Real-time PCR were carried out, in a 384-well plate format, with either the E/E5 pair ((5'- CCCC GCVCACATCACSGSCCTCGCSGTGAACATGCT-3' (SEQ ID NO: 33)/5'- GCAGGCKCCGTC SACS GTGTABCCGCCGCC-3' (SEQ ID NO: 34)) or E10/E pair ((5'-TGYGYSCCSSACSSSVSGSTGCTGCC-3' (SEQ ID NO: 35)/5'- ACGTTGCCGACSAGRTTSGTYTCCTCGAACCGAC-3' (SEQ ID NO: 36)) of primers. Specific PCR products were identified by melting curve analysis, each of the hits was confirmed by gel electrophoresis, and strains affording a specific PCR product with the E/E5, E/E10, or both primers were considered as potential enediyne producers. The PCR-amplified E/E5 and E/E10 fragments, as well as, an additional PCR amplified internal fragment of the *E* gene, with primer pair of 5'- TGTA AACGACGGCCAGTATGGGSTTCGGCGGSATCAAC-3' (SEQ ID NO: 37)/5'-CAGGAAACAGCTATGACCAGMGGNGAGTGGAANGCGTG-3' (SEQ ID NO: 38), were sequenced. Those strains that harbor enediyne PKS gene cassettes that are distinct to all enediynes known to date were identified as hits.

(iii) The 89 hits were further prioritized, on the basis of enediyne PKS gene cassette novelty and taxonomic distinctness of the producers, and the most promising 20 hits were identified for genome sequencing (Figure 10). The multilocus

sequence analysis approach was adopted, on the basis of selected house-keeping genes (*16S rRNA*, *recA*, *trpB*, *rpoB*), to characterize the taxonomy of the hits. Diverse representations from each of the individual phylogenetic clades were considered as one criterion to select strains for genome sequencing. It was shown that phylogenetic analysis of the enediyne PKS gene cassettes, on the basis of the internal fragment of *E*, *E5*, *E10*, or a combination of them, afforded distinct clades, each of which most likely encode enediynes with distinct structures (Figure 10). While the 11 known enediyne PKS gene cassettes are well dispersed, there are clades to which none of the known enediyne PKS gene cassettes fall, a provocative suggestion for novel enediynes. Phylogenetic representation of the enediyne PKS gene cassettes was considered as another criterion.

Genome sequencing to confirm the prioritized hits as new enediyne producers. The most unique 20 hits, selected on the basis of the above two criteria, have been subjected to genome sequencing. For genome sequencing, the mate-pair library protocols (with 5-6 kb inserts) have been optimized which greatly facilitated *de novo* genome assembly. Genome sequencing of the selected 20 hits has been completed, affording draft genomes, consisting on average 15-30 scaffolds that can be readily mined for natural product biosynthetic gene clusters (Figure 10). Various programs for automatic annotation of secondary metabolite biosynthetic gene clusters are known, each of which has different strengths and limitations. Given the high sequence homology and organizational conservation of the enediyne PKS gene cassettes, it was found that the anti-SMASH program is sufficient and most useful for rapid identification of new enediyne gene clusters (Medema, M.H. et al., *Nucleic Acids Res.* 2011, 39, W339-W346). Follow-up manual annotation corrects errors generated by the anti-SMASH program and affords accurate and complete annotation of the new enediyne clusters. From the 20 sequenced genomes, 19 enediyne clusters have been identified that are distinct to all enediyne clusters known to date, hence novel enediyne natural products, and confirmed one of them, CB02366, as a new C-1027 producer.

Fermentation optimization of the hit strains for new enediyne natural product production. (i) The hit strains were first subjected to medium and fermentation optimization, taking advantage of the extensive medium (>40) collection and experience

in culturing rare Actinomycetales and optimizing their fermentation for secondary metabolite, as well as, enediyne production. (ii) The regulation of the enediyne biosynthetic machinery was by inactivating repressors, overexpressing activators, or both. Various regulators have been identified within the enediyne gene clusters (Figure 9A) and their roles in regulating enediyne biosynthesis and improving enediyne production have been demonstrated. (iii) Functional expression of the enediyne biosynthetic gene clusters were followed at transcriptional level by RT-PCR following the transcription of enediyne PKS gene cassette, as exemplified by *pksE* in Figure 9H, as well as, other genes within the clusters in various natural product biosynthetic machineries (Yang, D. et al. *Appl. Microbiol. Biotechnol.* 2011, 89, 1709-1719; Smanski, M.J. et al. *J. Nat. Prod.* 2012, 75, 2158-2167). (iv) Heptaene production has been proven to be a sensitive phenotypic indicator for enediyne biosynthesis, and its utility was demonstrated following functional expression of the enediyne biosynthesis gene clusters (Figure 9F). Heptaene production was followed by HPLC analysis, to monitor the activation of enediyne biosynthetic gene clusters and optimization of enediyne production in both the native producers and their engineered recombinant variants (Figure 9G). (v) While the 11 known enediynes all feature a 9- or 10-membered enediyne core it is far from certain that the new endiynes will have to feature the same scaffolds, in particular those encoded by the enediyne PKS gene cassettes that fall into distinct phylogenetic clades (Figure 10). Unbiased comparative analysis of all metabolites in fermentation will be carried out. The metabolomics platform is ideally suited to help rapidly detect the new enediynes and optimize fermentations for their production as exemplified in Figures 9G and 9J. The genetic amenability for 66 of the 89 hits was confirmed and $\Delta pksE$ mutants were constructed for 9 of them (Figure 10). (vi) To ensure the discovery of the most potent enediynes, preliminary cytotoxicity assays of the new enediynes were conducted before committing large scale fermentation for their production, isolation, and structural elucidation. Only the most potent, promising new enediyne natural products were isolated and their structures determined by standard chromatographic and spectroscopic methods. (v) Fermentations were scaled up to produce and isolate sufficient quantities of the new enediynes for evaluation as anticancer ADC payload candidates, and the fermentation facility at TSRI should greatly facilitate these efforts. The regulatory

network of the new enediyne biosynthetic machinery was manipulated to improve their production if necessary. (vi) The cytotoxicity of the new enediynes were evaluated against selected cancer cell lines in house, and selected enediynes were also be sent to NCI for comprehensive evaluation in the NCI-60 cell line panel.

5 *Results and Discussion*

Enediynes as anticancer drugs and payloads in ADCs. The enediyne natural products are the most cytotoxic molecules in existence today, and their use as anticancer drugs has been demonstrated clinically. Although the natural enediynes have seen limited use as clinical drugs mainly because of substantial toxicity, various polymer-based delivery systems or ADCs have shown great clinical success or promise in anticancer therapy. Indeed, the poly(styrene-co-maleic acid)-conjugated NCS (SMANCS[®]) has been marketed since 1994 for use against hepatoma. Various ADCs have been successfully developed, including a CD33 mAB-CAL conjugate (i.e., MYLOTARG[®]) for acute myeloid leukemia (AML), a CD22 mAB-CAL conjugate (inotuzumab ozogamicin) for non-Hodgkin lymphoma, as well as, several mAB-C-1027 conjugates for hepatoma and mAB-UCM conjugates for selected tumors. (Pfizer voluntarily withdrew MYLOTARG[®] from the market in 2010, however, significant survival benefits observed in recent phase III trials suggest that MYLOTARG[®] may have an important future role in treating patients with good- or intermediate-risk AML). These examples clearly demonstrate that the enediynes can be developed into powerful drugs when their extremely potent cytotoxicity is harnessed and delivered to tumor cells. It is remarkable that among the 11 enediynes known to date, two (NCS and CAL) have been developed into clinic drugs and one (C-1027) is in clinical trial, representing an astonishing ~30% success rate with the enediyne class of natural products. Thus, developing innovative methods to discover new enediynes holds a great promise for anticancer drug discovery.

Interests in anticancer ADCs continue to grow. Inspired by the recent success of two ADCs (ADCETRIS[®] in 2011 and KADCYLA[®] in 2013), virtually every major pharmaceutical company with an oncology program now has an initiative on ADCs. Among the 30+ ADCs currently in development, however, the majority of them

(>90%) use one of the four available cytotoxic drugs, all of which are of natural product origin. The ADC field is therefore in critical need of new, highly potent cytotoxic payloads (1 nM to 10 pM), active in many tumor types, with improved physical, chemical, and biological properties. Thus, new enediynes that can be readily produced
5 by microbial fermentation would be extremely valuable assets in the development of safer, more effective ADCs with a validated mode of action.

The enediynes as ideal payload candidates for anticancer ADCs. ADCs provide the possibility of selectively ablating cancer cells by combining the specificity of a mAB for a target antigen with the delivery of a highly potent cytotoxic agent. The ideal
10 number of drug molecules per mAB for most current ADCs appears to be about four. Underconjugation can decrease potency of the resultant ADCs, whereas the conjugation of too many drugs per mAB can lead to decreased circulation half-life, reduced tolerability, and impaired antigen binding. The preferred payload molecules therefore have to be highly cytotoxic (IC₅₀s in the range of 1 nM to 10 pM) and ideally active in
15 many tumor types. The enediynes represent some of the most cytotoxic molecules in existence today (for example, the IC₅₀s of CAL and C-1027 towards selected cancer cell lines are in the range of 10 pM to 10⁻³ pM). While the enediynes are most known for their activity by DNA DSBs, ICL was discovered by the inventors, as an alternative mode of action for the enediyne family of anticancer agents and engineered C-1027 analogues
20 capable of DNA DSBs, ICLs, or both (Kennedy, D.R. et al., *Cancer Res.* 2007, 67, 773-781; Kennedy, D.R. et al., *Proc. Natl. Acad. Sci. USA* 2007, 104, 17632-17637). It was further demonstrated that the ICL property of the enediynes can be exploited to target solid tumors or other cancer cells under hypoxic environments, which do not respond well to enediynes that predominantly induce oxygen dependent DSBs. The exquisite
25 potency and mechanisms of action of the enediynes, as well as the demonstrated feasibility to engineer designer enediynes with altered activities (i.e., DSBs, ICLs, or both), make them ideal payload candidates for ADCs. However, only 11 enediynes are known to date, and most of them are produced in trace quantities, intrinsically unstable, produced by rare actinomycetes that are refractory to all means of genetic manipulations
30 for either titer improvement or analogue generation, or simply not available in sufficient quantities for a full evaluation as ADC payload candidates. New enediynes with varying

mechanisms and potency, functional groups for linkage, solubility to enable the reaction with antibodies, prolonged stability in formulation, that can be reliably produced in sufficient quantities by microbial fermentation of genetically amenable *Streptomyces* species, therefore will provide outstanding opportunities to address issues of potency, mechanism, permeability, tractability, stability, and efflux, critical challenges encountered with the known enediynes and other ADC payloads currently in development.

Enediyne biosynthesis and engineering. The enediyne natural products present an outstanding opportunity to (i) decipher the genetic and biochemical basis for the biosynthesis of complex natural products, (ii) explore ways to make novel analogues by manipulating genes governing their biosynthesis, and (iii) discover new enediyne natural products by mining microbial genomes for the trademark enediyne biosynthetic machineries. Significant progress has been made in enediyne biosynthesis and engineering in the last decade. Highlights include: (i) cloning and characterization of the C-1027, NCS, MDP, KED, SPO, CYA and CYN biosynthetic gene clusters as examples of nine-membered enediyne natural products, (ii) cloning and characterization of the CAL, ESP (partial), and DYN biosynthetic gene clusters as examples of 10-membered enediyne natural products, (iii) establishment of the polyketide origin for all enediynes and discovery of the enediyne PKS as an acyl carrier protein (ACP)-dependent, self-phosphopantetheinylating, iteratively acting, type I polyketide synthase (PKS) that produces a linear polyene to initiate both nine- and 10-membered enediyne core biosynthesis, (iv) characterization of numerous enzymes from the enediyne biosynthetic machineries that catalyze novel chemistries, (v) engineered production of novel enediyne analogues with altered modes of action, and exploitation of the ICL property of the enediynes to target solid tumors or other cancer cells under hypoxic environments, (vi) manipulation of the enediyne biosynthetic machinery for titer improvement and production of selected enediynes in sufficient quantities to support further mechanistic studies and preclinical development, (vii) development of methodologies to screen microorganisms for the discovery of new enediyne natural products, and (viii) discovery that, in spite of the fact that only 11 enediyne natural products (14 if including the ones isolated in the cycloaromatized form) have been isolated to date, the biosynthetic

potential of Actinomycetales to produce enediynes is greatly underappreciated. These findings have laid the foundation to explore microbial genomics for the discovery of new enediyne natural products. Thus, (i) genome survey of microbial strain collections for the hallmark enediyne PKS gene cassettes allows for rapid identification of potential enediyne producers, (ii) genome sequencing of the potential producers for the enediyne biosynthetic gene clusters could allow accurate prediction of the structural novelty of the new enediynes, and (iii) genetic manipulation and fermentation optimization of the most promising producers would ultimately allow efficient production, isolation, and structural and biological characterization of new enediyne natural products.

10 *Strain prioritization for novel natural product discovery.* Traditional microbial natural product discovery programs start from fermenting each strain individually, often in multiple media, followed by preparation of crude extracts. There are two primary approaches to search for novel natural products from extracts: bioassay-guided fractionation and chemical profiling of agents possessing unique structural features and/or novelty as important representatives of different chemical classes. In both cases, a molecule of interest must be produced in sufficient amounts in order to permit isolation, purification, and characterization on a reasonable timeframe. The ultimate success in discovering a new natural product typically requires three principal steps: dereplication of known compounds to avoid duplication of effort, isolation and purification of the targeted molecules from a highly complex matrix, and structural elucidation of the purified natural product. This traditional sequence of steps still characterizes new natural product discovery from crude extracts today. While successful, it is a tedious and laborious process. As exemplified by the discovery of platensimycin from *Streptomyces platensis* MA7327 in 2006 and platencin from *Streptomyces platensis* MA7339 in 2007, Merck screened 250,000 crude extracts made from 83,000 strains, each of which was fermented in three different media. The process could be significantly shortened and more cost effective should Merck know in advance the biosynthetic potential of the strains in its collection so that resources could be devoted preferentially to interrogate only the strains that hold the highest promise in producing novel natural products.

Complementary to traditional approaches, the progress made in the last two decades in connecting natural products to the genes that encode their biosynthesis has fundamentally changed the landscape of natural products research and sparked the emergence of a suite of contemporary approaches to natural product discovery. Thus, genes have become as important as chemistry in categorizing known natural products and identifying likely new ones yet to be discovered. Advances in microbial genomics have unequivocally demonstrated that ~90% of the natural product biosynthetic capacity is missing, even from the workhorse producers, the Actinobacteria. To gain access to this untapped reservoir of potentially new natural products, two principal strategies have been applied to induce these “cryptic biosynthetic pathways”. The so-called ‘epigenetic’-related approaches include challenging the microorganisms through culture conditions, nutritional or environmental factors, external cues, and stress, as well as, exploiting interspecies crosstalk. The genomics-based approaches include mining the genomes to predict metabolite structures, engineering the pathways by manipulating global and/or pathway-specific regulators, and expressing the cryptic pathways in selected heterologous hosts. While each of the various approaches has different strengths and weaknesses, they have been successful in yielding cryptic natural products but only on a case-by-case basis and are far from being of practical use for natural product discovery. Thus, in spite of the rapid advances in DNA sequencing technologies and bioinformatics, it is still unlikely to sequence all strains within a collection as a practical means to discover new natural products. Clearly, a high throughput method is needed to rapidly survey the biosynthetic potential of our strain collection. Strains that harbor the highest biosynthetic potential can then be identified, prioritized, and subjected to epigenetic- and/or genomics-based approaches to induce all cryptic biosynthetic pathways for novel natural product discovery.

An innovative genome survey strategy was developed herein, that can be applied to rapidly identify strains, from the Actinomycetale collection, that are of high likelihood to produce enediyne natural products. Remarkably, a virtual survey of the bacterial genomes currently available in the public databases revealed that the Actinomycetales are the most prolific enediyne producers. A genome survey of 3,200 strains from the Actinomycetale collection was conducted, identifying 89 potential

enediyne producers (hits). It was demonstrated, by genome sequencing, that these hits are true enediyne producers, containing gene clusters that are distinct to all enediyne clusters known to date and, by genetic manipulation and fermentation optimization, that the most promising hits can be activated to produce the new enediyne natural products.

5 *Enediyne PKS gene cassettes as a beacon for enediyne producers.* In spite of their remarkable structural diversity, the 14 known enediyne natural products all feature a nine- or 10-membered enediyne core. Comparative bioinformatics analyses of the seven nine- (NCS, C-1027, MDP, KED, SPO, CYA, and CYN) and three 10-membered (CAL, DYN, and ESP) enediyne PKS loci revealed a set of five genes
10 common to all enediynes (i.e., the enediyne PKS gene cassette consisting of *E10/E/E5/E4/E3*); no apparent conservation was observed beyond the enediyne PKS gene cassettes, accounting for the structural diversity characteristic for the periphery moieties of the enediynes. This remarkable sequence homology has resulted in a unified model for the enediyne PKS cassette to catalyze the formation of both nine- and 10-membered
15 enediyne cores. This model has been further supported by the fact that: (i) PKS-TE (i.e., E–E10) catalyze the biosynthesis of heptaene as a shunt metabolite in the absence of the associated enzymes, (ii) heptaene is co-produced in fermentations of known enediyne producers, and (iii) heptaene production is PKS-dependent—inactivation of the *pksE* gene abolished heptaene production and complementation of the $\Delta pksE$ mutation with a
20 functional *pksE* restored its production. These observations prompted the selection of genes within the enediyne PKS cassettes as probes to survey genomes for the presence of enediyne biosynthetic machinery. Once the new enediyne producers are identified, heptaene production, complementary to the biochemical induction assay (BIA) for enediyne production, could be used as a sensitive phenotypic indicator to follow enediyne
25 production upon fermentation optimization.

Genetic manipulation of Actinomycetales to activate enediyne biosynthesis and production. There are minimally four requirements for implementing metabolic pathway engineering strategies to natural product discovery and structural diversity. These are: (i) the gene clusters encoding the production of a particular natural product or
30 family of natural products, (ii) genetic and biochemical characterizations of the biosynthetic machinery for the targeted natural products to a degree that combinatorial

biosynthesis principles can be rationally applied to engineer the novel analogues, (iii) expedient genetic systems for *in vivo* manipulation of genes governing the production of the target molecules in either native producers or heterologous hosts, and (iv) production of the natural products or engineered analogues to levels that are sufficient for isolation and structural and biological characterization. Although each of these requirements is essential, establishing an expedient genetic system for *in vivo* manipulation of the biosynthetic machinery of the targeted metabolites is importance (Galm U., Shen, B. *Exp. Opinion Drug Dis.* 2006, 1, 409-437; Van Lanen S.G., Shen, B. *Drug Disc. Today: Technologies* 2006, 3, 285-292; Van Lanen, S.G.; Shen, B. *Curr. Opinion Drug Discov. Develop.* 2008, 11, 186-195). Thus, a decision and opportunity for innovation in manipulating enediyne biosynthesis is the selection of the producers that are compatible with the expedient technologies and tools of recombinant DNA work in *Streptomyces* species and related organisms that have been developed in the past two decades. The CAL, DYN, and ESP (partial) clusters were cloned from *M. echinospora*, *M. chersina*, and *A. verrucosospora*, respectively, and genetic manipulations in *Micromonospora* and *Actinomadura* are known to be notoriously difficult. As a result, the ESP cluster is incomplete, and the boundaries of both the CAL and DYN clusters have yet to be determined experimentally. In contrast, biosynthesis and engineering of C-1027, NCS, and UCM have been greatly facilitated by the expedient genetic systems in *S. globisporus*, *S. carzinostaticus*, and *S. sp.*, respectively. Accordingly, *Streptomyces* in the Actinomycetale strain collection was selected, and this selection is vital to overcoming the current challenges of, and meeting future objectives for, enediyne discovery, biosynthesis, and engineering in their native producers.

The availability of four 9-membered (C-1027, NCS, MDP, KED) and four 10-membered [CAL, ESP (partial), DYN, and UCM] enediyne gene clusters, as well as the three additional clusters encoding the biosynthesis of the cycloaromatized enediyne natural products of sporolides and cyanosporasides, and the commanding role the inventor's laboratory leads in advancing enediyne biosynthesis and engineering provided a unique opportunity to investigate the molecular basis of enediyne biosynthesis by a comparative genomics approach. Thus, by comparing the gene clusters between the 9- and 10-membered enediynes, a unified model was formulated for the enediyne PKS

cassette to catalyze the formation of both 9- and 10-membered enediyne cores (Zhang, J. et al. *Proc. Natl. Acad. Sci. USA* 2008, 105, 1460-1465; Horsman G.P., et al. *Proc. Natl. Acad. Sci. USA* 2010, 107, 11331-11335), on which the current genome survey strategy for enediyne discovery was developed. By comparing metabolite profiles of the enediyne native producers, selected mutant strains, and recombinant strains expressing selected genes within the enediyne PKS cassette, a metabolomics method was established to follow the biosynthesis of heptaene as a sensitive phenotypic indicator for enediyne production. By manipulating the regulatory genes within the C-1027 biosynthetic gene cluster, C-1027 production was significantly improved. Application of the comparative genomics approach to analyze the new enediyne clusters in this application promises to reveal equally informative insights into their structures, biosynthesis, and regulations. The fact that most of the new enediyne producers discovered are of *Streptomyces* origin ensures that the extensive genetic tools available in *Streptomyces* can be readily applied to facilitate the discovery and production of the new enediynes in the native producers.

Preparation of complex natural products such as the enediynes and their analogues by total synthesis poses a monumental challenge to synthetic chemists. Combinatorial biosynthesis offers an excellent alternative to produce natural products and their analogues biosynthetically. The expedient tools for recombinant DNA work in *Streptomyces* and related microorganisms developed in the past two decades have made it possible to apply genetic principles to meet the drug discovery and development challenges in these organisms. Target metabolites can be produced by recombinant organisms that are amenable to large-scale fermentation, thus lowering production costs. Application of combinatorial biosynthesis strategies to the biosynthetic machineries in the new enediyne producers from our Actinomycetale collection promises to enable the discovery of novel enediynes and production of novel enediyne analogues.

Actinomycetales as the most prolific enediyne producers. To validate the utility of the selected genes within the enediyne PKS cassette as probes, a virtual survey of the entire GenBank was carried out, using each of the five genes within the enediyne cassette, alone or in combination, as queries, for genes encoding enediyne biosynthetic machineries. Several important lessons were learned from these surveys. (i) All 11 confirmed enediyne biosynthetic machineries were identified (Figure 9A), validating the

utility and specificity of the genes with the enediyne PKS cassette as probes. (ii) While each of the five genes alone yields essentially the same outputs, *E5*, *E*, or *E10* were preferred, and the combination of *E5/E* or *E/E10* afforded the best results. (iii) Together with the 11 known enediyne biosynthetic machineries, 55 additional enediyne PKS cassettes were also identified from organisms not known as enediyne producers, consistent with the early findings that the biosynthetic potential of enediynes is significantly underappreciated (i.e., a total of 66 enediyne biosynthetic loci from the publicly accessible GenBank database as of February 24, 2014). (iv) All of the 66 loci are of bacterial origin, and most remarkably, 55 of the 66 loci are in the order of Actinomycetales, revealing the Actinomycetales as the most prolific enediyne producers.

A high-throughput method to survey Actinomycetale genomes for novel enediyne producers. Inspired by the accuracy and specificity observed in the virtual screening, a high-throughput method was developed to survey the genes encoding the enediyne PKS cassettes and applied it to the Actinomycetale collection to identify potential new enediyne producers. (i) Close examination of the enediyne PKS gene cassettes showed that, while the five genes are absolutely conserved among 10 of the 11 known enediynes (the ESP cluster is incomplete hence cannot be included for comparison), there are subtle variations in their relative organization – (a) *E5/E/E10* all clustered (as in NCS, MDP, SPO, CYA, CAL, DYN, UCM, ESP), (b) *E5/E* clustered but *E10* separated (as in C-1027), or (c) *E/E10* clustered but *E5* separated (as in KED, CYN) (Figure 9A). Two sets of PCR primers were designed, specifically targeting *E5/E* or *E/E10*, respectively (Figure 9B). (ii) The feasibility to amplify the enediyne PKS gene cassettes by PCR was shown but these early experiments were low throughput, requiring analysis of each of the PCR products by gel electrophoresis. To develop a high-throughput method, genomic DNAs were prepared for each of the strains in the collection, normalized their concentrations, and arrayed the DNAs into a 384-well plate format. Real-time PCR was chosen, in a 384-well plate format, where specific PCR products were rapidly identified by melting curve analysis (Figure 9C). The putative hits were then confirmed by gel electrophoresis (Figure 9D), and the identity of hits as the targeted enediyne PKS gene cassettes was finally established by DNA sequencing.

Genome survey of our Actinomycetale collection to identify novel enediyne producers. (i) A genome survey of 3,200 strains has now been completed at the Scripps Research Institute's (TSRI) Actinomycetale collection and 89 novel enediyne producers (Figures 9B, 9C, 9D and Figure 10) were identified. (ii) The two sets of PCR primers, specifically targeting *E5/E* or *E/E10*, (Figure 9B), are complementary. Hits that were identified by both sets of the primers featured the enediyne PKS gene cassettes with *E5/E/E10* clustered together, while hits that can only be identified by one of the two sets of primers featured enediyne PKS gene cassette with either *E5* or *E10* separated from the *E* gene. (iii) The identity of the enediyne PKS gene cassettes from the 89 newly identified enediyne producers have been confirmed by DNA sequencing of *E5*, *E10*, and a 1-kb internal fragment of *E*. They differ from all the enediyne PKS gene cassettes currently known (Figure 10). (iv) Genome sequencing for 20 of the 89 hits at the Next Generation Sequencing (NGS) Core Facility, TSRI have been completed, confirming that they each contain an enediyne biosynthetic gene cluster and therefore are true enediyne producers. Among the 20 clusters, 19 are distinct to all enediyne clusters known to date, indicative of novel structural features/functional groups for these new enediynes, and one (CB02366) has the same genetic organization as the C-1027 cluster from *S. globisporus*; CB02366 was subsequently confirmed as a new C-1027 producer. Interestingly, CB02366 would be classified as an *S. griseus* species on the basis of selected house-keeping genes (*16S rRNA*, *recA*, *trpB*, *rpoB*), but the original C-1027 producer is known as *S. globisporus*, although the two C-1027 clusters are highly homologous (~90%/~94% identity at DNA/amino acid sequences). The finding of near identical gene clusters from taxonomically distinct species supports the strategy herein, in constructing the TSRI Actinomycetale collection, which has taken geographic and ecological, in addition to taxonomical, diversities into consideration. (v) The 19 new enediyne gene clusters are rich in open reading frames that are unprecedented not only in gene clusters that encode production of the known enediynes but also in gene clusters that encode production of other natural products, promising the discovery of new chemistries and biosynthetic knowledge for natural products in general.

Genetic manipulation and fermentation optimization for enediyne production. The sequence homology and organizational conservation of the enediyne

PKS gene cassettes also provided the inspiration to formulate a unified mechanism for enediyne biosynthesis. (i) It was confirmed that both 9- and 10-membered PKS-thioesterase (TE) (i.e., E-E10, Figure 9A) produce a common polyene intermediate to initiate the enediyne core biosynthesis, and in the absence of the associate enzymes, PKS-TE catalyze the biosynthesis of heptaene as a shunt metabolite (Figure 9E). (ii) Heptaene was discovered in fermentations of known enediyne producers and established that heptaene production was PKS-dependent – inactivation of PKS, as in the *ΔsgcE* mutant strain, abolished heptaene production and complementation of the *ΔsgcE* mutation with a functional *sgcE* restored its production (Figure 9F). These observations allowed for following heptaene as a sensitive phenotypic indicator for enediyne production. Preliminary screening, by HPLC analysis, of the fermentations of 66 of the 89 hits, each in five different media, indeed revealed that at least 15 produced heptaene (Figure 9G and Figure 10). (iii) Correlation between heptaene production and functional expression of the enediyne PKS gene cassette was confirmed by RT-PCR transcriptional analysis (Figure 9H). These findings evidence that the enediyne biosynthetic machinery in these 15 producers are most likely functional under the conditions examined, and further fermentation optimization of the other hits could be readily followed by RT-PCR for the expression of the *E* gene or other genes within the enediyne gene cluster and by HPLC analysis for heptaene production.

It was also shown that enediyne production can be improved by manipulating pathway regulation, and enediyne overproduction is often accompanied by the overproduction of heptaene, a finding that further supports heptaene as a phenotypic indicator for enediyne production. Thus, inactivation of a repressor and overexpression of an activator afforded an engineered strain *S. globisporus* SB1023 that overproduces heptaene and C-1027 in titers that are 24-fold and 7-fold higher, respectively, than those from the *S. globisporus* wild-type (Figure 9I). Titer improvement by manipulating the enediyne biosynthetic machinery in these new producers should greatly facilitate their production, isolation, and structural elucidation. (v) Taxonomic analysis of 55 of the 89 hits, on the basis of selected house-keeping genes (16S rRNA, *recA*, *trpB*, *rpoB*), classified 50 (91%) of them as *Streptomyces*. The fact that most of the new enediyne producers identified are *Streptomyces* ensures that the expedient recombinant DNA

technologies and genetic tools available for *Streptomyces* can be readily adopted by the current project. In preliminary experiments with 74 of the 89 new enediyne producers, the genetic amenability for 66 (89%) strains were confirmed and constructed the $\Delta pksE$ mutants for nine of them (Figure 10). (vi) With the ability to generate enediyne knockout mutant strains (i.e., $\Delta pksE$ mutants) as negative controls, it also demonstrated a metabolomics approach, by comparison of the metabolite profiles between the enediyne producing wild-type and enediyne knockout mutant strains, to rapidly optimize fermentation conditions to produce the new enediynes. As exemplified in Figure 9J, by comparison of the metabolite profiles between the *S. sp.* wild-type and $\Delta ucmE$ mutant strains, three metabolites were quickly correlated to the UCM biosynthetic machinery, one of which, produced in trace quantity (<20 $\mu\text{g/L}$), was subsequently confirmed to be UCM by bioassay against *M. luteus* and high resolution mass spectrometry analysis. Again, the accessibility to the state-of-the-art recombinant DNA technologies and genetic tools available in *Streptomyces* will greatly facilitate the effort to generate the $\Delta pksE$ mutants in the new producers for enediyne discovery by the metabolomics approach. (vii) Finally, production was confirmed and C-1027 was isolated from CB02366. The discovery of CB02366 as a new C-1027 producer and fermentation optimization of CB02366, leading to the eventual production, isolation, and characterization of C-1027, served as an ultimate demonstration that follow up studies on the other hits will result in the discovery, production, and characterization of new enediynes.

Given the extreme potencies of the enediyne family of natural products, their proven utility as cytotoxic payloads of anticancer ADCs, and the astonishing success rate of these rare natural products in the clinic, the discovery of new enediynes, as well as alternative producers of known enediynes with high titers or better growth characteristics, is a critical step in the development of enediynes as novel anticancer drugs. Thorough bioinformatics analysis of microbial genomes reveals that enediyne natural product gene clusters are widely distributed among Actinomycetales (50 out of 796 sequenced bacterial genomes, accounting for 6.3%), and yet remain underexplored, as the newly identified gene clusters are non-redundant and highly unique. The abundant and highly diverse enediyne gene clusters found within public databases should provoke

the interests of scientific communities in pursuing novel enediynes natural products from microbial strain collections, in particular from Actinomycetales.

Combining traditional natural product discovery methods with genomics-based technologies and bioinformatics will expedite the discovery process of novel enediynes. High-throughput methods to rapidly screen a large microbial strain collection for the conserved genes within the enediyne PKS cassettes are essential to allocate resources to strains that show biosynthetic potential for enediyne production. DNA sequencing of the conserved *pksE* genes or alternatively, the whole genome, allows further prioritization by identifying redundant or unique enediyne biosynthetic gene clusters through bioinformatics analysis (i.e., phylogenetic analysis, cluster annotation, genome neighborhood network). Once strains are prioritized, fermentation in multiple media and preparation of crude extracts allow chemical profiling of heptaene production, and bioassays for DNA damage activity (i.e., by BIA) can then be utilized to determine which strains readily produce enediynes or enediyne-like metabolites. Fractionation of promising crude extracts guided by the BIA allows a simple detection method for isolating novel enediynes. With the conserved enediyne PKS gene cassettes known, $\Delta pksE$ deletion mutants can be easily constructed to correlate putative gene clusters with the production of heptaene or DNA-damaging natural products. Silent or cryptic gene clusters can be overcome by the genetic manipulation of positive or negative regulators found within the enediyne gene clusters or by extensive fermentation optimization. The expedient technologies and tools of recombinant DNA work in *Streptomyces* species and related actinomycetes that have been developed in the past two decades will facilitate these efforts.

It is evident from the rarity of known enediyne natural products that future enediyne discovery efforts must be approached using new and innovative means as described herein. The advancement of DNA sequencing technologies and genomics-based approaches in recent years has opened the door for a multi-disciplinary approach towards the discovery of natural products and enediynes in particular. The potential of the enediynes as relevant clinical drugs and the unique chemistry and enzymology involved in their biosynthesis should be of great interest to cancer biologists, natural product chemists, biochemists, enzymologists, structural biologists, and the community as a

whole. Enediynes will continue to play an important role in natural product biosynthesis, engineering, and drug discovery.

CLAIMS

What is claimed

1. A method of high throughput screening for identifying compounds encoded by an organism comprising:

isolating genomic DNA (gDNA) from the organism;

targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes;

determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences;

comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying a nucleic acid sequence encoding a compound by an organism.

2. A method of high throughput screening for identifying organisms encoding a compound comprising:

isolating genomic DNA (gDNA) from the organism;

targeting and amplifying one or more regions of gDNA comprising conserved nucleic acid sequences or conserved genes;

determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences;

comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby identifying organisms encoding a compound.

3. A method of high throughput screening for prioritizing organisms encoding compounds comprising:

isolating genomic DNA (gDNA) from the organism;

targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes;

determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences;

comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; thereby prioritizing organisms encoding compounds.

4. The method of claims 1, 2, or 3, wherein the gDNA is labeled with a detectable label.

5. The method of claims 1, 2 or 3, wherein the conserved nucleic acid sequences or conserved genes encode for one or more biosynthetic molecules, mutants or variants thereof.

6. The method of claims 5, wherein the biosynthetic molecules are natural and/or synthetic.

7. The method of claim 5, wherein the synthetic molecules are encoded by one or more nucleic acid sequences transfected into an organism.

8. The method of claims 1, 2, or 3, optionally comprising culturing the organisms under varying culture conditions and culture media.

9. The method of claims 1, 2, or 3, wherein the organism encodes for mutated natural or synthetic biosynthetic molecules.

10. The method of claims 1, 2 or 3, wherein the conserved nucleic acid sequences are targeted by primers having a sequence identity of at least about 50% with at least five consecutive nucleic acid bases of a 5' to 3' or 3' to 5' strand of the regions comprising

the conserved nucleic acid sequences, said primers optionally comprising a detectable label.

11. The method of claims 1, 2 or 3, wherein the regions comprising the conserved nucleic acid sequences or conserved genes are amplified by real time polymerase chain reaction (RT-PCR).

12. The method of claim 4, wherein a detectable label comprises: nucleic acid intercalating dyes, fluorophores, radio labeled molecules, fluorescent agents, molecular beacons, chemiluminescent agents, luminescent agents, electron-dense reagents, enzymes, biotin, digoxigenin, haptens or peptides.

13. A method of high throughput screening for prioritizing bacterial strains producing diterpenoids comprising:

isolating genomic DNA (gDNA) from the organism;

targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes;

determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences;

comparing the melting temperatures of the amplified conserved nucleic acid sequences or conserved genes to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; and, identifying a nucleic acid sequence encoding a diterpenoid by a bacterial strain; thereby prioritizing bacterial strains producing diterpenoids.

14. The method of claim 13, wherein the gDNA is labeled with a detectable label.

15. The method of claim 13, wherein the bacterial strain is an Actinomycetale strain.

16. The method of claim 13, wherein the conserved nucleic acid sequences or conserved genes encode for one or more diterpene synthase molecules, mutants or variants thereof.

17. The method of claim 16, wherein the conserved diterpene synthase sequences are targeted by primers having a sequence identity of at least about 50% with at least five consecutive nucleic acid bases of a 5' to 3' or 3' to 5' strand of the conserved diterpene synthase sequences, said primers optionally comprising a detectable label.

18. The method of claim 17, wherein the primers comprise any one or combinations of primers set forth in SEQ ID NOS: 1-127.

19. The method of claim 13, wherein the regions comprising the conserved nucleic acid sequences are amplified by real time polymerase chain reaction (RT-PCR).

20. The method of claim 14, wherein a detectable label comprises: nucleic acid intercalating dyes, fluorophores, radio labeled molecules, fluororescent agents, molecular beacons, chemiluminescent agents, luminescent agents, electron-dense reagents, enzymes, biotin, digoxigenin, haptens or peptides.

21. A method of high throughput screening for prioritizing bacterial strains producing enediynes comprising:

isolating genomic DNA (gDNA) from the organism;

targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes;

determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences;

comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control; and identifying a nucleic acid sequence encoding an enediyne by a bacterial strain; thereby prioritizing bacterial strains producing enediynes.

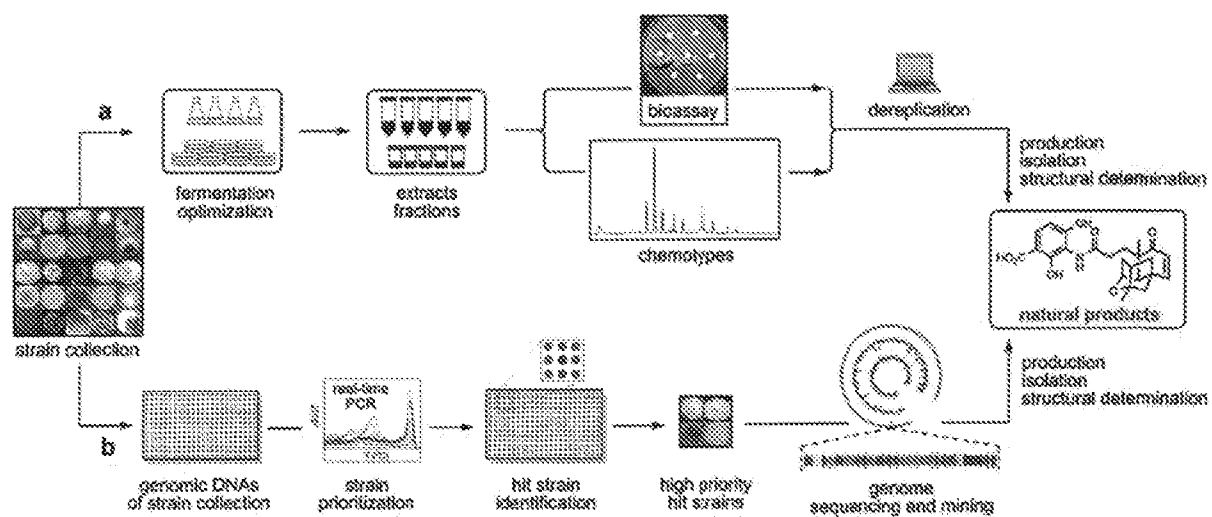
22. The method of claim 21, wherein the gDNA is labeled with a detectable label.
23. The method of claim 21, wherein the bacterial strain is an Actinomycetale strain.
24. The method of claim 21, wherein the conserved nucleic acid sequences or conserved genes encode for one or more enediyne molecules, mutants or variants thereof.
25. The method of claim 24, wherein the conserved nucleic acid sequences or conserved genes comprise: polyketide synthase (PKS) nucleic acid sequences, mutants or variants thereof.
26. The method of claim 21, wherein the conserved nucleic acid sequences or conserved genes encoding one or more enediyne molecules are targeted by primers having a sequence identity of at least about 50% with at least five consecutive nucleic acid bases of a 5' to 3' or 3' to 5' strand of the conserved enediyne nucleic acid producer sequences, said primers optionally comprising detectable label.
27. The method of claim 26, wherein the primers comprise any one or combinations of primers set forth in SEQ ID NOS: 33-38.
28. The method of claim 21, wherein the regions comprising the conserved nucleic acid sequences or conserved genes are amplified by real time polymerase chain reaction (RT-PCR).
29. The method of claim 22, wherein a detectable label comprises: nucleic acid intercalating dyes, fluorophores, radio labeled molecules, fluorescent agents, molecular beacons, chemiluminescent agents, luminescent agents, electron-dense reagents, enzymes, biotin, digoxigenin, haptens or peptides.
30. A compound identified by the methods of claims 1, 2, 3, 13 or 21.

31. A pharmaceutical composition comprising a compound identified by the methods of claims 1, 2, 3, 13 or 21.

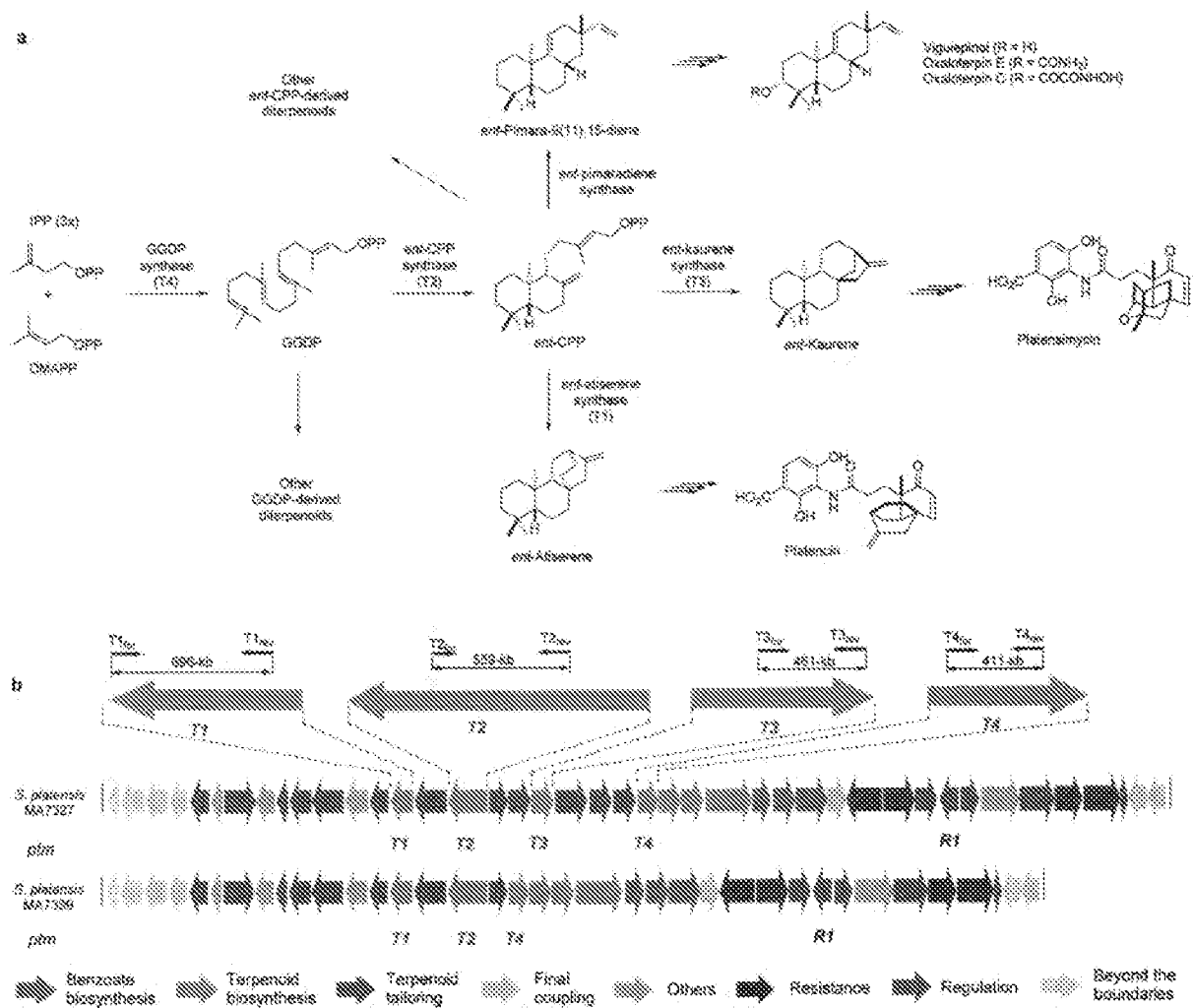
32. An oligonucleotide set forth as SEQ ID NOS: 1-38, mutants, variants or complementary sequences thereof.

33. An oligonucleotide having at least about 50% sequence identity to one or more oligonucleotides set forth as SEQ ID NOS: 1-38, mutants, variants or complementary sequences thereof.

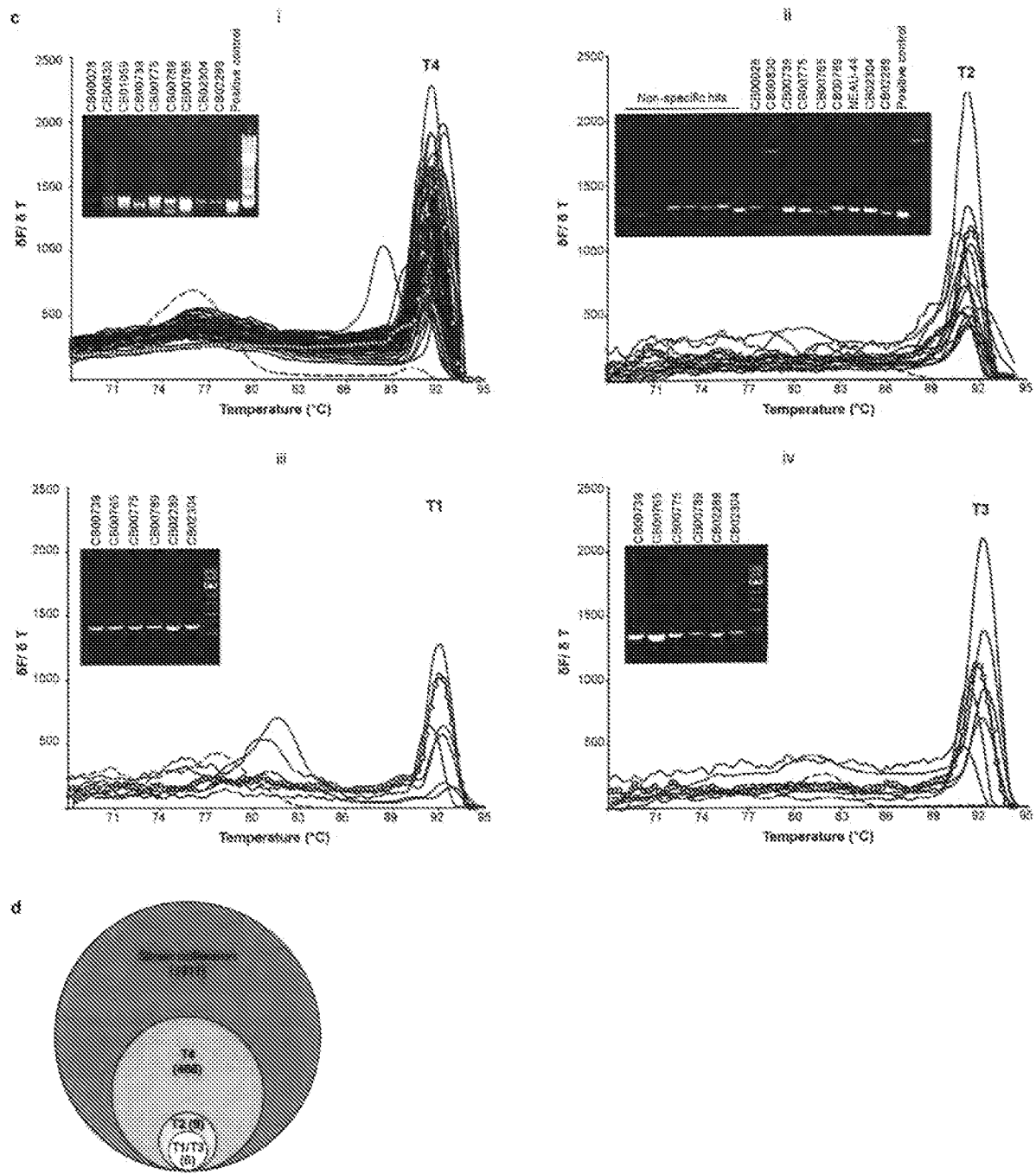
34. An oligonucleotide or peptide encoded therefrom comprising one or more sequences set forth as SEQ ID NOS: 1-127.



FIGURES 1A, 1B



FIGURES 2A, 2B



FIGURES 2C, 2D



FIGURES 3A, 3B, 3C

[illegible]

Sense primer region:

Conserved residues








rtmf1 181 ATCTGTGGGAGGACATCTCCAGAGAACCGGCTTACCGCGGGGTGTACATCCGGTGGGACACC
 rtmf1 181 ATCTGTGGGAGGACATCTCCAGAGAACCGGCTTACCGCGGGGTGTACATCCGGTGGGACACC
 CCGGGGTGTACATCCGGTGGGAC

Antelope point region

~~CONFIDENTIAL~~

gtaT1 000 CTCCTCCCTGGGAGTGGGGTGGGACCTCTCTCTCTGCGGCATCCGCTCTCTCTGCGGCATCGA
 gtaT1 000 CTCCTCCCTGGGAGTGGGGTGGGACCTCTCTCTCTCTGCGGCATCCGCTCTCTCTGCGGCATCGA
 GACCTCTCTCTCTCTGCGGCATCG

[illegible]

Sense primer region:

Conserved residues

5 2 4 4 2

aka3985a_orf2	603	AGGCTTCCGCGCGCGCCACCTCTGGCTGGCTGGCCGATCCGCGCTCATACGGAACG
swt2_orf2	600	AGGCTTCCGCGCGCGCCACCGCGCGCTGGCTGGCTGGCCGATCCGCGCGACACGGAACG
swt1_orf2	645	TTGCTTCCGCGCGCGCCACCGCGCGCTGGCTGGCCGATCCGCGCGACACGGAACG
ptmT2	665	CTGCTCCGCGCGCGCCACCGCGCGCTGGCTGGCGCGCTGGCGGATCCGCGCGCG
ptmT1	665	CTGCTCCGCGCGCGCCACCGCGCGCTGGCTGGCGCGCGCTGGCGGATCCGCGCGCG
		CGCTTCCGCGCGCGCCACG

Accession primer region:

~~converted residues~~








sko3508_orf2	1106	ATCAGGACACCGGACCGCCCGGTGGATGGACAAGTGGCAGCGCCTCTCCCGTACTACG
swt2_orf2	1103	ATCAGGACACCGGACCGCCCGGTGGATGGACAAGTGGCAGCGCCTCTCCCGTACTACG
swt1_orf2	1148	ATCAGGACACCGGACCGCCCGCTGGATGGACAAGTGGCAGCGCCTCCCGTACTACG
ptmT2	1173	ATCCAGCTGCCCCGACCGGCGCTGGATGGACAAGTGGCAGCGCCTCGCCGTACTACG
ptnT2	1173	ATCCAGCTGCCCCGACCGGCGCTGGATGGACAAGTGGCAGCGCCTCGCCGTACTACG

FIGURES 4A, 4B

c.

PtmT4 73 L Y L L D G R L D ... (137) ... A N D L R T W A K 227
 217 ccttatctgtctogaaagggcaggctogaa... (411) ... gccaaagatctgctggacctggggccaag
 Sense primer region Antisense primer region

d.

PtmT4	86	LVNPFLLINDVVMGSLERR... (106) ...	SENHLEMAPOLYDDVIGIWDPA... 246
PtnT4	96	LVNPFLLINDVVMGSLERR... (106) ...	SENHLEMAPOLYDDVIGIWDPA... 246
Kgris_Orf1	86	LVNPFLLINDVVMGSLERR... (106) ...	SENHLEMAPOLYDDVIGIWDPA... 246
PlaT4	106	SENHLEMAPOLYDDVIGIWDPA... (106) ...	SENHLEMAPOLYDDVIGIWDPA... 256
Bra2	77	TVNPA... (112) ...	SENHLEMAPOLYDDVIGIWDPA... 256
Sko3988_Orf4	120	SENHLEMAPOLYDDVIGIWDPA... (122) ...	SENHLEMAPOLYDDVIGIWDPA... 266
consensus	121	..*... ..*	*

Sense primer region Antisense primer region

Sense primer design:

Conserved residues			L X N D D V/I M/I D
ptmT4	271	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	
ptnT4	271	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	
kgris_orf1	270	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	
plaT4	279	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	
bra2	264	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	
sko3988_orf4	269	GGCGCGCCGCTCGAACTGGTCCACAACTTCTCGCTGTTGCACGACGACGTCATGGACCGC	

CTGRTSCACGACGATVTSATGGAC

Antisense primer design:

Conserved residues		F Q X X D D X L
ptmT4	675	CCTGGGATGGGCTTCCAGCTGGTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT
ptnT4	675	CCTGGGATGGGCTTCCAGCTGGTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT
kgris_orf1	674	CCTGGGATGGGCTTCCAGCTGGTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT
plaT4	682	CCTGGGATGGGCTTCCAGCTGGTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT
bra2	636	CCTGGGATGGGCTTCCAGCTGGTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT
sko3988_orf4	751	GATCGGAGAGGCTTCCAACTGGGCTGGGATGACGCTCTGGGATCTGGGAGAACCCGGCGCT

TTCARHWBRYSGAGACVTVMISGGC

FIGURES 4C, 4D

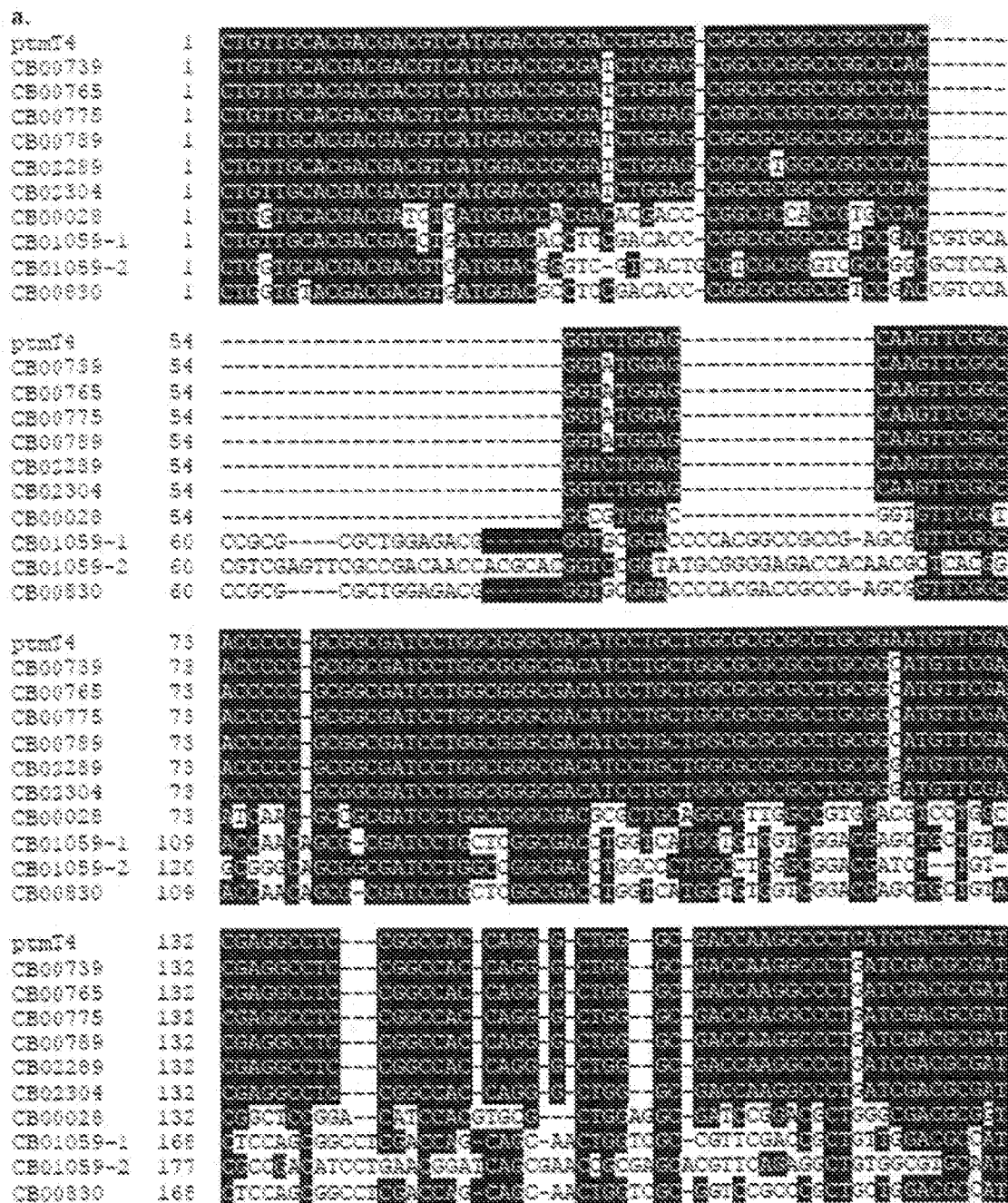


FIGURE 5A

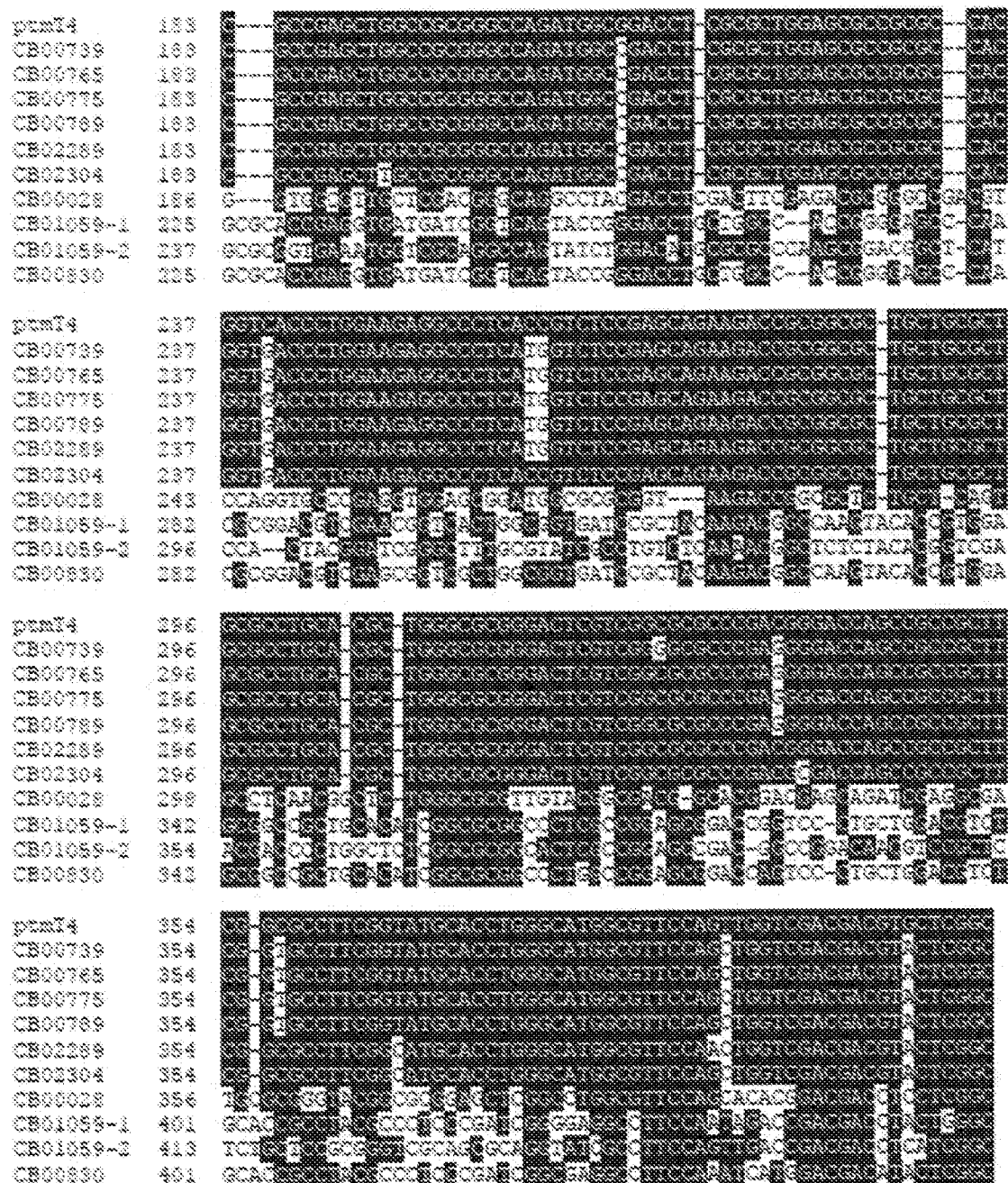


FIGURE 5A (CONTINUED)

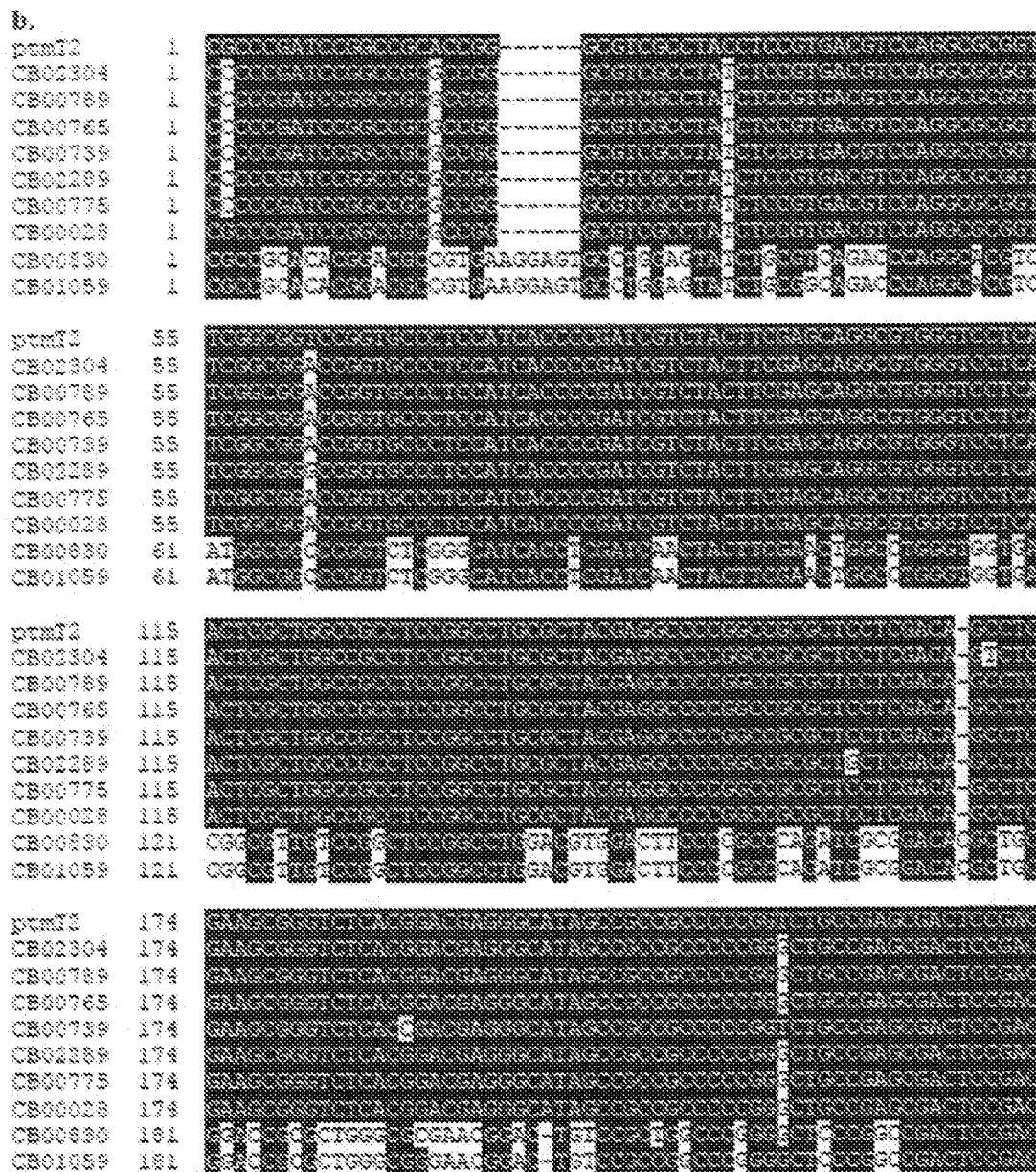


FIGURE 5B

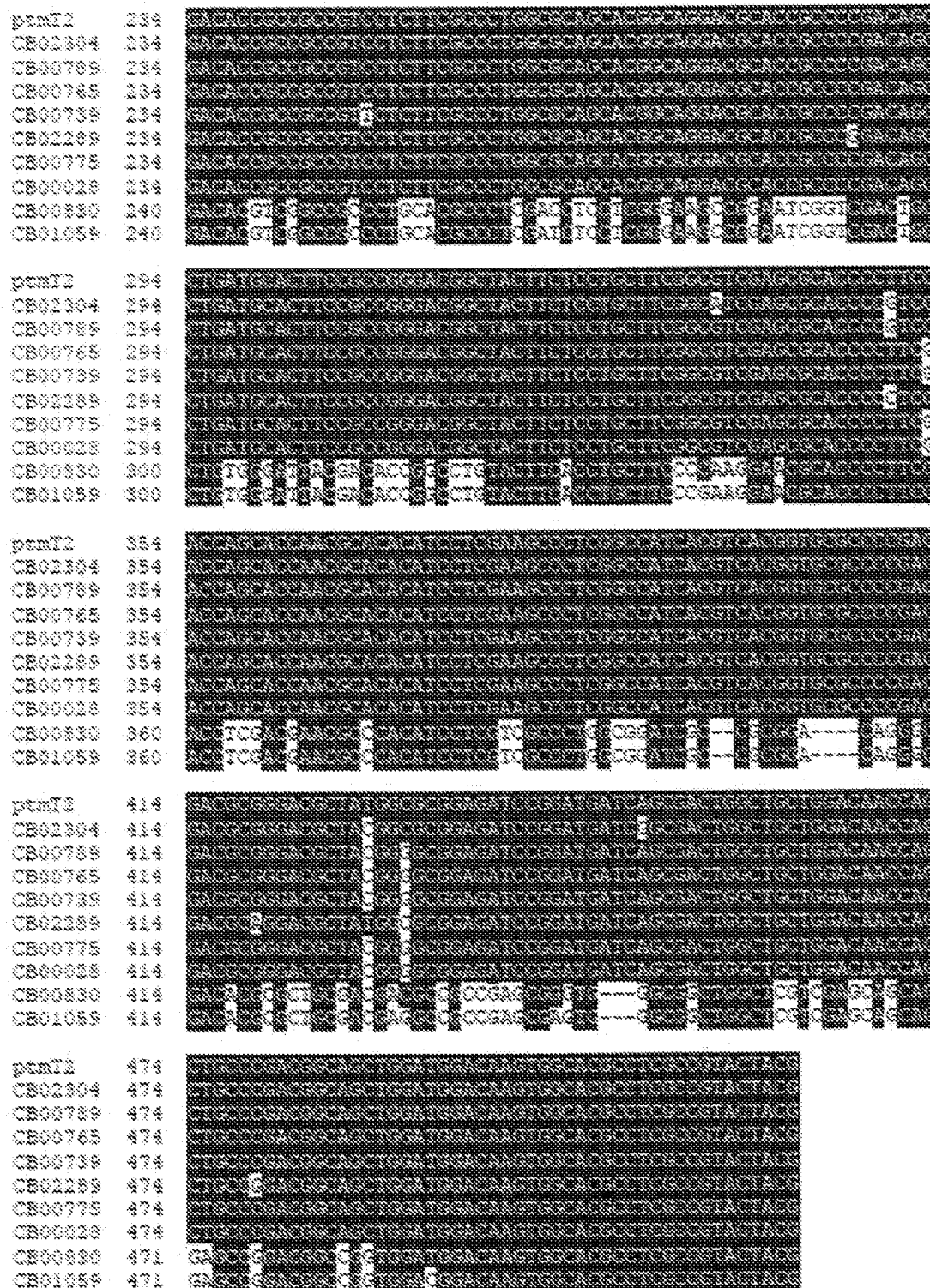


FIGURE 5B (CONTINUED)

E.		
ptmT3	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB00739	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB00765	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB00775	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB00789	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB02289	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
CB02304	1	ATTTCTCGACCGGCTCGACGACTTGGAGCACTACGGCACCACCGCCGAGGACGTC
ptmT3 61		
CB00739	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
CB00765	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
CB00775	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
CB00789	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
CB02289	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
CB02304	61	ACCTGGGCGGATCGCGTACTTCCGCGCGGCGTGGCGAGGTCGAGGCGAGGCTTC
ptmT3 121		
CB00739	121	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
CB00765	121	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
CB00775	121	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
CB00789	131	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
CB02289	121	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
CB02304	121	TGGAGAGCTCGCTCGCGGCACTTGGTGGAGGAGTGGCGGCTCGAGGAGCTTGGAG
ptmT3 181		
CB00739	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
CB00765	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
CB00775	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
CB00789	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
CB02289	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
CB02304	181	TGCTGGAGCGTTGGTGTGGCTTGGTGGAGGAGGAGTGGCTGGAGGAGGAGTGG
ptmT3 241		
CB00739	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
CB00765	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
CB00775	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
CB00789	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
CB02289	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
CB02304	241	TGCTCGAGCGGATCGCGGAGGAGTGGAGGAGTGGAGGAGTGGAGGAGTGGAG
ptmT3 301		
CB00739	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG
CB00765	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG
CB00775	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG
CB00789	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG
CB02289	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG
CB02304	301	GAACTGGCTTACCGGACGCTCACTACCGAGTGTGGGCTGGCACTGGAGGAGTGG

FIGURE 5C

ptmT3	361	GGCGAGCGCTCGGACAGCTCGGGGCGAGGAGCGGCTCGACGCGGCTCTCGTCCCGCGGTGG
CB00739	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGCGGTGG
CB00775	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGCGGTGG
CB00789	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGCGGTGG
CB02304	361	GACGAGCGCTCGGACAGCGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

ptmT3	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00739	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00775	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00789	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02304	421	AGGAGCGCTCGGCTCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

FIGURE 5C (CONTINUED)

d.

ptmT1	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02304	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00775	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00789	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00739	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	1	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

ptmT1	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02304	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00775	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00789	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00739	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	61	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

ptmT1	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02304	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00775	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00789	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00739	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	121	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

ptmT1	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02304	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00765	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00775	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00789	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB00739	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG
CB02289	181	TGGGCTGGGACATCGGAGCGGACAGCGGCTCGGAGCGGCTCGACGCGGCTCTCGTCCCGAGCTCG

FIGURE 5D

FIGURE 5D (CONTINUED)

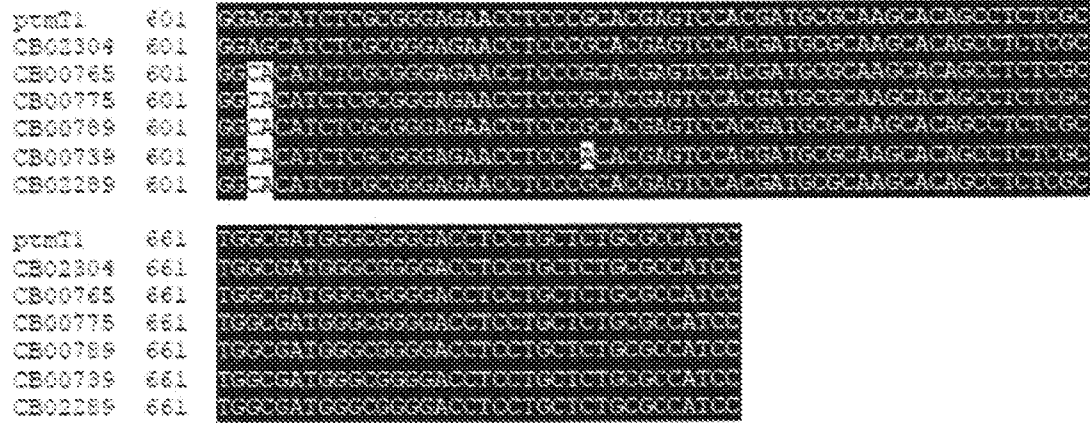


FIGURE 5D (CONTINUED)

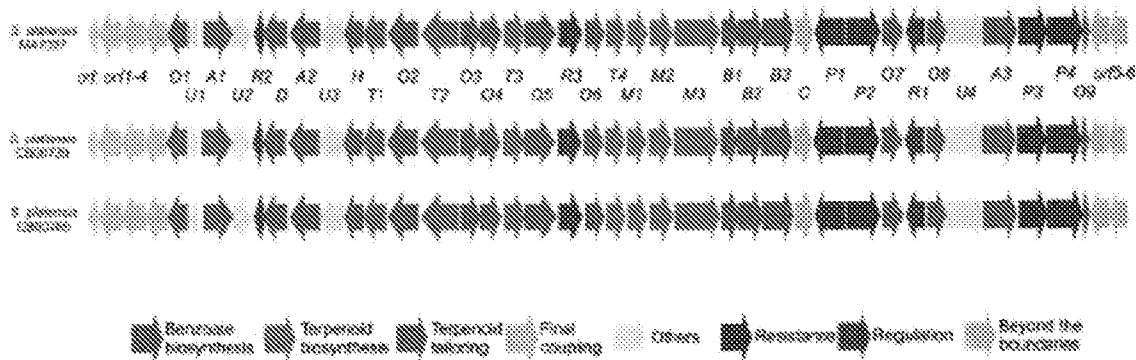
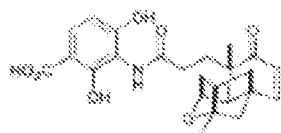


FIGURE 6

a.

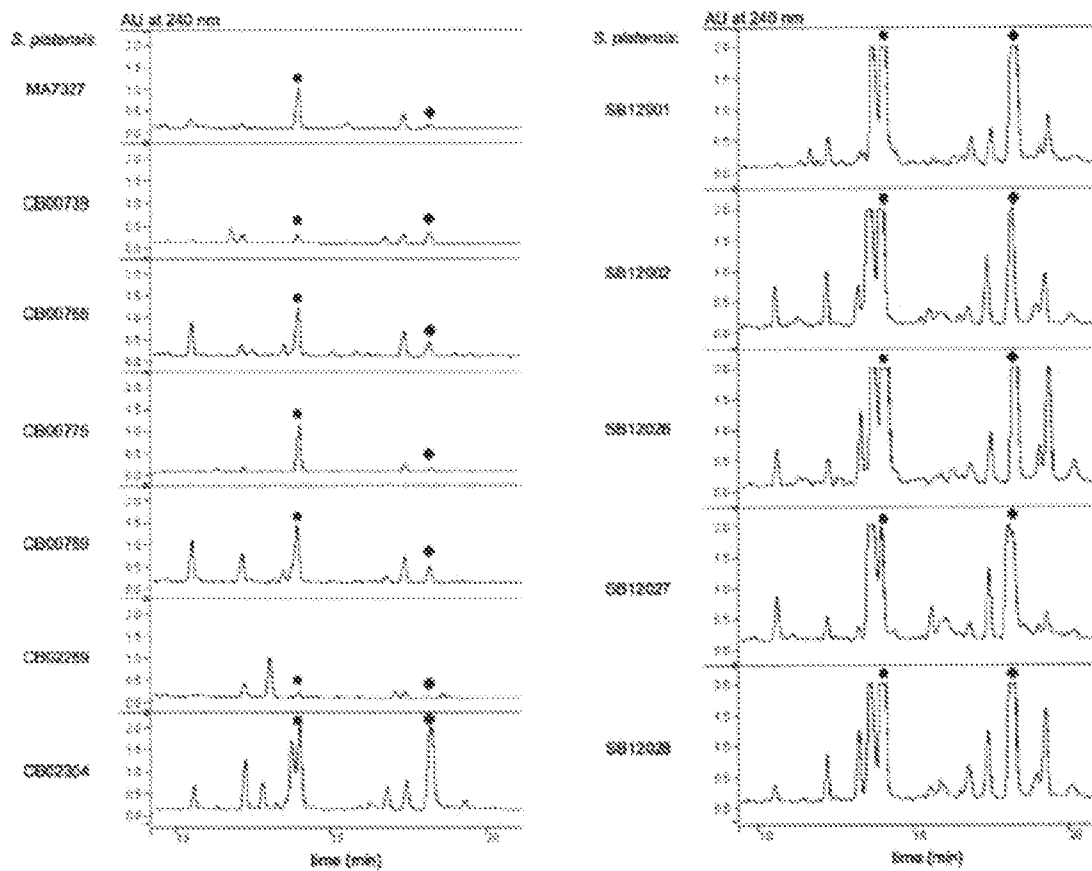


Platensimycin



Platencin

b.



FIGURES 7A, 7B

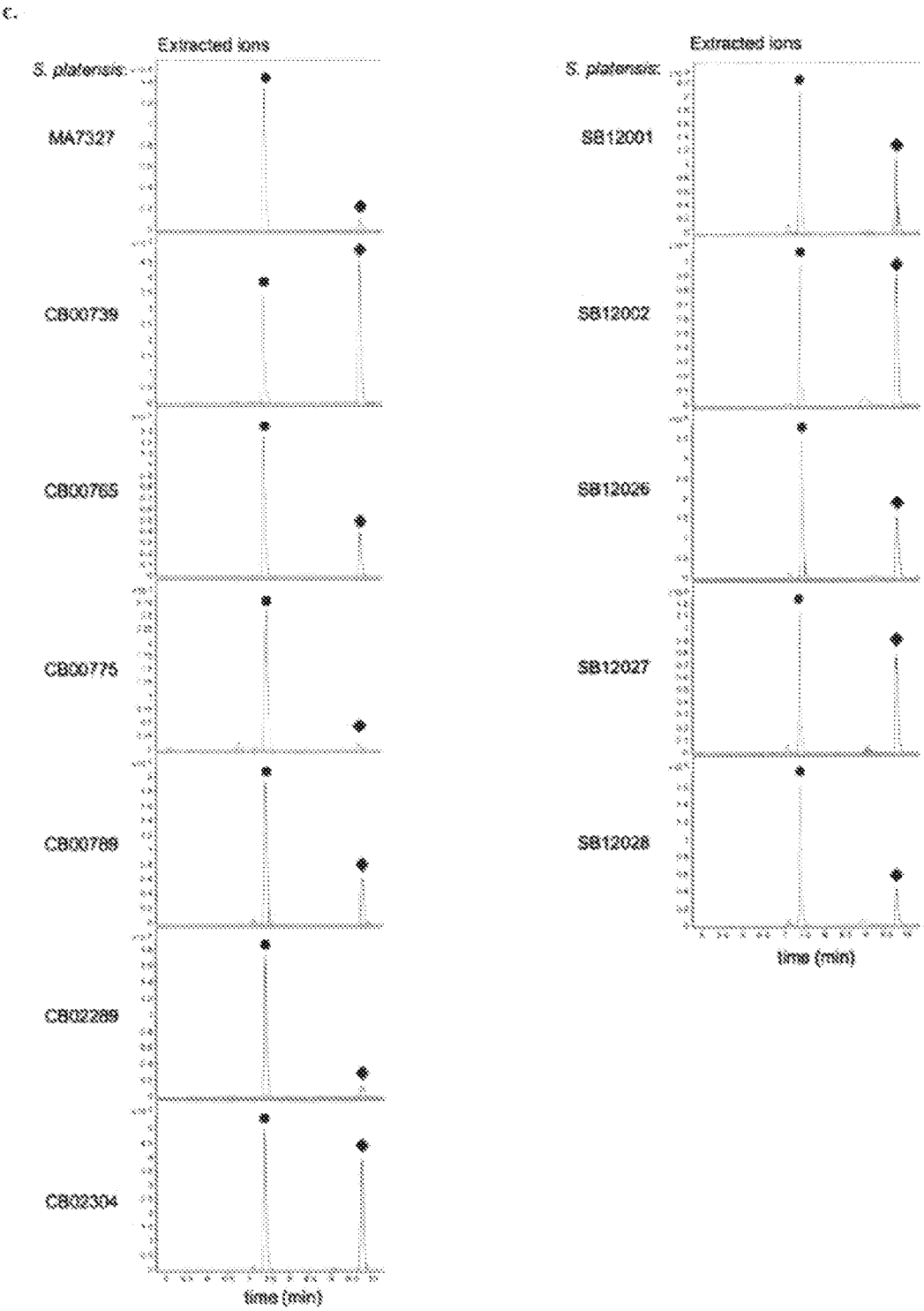
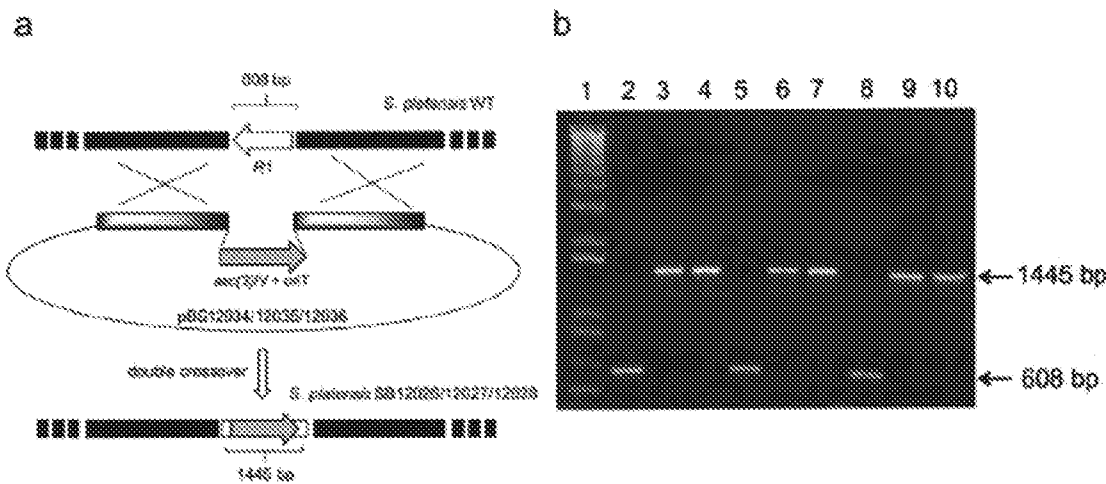
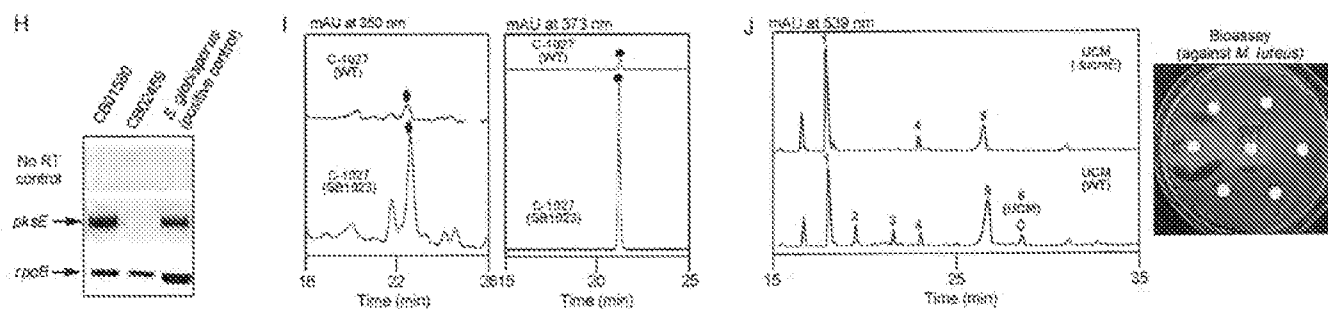
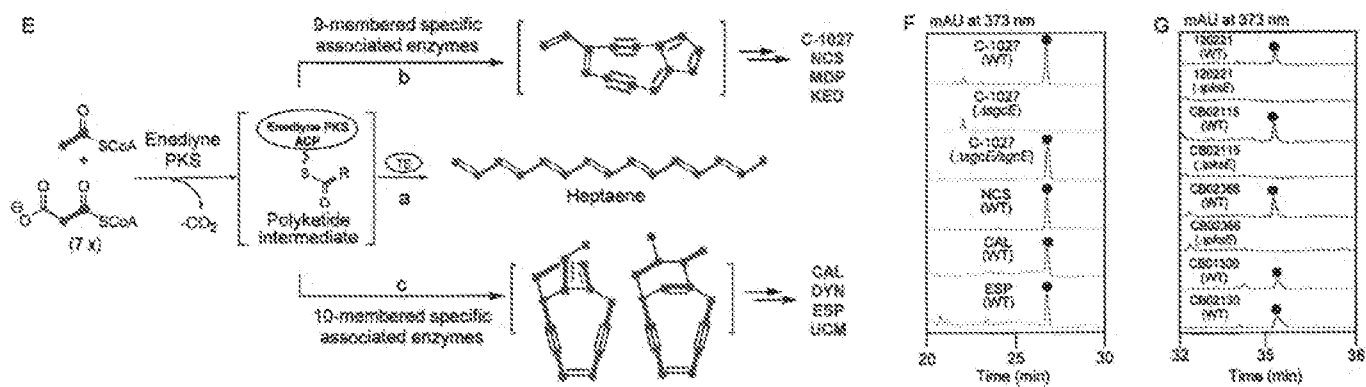
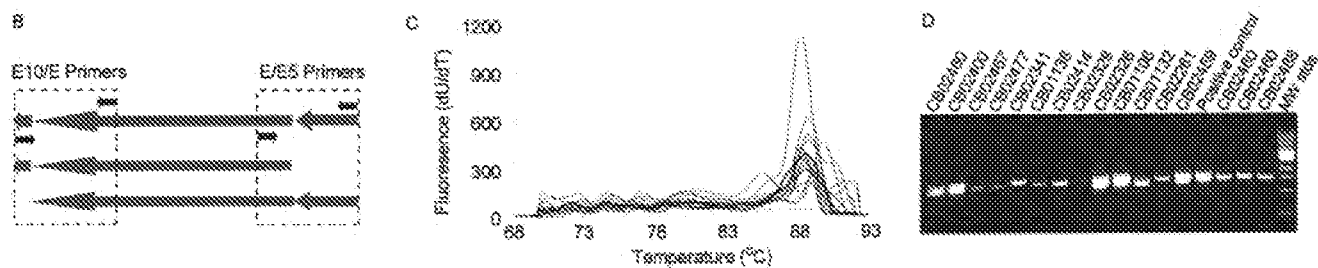


FIGURE 7C



FIGURES 8A, 8B

FIGURE 9A



A

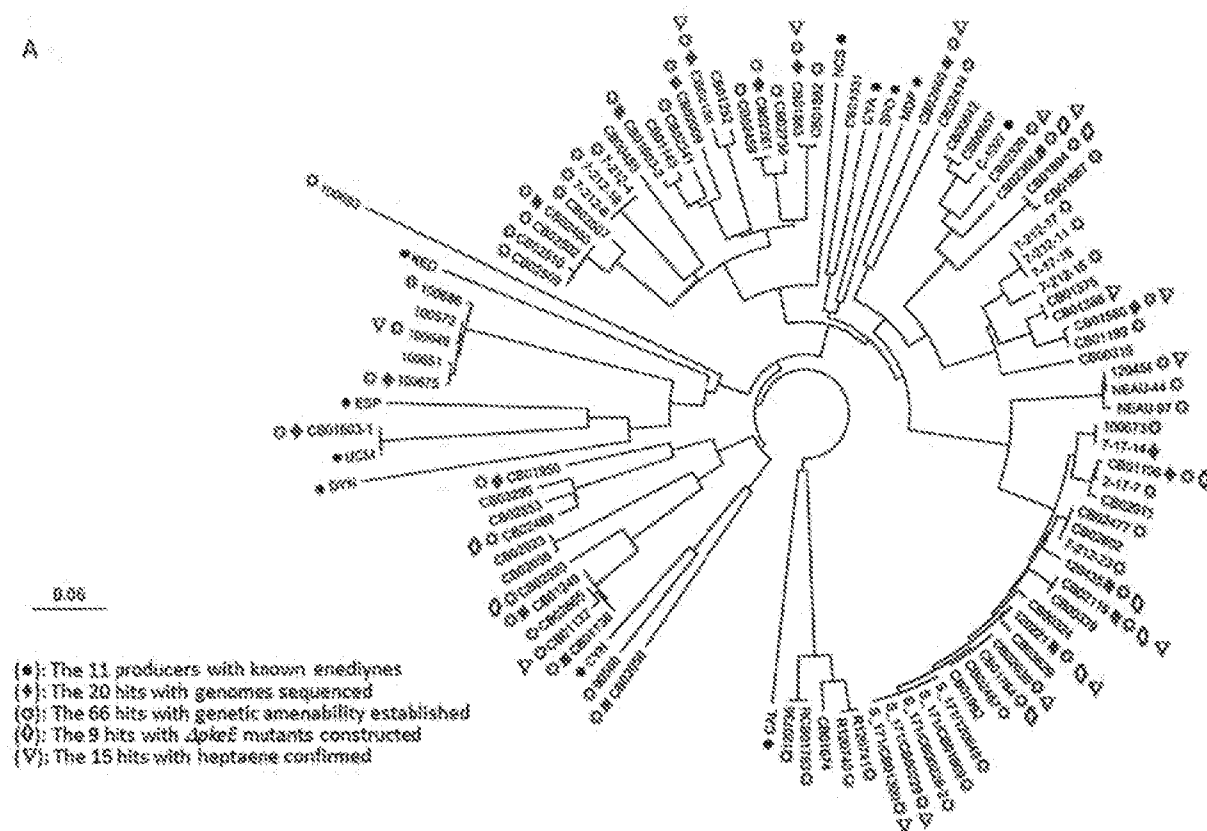


FIGURE 10

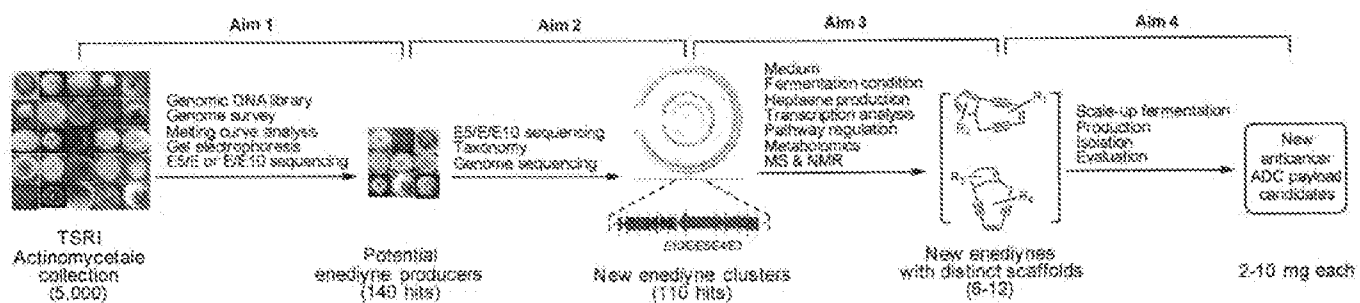


FIGURE 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/37455

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12Q 1/68; C12N 1/20; C07H 21/04 (2015.01)

CPC - C12Q 1/686, 1/689, 1/6876; C12R 1/465, 1/585, 1/545

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC(8)- C12Q 1/68; C12N 1/20; C07H 21/04 (2015.01)

CPC- C12Q 1/686, 1/689, 1/6876; C12R 1/465, 1/585, 1/545

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

CPC: C12Q 1/686, 1/689, 1/6876; C12R 1/465, 1/585, 1/545 (text search)

UPSC: 435/6.11, 6.12, 252.1, 252.35, 253.5; 536/24.32, 24.33

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Electronic data bases: PatBase; Google Patents; Google Scholar; GenCore Sequence Search (NT)

Search terms: high resolution melting, Tm deviation from control, RT-PCR, evolutionary relation, detectable label, pathway specific biosynthetic gene cluster, Actinomycetes, Streptomyces griseus CB00830, S. platensis, diterpene synthase genes (DTS)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ----- Y	US 2003/0215798 A1 (SHORT et al.) 20 November 2003 (20.11.2003). Especially para [0347], SEQ ID NO: 1	32-34 ----- 18
Y	XIE et al. Biosynthetic Potential-Based Strain Prioritization for Natural Product Discovery: A Showcase for Diterpenoid-Producing Actinomycetes. J Nat Prod ePub 31 January 2014 Vol 77 No 2 Pages 377-387 . and online Suppl. Data. Especially pg 379 col 1 para 3, pg 379 fig 1, pg 380 col 2 para 2, pg 381 fig 3, pg 383 col 2 para 4, pg 385 col 1 para 5, pg S2 tables S1 and S2, pg S8 fig S2, pg S10 fig S6.	1-20, (30,31)/(1-3,13)
Y	SCHMIDERER et al. DNA-based identification of Helleborus niger by high-resolution melting analysis. Planta Med November 2010 Vol 76 No 16 Pages 1934-1937. Especially pg 1934 col 1 para 2	1-20, (30,31)/(1-3,13)
Y	SMANSKI et al. Bacterial diterpene synthases: new opportunities for mechanistic enzymology and engineered biosynthesis. Curr Opin Chem Biol April 2012 Vol 16 No 1-2 Pages 1934-1937. Especially Pg 132 col 2 para 2.	7
X,P	HINDRA et al. Strain prioritization for natural product discovery by a high-throughput real-time PCR method. J Nat Prod Epub 19 September 2014 Vol 77 No 10 Pages 2296-2303. entire article.	1-20, (30,31)/(1-3,13)

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

21 October 2015 (21.10.2015)

Date of mailing of the international search report

04 DEC 2015

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/37455

Box No. 1 Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

- a. ☒ forming part of the international application as filed:
 - ☒ in the form of an Annex C/ST.25 text file.
 - ☒ on paper or in the form of an image file.
- b. ☐ furnished together with the international application under PCT Rule 13*ter*. 1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
- c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
 - ☐ in the form of an Annex C/ST.25 text file (Rule 13*ter*. 1(a)).
 - ☐ on paper or in the form of an image file (Rule 13*ter*. 1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

GenCore ver 6.4.1 SEQ ID NOs: 1, 2, 3, 4, 9, 10 were specifically searched as part of the ISR.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/37455

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

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

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

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

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

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

This International Searching Authority found multiple inventions in this international application, as follows:
-----go to Extra Sheet for continuation-----

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-34, limited to the first named compound (claim 18, diterpenoid) and first named sequence (SEQ ID NO: 1)
(claims 1-20, (30,31) (in part), 32-34)

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/37455

-----continuation of Box III (Lack of Unity of Invention)-----

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I+: Claims 1-34, drawn to a method of or a composition for high throughput screening to identify organisms that produce a particular compound.

The method of or composition for identifying organisms that produce a particular compound is limited to the first named compound, diterpenoid (claim 13), and the first named primer sequence, SEQ ID NO: 1 (claim 18) used to identify a gene in the diterpenoid synthetic pathway [primer for ptm/ptnT1, see instant application pg 37 table 2]. It is believed that claims 1-20, (30,31) (in part), 32-34 read on this first named invention and thus these claims will be searched without fee to the extent that they encompass a diterpenoid compound and primer SEQ ID NO: 1. Additional compounds and a primer used to identify a nucleic acid encoding a gene in the compound's synthetic pathway will be searched upon payment of additional fees. Applicant must specify the claims that encompass any additional elected target gene(s). Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be: an enediynes compound (claim 21) and a primer used to identify a nucleic acid encoding a gene in the enediynes synthetic pathway, SEQ ID NO: 33 (claim 27) (claims 1-12, 21-29, (30, 31) (in part), 32-34).

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

Special Technical Features:

Among the inventions listed as Groups I+ is the specific sequences recited therein. The inventions do not share a special technical feature, because no significant structural similarities can readily be ascertained among sequences.

Among the inventions listed as Groups I+ is the specific compounds recited therein (e.g. diterpenoid, enediynes). The inventions do not share a special technical feature, because no significant structural similarities can readily be ascertained among compounds.

Common Technical Features:

1. isolating genomic DNA (gDNA) from the organism
2. targeting and amplifying one or more regions of gDNA comprising conserved nucleic acid sequences or conserved genes
3. determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences
4. comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control
5. identifying organisms encoding a compound

However, said common technical feature does not represent a contribution over the prior art, and is obvious over the publication titled "Biosynthetic Potential-Based Strain Prioritization for Natural Product Discovery: A Showcase for Diterpenoid-Producing Actinomycetes" by XIE et al. (hereinafter "Xie") [published 28 February 2014 in J Nat Prod Vol 77 No 2 Pages 377-387], in view of the publication titled "DNA-based identification of *Helleborus niger* by high-resolution melting analysis" by SCHMIDERER (hereinafter "Schmiderer") [published November 2010 in Planta Med Vol 76 No 16 Pages 1934-1937], in view of US 2003/0215798 A1 to (SHORT et al.) (hereinafter "Short").

Concerning common technical features #1, #2 and #5, Xie teaches isolating genomic DNA (gDNA) from the organism (abstract; "surveying 100 strains, randomly selected from our actinomycete collection, for their biosynthetic potential") and targeting and amplifying one or more regions of gDNA comprising conserved nucleic acid sequences or conserved genes (pg 379 fig 1A,C; "Biosynthetic potential-based strain prioritization for natural product discovery: (A) amplification of genes characteristic of representative biosynthetic machinery from a pool of genomic DNA by PCR using pathway-specific primers [...] survey of the biosynthetic machinery of the targeted class of natural products by screening a genomic DNA array of the strains with the pool of pathway-specific probes") and identifying organisms encoding a compound (pg 379 fig 1D; "fermentation optimization of "hit" or "talented" stains for production and isolation of the targeted natural products. Five strains are depicted, with strains 1-4 harboring the biosynthetic machineries of the targeted class of natural products").

-----continued on next sheet-----

-----continued from previous sheet-----

Xie does not specifically teach common technical features #3 or #4. However, Schimderer teaches determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences (pg 1934 col 1 para 2; "The underground parts of *Helleborus niger* (rootstock and roots) are used in homeopathy in cases of meningitis, convulsions, hydrocephalus, dropsy, and cachexia due to tumours"; pg 1936 fig 2; "HRM [i.e. high resolution melting] based on the *matK* sequence. Group I: *H. cyclophyllus*, *H. dumetorum*, *H. multifidus*, *H. odoratus* (p.p.), *H. purpurascens*, *H. viridis*; group II: *H. bocconei*, *H. caucasicus*, *H. foetidus*, *H. odoratus* (p.p.), *H. orientalis*; group III: *H. lividus*; group IV: *H. niger*; group V: *H. argutifolius* (group comprising the target species *H. niger* underlined)" [T_m 's indicated in figure]) and comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control (pg 1936 fig 2; pg 1934 col 1 para 4; In this paper we report a method to distinguish *Helleborus niger* from other species of the genus *Helleborus*. In the two chloroplast regions *matK* and the *trnL-trnF* intergenic spacer, species-specific differences in DNA sequence could be detected, and primers specifically adapted to HRM were developed and tested in an extensive set of species (Table 1). The assay was designed to amplify all species from the genus *Helleborus* and to subsequently identify the species by their melting curves. T_m values [melting temperature at the inflection point(s)]").

It would have been obvious to an artisan of ordinary skill to have incorporated high resolution melting curves and determination of T_m and selecting melting curves with a deviation of 0.0001°C to 5°C from control, as taught by Schimderer, into the method of high throughput screening for identifying compounds encoded by an organism taught by Xie.

Concerning Group I+, claims 1, 2, and 3, Xie teaches isolating genomic DNA (gDNA) from the organism (abstract) targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes (pg 379 fig 1A, C); [claim 1]thereby identifying a nucleic acid sequence encoding a compound by an organism (pg 379 fig 1C,D), or [claim 2] thereby identifying organisms encoding a compound (pg 379 fig 1C,D) or [claim 3] thereby prioritizing organisms encoding compounds (pg 379 fig 1C,D). Xie does not teach determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences or comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control. However, Schimderer teaches these claim limitations (see common technical features #3 and #4, as discussed above). It would have been obvious to an artisan of ordinary skill to have incorporated high resolution melting curves and determination of T_m and selecting melting curves with a deviation of 0.0001°C to 5°C from control, as taught by Schimderer, into the method of high throughput screening for identifying compounds encoded by an organism taught by Xie.

Concerning Group I+ claim 13, Xie teaches a method of high throughput screening for prioritizing bacterial strains producing diterpenoids (pg 379 col 1 para 3; Figure 1 depicts the design of the method to survey biosynthetic potential in microorganisms and application of the method to prioritize strains for targeted natural product discovery. We first selected 16 representative strains, known to produce [...] diterpenoids (two) comprising isolating genomic DNA (gDNA) from the organism (abstract) targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes (pg 379 fig 1A, C); identifying a nucleic acid sequence encoding a diterpenoid by a bacterial strain; thereby prioritizing bacterial strains producing diterpenoids. (pg 379 col 2 para 1; While the DTS pool of probes hybridized only to the two known diterpenoid producers, hybridizations using the other three pools of probes revealed that many of the 16 strains"). Xie does not teach determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences or comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control. However, Schimderer teaches these claim limitations (see common technical features #3 and #4, as discussed above). It would have been obvious to an artisan of ordinary skill to have incorporated high resolution melting curves and determination of T_m and selecting melting curves with a deviation of 0.0001°C to 5°C from control, as taught by Schimderer, into the method of high throughput screening for identifying diterpenoids encoded by an organism taught by Xie.

Concerning Group I+ claim 20, Xie teaches a method of high throughput screening for prioritizing bacterial strains producing enediynes (pg 379 col 1 para 3; Figure 1 depicts the design of the method to survey biosynthetic potential in microorganisms and application of the method to prioritize strains for targeted natural product discovery. We first selected 16 representative strains, known to produce [...] i.e., type I PKS for reduced polyketides, type II PKSs for aromatic polyketides; pg 378 col 2 para 3; elements of the biosynthetic machineries that are highly conserved across the entire family of natural products, as exemplified by the type I polyketide synthases (PKSs) for reduced polyketides") comprising isolating genomic DNA (gDNA) from the organism (abstract) targeting and amplifying regions of gDNA comprising conserved nucleic acid sequences or conserved genes (pg 379 fig 1A, C); identifying a nucleic acid sequence encoding an enediyne by a bacterial strain; thereby prioritizing bacterial strains producing enediynes (pg 379 col 2 para 1; Ten randomly selected clones from each pool of the probes were sequenced, and all of them were confirmed to be the targeted pathway specific genes, with eight variants each for type I and type II PKSs [...] Finally, the genomic DNAs of the 16 strains were arrayed on a nylon membrane and screened with each pool of pathway-specific probes. Six of the six reduced polyketide producers, four of the four aromatic polyketide producers"). Xie does not teach determining melting temperatures (T_m) of the amplified conserved nucleic acid sequences or comparing the melting temperatures of the amplified conserved nucleic acid sequences to the melting temperature of a control and selecting the conserved nucleic acid sequences having a melting temperature within about 0.0001°C to about 5°C deviation from the control. However, Schimderer teaches these claim limitations (see common technical features #3 and #4, as discussed above). It would have been obvious to an artisan of ordinary skill to have incorporated high resolution melting curves and determination of T_m and selecting melting curves with a deviation of 0.0001°C to 5°C from control, as taught by Schimderer, into the method of high throughput screening for identifying enediynes encoded by an organism taught by Xie.

-----continued on next sheet-----

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 15/37455

-----continued from previous sheet-----

Concerning Group I+, claims 32, an oligonucleotide set forth as SEQ ID NOS: 1-38, mutants, variants or complementary sequences thereof, Short teaches SEQ ID NO: 9 (para [0347], SEQ ID NO: 1, 5'-agagttgacctggctcag-3', 20 nt, 100% sequence identity)

Concerning Group I+, claim 33, an oligonucleotide having at least about 50% sequence identity to one or more oligonucleotides set forth as SEQ ID NOS: 1-38, mutants, variants or complementary sequences thereof, Short teaches SEQ ID NO: 9 (para [0347], SEQ ID NO: 1, 5'-agagttgacctggctcag-3', 20 nt, 100% sequence identity)

Concerning Group I+, claim 34, an oligonucleotide or peptide encoded therefrom comprising one or more sequences set forth as SEQ ID NOS: 1-127, Short teaches SEQ ID NO: 9 (para [0347], SEQ ID NO: 1, 5'-agagttgacctggctcag-3', 20 nt, 100% sequence identity)

As the common technical features were known in the art at the time of the invention, they cannot be considered common special technical features that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Group I+ lacks unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature