Generate a bioactive natural product library by mining bacterial cytochrome P450 patterns

Xiangyang Liu *

UNT System College of Pharmacy, University of North Texas Health Science Center, Fort Worth, TX 76107, USA

ABSTRACT

The increased number of annotated bacterial genomes provides a vast resource for genome mining. Several bacterial natural products with epoxide groups have been identified as pre-mRNA spliceosome inhibitors and antitumor compounds through genome mining. These epoxide-containing natural products feature a common biosynthetic characteristic that cytochrome P450s (CYPs) and its patterns such as epoxidases are employed in the tailoring reactions. The tailoring enzyme patterns are essential to both biological activities and structural diversity of natural products, and can be used for enzyme pattern-based genome mining. Recent development of direct cloning, heterologous expression, manipulation of the biosynthetic pathways and the CRISPR-CAS9 system have provided molecular biology tools to turn on or pull out nascent biosynthetic gene clusters to generate a microbial natural product library. This review focuses on a library of epoxide-containing natural products and their associated CYPs, with the intention to provide strategies on diversifying the structures of CYP-catalyzed bioactive natural products. It is conceivable that a library of diversified bioactive natural products will be created by pattern-based genome mining, direct cloning and heterologous expression as well as the genomic manipulation.

© 2016 The authors. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The conception of the natural product library originated from the advocacy for the Biological Resource Center (BRC) via the Organization for the Economic Cooperation and Development (OECD) in 1999. During the past 15 years, the production of a library of microbial natural products has been mainly driven by the development of high-throughput screening methods and the increasing number of reported marine natural products. Broadly speaking, researches on natural product libraries have included the activities of creating a genomic DNA prioritized strain library, a metagenome library, a combinational library, a crude extract library, a fraction library and a pure compound library. Additionally, the bioactive compound library is commercially available from Selleckchem, OTAVA and Cromadex. Strategies for developing a high quality microbial natural product library include diversification and de-replication of strain, natural product structures and high throughput screening methods. Even so, the low outcomes from the creation of a crude extract library and a fraction library are far from the expectation in drug discovery. These facts have diminished the interests of big pharma in the natural product library.

Nature is always full of diversity and natural product biosynthesis is no exception. The “co-linearity” rule, and the diversity and variations in nonribosomal peptide synthetases (NRPSs), polyketide synthases (PKSs), hybrid NRPS/PKS systems and the beyond have been greatly explored. During last two decades, well over hundreds of natural product biosynthetic pathways and thousands of natural scaffolds such as nonribosomal peptide, polyketide, terpenoid and oligosaccharide have been characterized. Except for the enzymes that catalyze formation of chemical scaffolds, there is a group of enzymes that decorate the functional group density, acting sequentially to transfer the building blocks into complex molecular structures. The tailoring enzymes can be classified into nonoxidative and oxidative enzymes. Nonoxidative enzymes include...
acyltransferases, methyltransferases and glycine transferases, while oxidative enzymes encompass those who introduce water-solubilizing oxygen functionalities during maturation and are probably most consequential for introduction of both scaffold and functional group complexity. While considering how to diversify nature’s small molecule inventory, the oxygenase enzymes have proven to be most valuable to characterize with the twin goals of prediction of new natural product scaffolds and combinatorial engineering of intermediates.\textsuperscript{33,34} The most prominent of the oxidative enzymes are the heme iron cytochrome P450 (CYPs) monooxygenases found in microbes such as Streptomyces species,\textsuperscript{35} Bacillus species\textsuperscript{36} and Cyanobacteria,\textsuperscript{37} etc.\textsuperscript{38}

Traditionally, natural product genome mining means “use analysis of DNA sequence data to predict structural elements of new natural products and then use this information to design strategies for rapidly identifying, purifying, and structurally characterizing the compounds”.\textsuperscript{39,40} However, the peculiarities of biosynthetic pathways indicate that textbook “co-linearity” rules cannot be applied to deduce structures from all DNA data.\textsuperscript{41,42} Thanks to the development of gene cluster prediction software such as antiSMASH,\textsuperscript{43} genome mining has become a quick and inexpensive way to analyze the biosynthetic potential of sequenced microbes. The current genome mining approaches include sequence-based genome mining\textsuperscript{44,45} bioactivity-guided genome mining,\textsuperscript{46} enzyme-based genome mining,\textsuperscript{47} pattern-based genome mining\textsuperscript{48,49} and genome neighborhood network analysis.\textsuperscript{50,51} Sequence-based genome mining is designed to detect and extract carboxyl (C) – and keto-synthase (KS) – domains from DNA or amino acid sequence data. The high degrees of sequence similarity (E-value <10e-40 for eSNAPD or 85–90% protein sequence identity for NaPDOS) suggest that identified biosynthetic gene clusters are responsible for the biosynthesis of similar natural products resulting from reference gene clusters.\textsuperscript{52,44,52} For bioactivity-guided genome mining, the conserved structural motif of bioactive natural products can be used as a reference to perform genome mining. For example, the enediyne PKS (PKSE) is proposed to be involved in the formation of the highly reactive chromophore ring structure (or “warhead”) found in all enediyynes. By PKSE based genome mining, the enediyne biosynthetic gene clusters have been identified from sequenced actinomycete genomes.\textsuperscript{53} FK228 is an antitumor drug with tumor cytotoxicity resulting from the functional disulfide group, which is catalyzed by the FAD-dependent disulfide oxidoreductase.\textsuperscript{44,45} BLAST analysis of FAD-dependent disulfide oxidoreductase has led to the discovery of a homologous gene cluster of thailandepsins in Burkholderia thailandensis E264.\textsuperscript{56} Enzyme-based genome mining searches conserved synthase domains against the NCBI database of sequenced bacterial genomes in order to obtain the presumptive enzyme sequences.\textsuperscript{47} The “pattern-based genome mining” refers to the connection of MS/MS fragmentation pattern to the biosynthetic pathways genome mining and de-replication of certain bacterial species. For example, the MS/MS fragmentation pattern of the 287,492 for arenicicol A production was used to identify the uncharacterized gene cluster in S. pacifica strains CNO-748 and CNT-138.\textsuperscript{48} The “genome neighborhood networks (GNN) analysis” is a bioinformatics strategy to predict enzymatic functions on a large scale based on their genomic context. In this case, bioinformatics of PepeM and phosphate GNN were applied for 278 sequenced bacterial genomes and led to the discovery of 19 new phosphate natural products.\textsuperscript{59} Enediyne GNNs were generated for the virtual screening of the sequenced bacterial genomes resulted in 87 potential enediyne gene clusters from 78 different bacteria strains.\textsuperscript{50} The pattern-based genome mining and the genome neighborhood analysis provide a comprehensive method to identify the biosynthetic gene cluster of sequenced bacterial strains.

With the physiological and medical Nobel Prize awarding to natural product avermectin and artemisinin, the discovery of natural products has entered a golden age.\textsuperscript{58} Our long term goal is to create a high quality, diversified natural product library. In the postgenomic era, the challenges to generate a microbial natural product library have been switched from the traditional de-replication strategies to issues of how to translate the annotated biosynthetic gene clusters of interest to a bioactive natural product library. The current review will focus on genome mining of CYPs which are involved in the biosynthetic gene clusters of bacterial secondary metabolites, especially those with epoxide functional groups, with an intention to share the considerations to build a diversified natural product library through CYP pattern-based genome mining, direct cloning and heterologous expression, and genome manipulation.

2. Genome mining of CYP-catalyzed bacterial natural products

This section will cover the introduction of CYP, two genome mining methods, and application of genome mining in two genera.

2.1. The importance of microbial CYPs

Microbial natural products catalyzed by CYP biosynthetic pathways have diversified biological activities including antitumor activities, antibacterial activities, antifungal activities, anti-HIV activities, anti-parasitic and anti-cholesterol activities. Natural products such as pladienolides/FD-895,\textsuperscript{59,60} GEX1/herboxidine,\textsuperscript{61} FR901464 (FR)/spliceostatins/thailanstatins\textsuperscript{62} (Fig. 1A) are known to have antitumor activities by targeting pre-mRNA spliceosome. Specifically, pladienolide B and spliceostatin A have been reported to express their antiproliferative activities against tumor cells at the nM range by targeting the splicing factor 3B subunit 1 (SF3B1) with the help of three-membered epoxide groups. Furthermore, CYP-catalyzed natural products with antitumor activities may have different mechanisms. For instance, epothilone A expresses the tubulin-binding activity,\textsuperscript{63} griseorhadin A is telomerase inhibitor,\textsuperscript{64} tirandamycin is RNA polymerase inhibitor\textsuperscript{65} and epoxomicin is proteasome-inhibitor\textsuperscript{66} etc (Fig. 1B).

Microbial CYPs are one of the most widely distributed groups of tailoring enzymes to catalyze the formation of the final bioactive natural products.\textsuperscript{51,35,37,67–69} CYPs catalyze the hydroxylation and/or epoxidation reactions in the late stages of biosynthesis after macroclide formation by PKSs.\textsuperscript{70–72} Reactions catalyzed by CYP monooxygenases include hydroxylation of saturated C–H bonds, epoxidation of C=C double bonds, and oxidative decarboxylation (Table 1).\textsuperscript{106} For example, The CYP PdB hydroxylates 6-deoxy pladienolid B to pladienolid B in the pladienolid biosynthetic pathway.\textsuperscript{59} The CYP HerG catalyzes the stereospecific hydroxylation at C-18 of herboxidine.\textsuperscript{61} For FR biosynthesis, the CYP Fri9R is not only involved in the hydroxylation at C-4 of FR901464, but is also related to the formation of the hemiketal group at C-1.\textsuperscript{62} CYPs in the biosynthetic pathways of natural products often work together with its patterns such as transferases and epoxidases to catalyze the production of epoxide groups. Moreover, dual-function CYPs have been reported to catalyze sequential epoxidation and hydroxylation of the same substrate.\textsuperscript{65,68,107}

2.2. Bioinformatics of microbial CYPs

The information on three-dimensional structures is essential to understand the molecular basis for substrate recognition and specificity of CYPs. It is universally thought that three elements are essential in all CYPs sequences: (a) the conserved cysteine, which is the fifth ligand to the heme Fe atom and can be represented as FXGXXXGXG, (b) the EXXR motif forming a charge pair in the K helix (possibly involved in heme binding) and (c) overall CYP fold topology, including stability.\textsuperscript{108} However, there are some unusual examples for bacterial CYPs, such as variations at conserved motifs,
heme incorporation and topology, substrate binding and functionalities. For example, the amino acid sequence of CYP157C1 contains EQSLW in place of the conserved EXXR.

The conserved threonine in CYP107A1 is not present and a hydroxyl group of the substrate 6-deoxyerythronolide B can directly donate a hydrogen bond to the Fe-linked dioxygen for proton transfer.

Different from CYP107A1, CYP158A2 binds two molecules of flaviolin in its active site, and the 2-OH group of flaviolin is responsible for anchoring the substrate in the active site, while the 5-OH and 7-OH stabilize water molecules which are important for catalysis.

For the discovery of bacterial CYPs in sequenced genomes, genes encoding the CYP heme binding domain (FXXGXXXCXG) can be screened for the presence of a highly conserved threonine in the putative I-helix, which is proposed to be involved in CYP oxygen activation in most CYPs, and the conserved EXXR motif is located in the K-helix. The sequences of polypeptides containing all three

---

Table 1

| CYPs involved in the biosynthesis of microbial natural products. |
|---------------------------------------------------------------|
| **Hydroxylation of saturated C—H bonds**                      |
| **CYP name as annotated** | Protein identifier | Identity% | AA overlap |
| **AmphL** | 396 | AAK73504 | CYP107E | AAK73504 | 100% | 396 |
| **AziB1** | 401 | B4XV99.1 | CYP107-like | WP,018771011.1 | 34% | 388 |
| **ChmPl** | 407 | AAS97447 | CYP107B | AAS97447 | 100% | 407 |
| **EpiK** | 401 | AAB93855 | CYP162-like | WP,030872520.1 | 40% | 399 |
| **EryF** | 404 | Q00441.2 | CYP107A | WP,0099503971 | 100% | 404 |
| **EryK** | 391 | CYP113A | CYP113A-like | WP,009950895.1 | 100% | 391 |
| **HerG** | 422 | AEZ64507 | CYP107B-like | WP,016577953.1 | 48% | 404 |
| **MeI** | 456 | AAM67914 | CYP171A | AAM67914 | 100% | 456 |
| **MycCl** | 383 | BAC37023 | CYP105U | BAC37023 | 100% | 383 |
| **NcsB3** | 410 | AAS77997 | CYP15A | AAS77997 | 100% | 410 |
| **OxyD** | 394 | CDD33151 | CYP146A | CDD33151 | 100% | 394 |
| **Pikc** | 416 | B6T605 | CYP107L | AAC4105 | 100% | 416 |
| **PilD** | 399 | BA02272 | CYP107B-like | BA02272 | 100% | 399 |
| **PncC** | 399 | BAC8123 | CYP105P | WP,010981850.1 | 100% | 399 |
| **TyHi** | 436 | AAD41818 | CYP105U | AAD41818 | 100% | 436 |
| **TyI** | 417 | AAZ21341 | CYP107-like | AAA21341 | 100% | 417 |
| **ZbmVIlc** | 416 | AGC60779 | CYP185-like | WP,030613068.1 | 51% | 430 |
| **TtxC** | 395 | AAL36838.1 | CYP105A | WP,009733936.1 | 41% | 406 |

| **Dual function, hydroxylation and epoxidation**              |
|---------------------------------------------------------------|
| **GsfF** | 414 | BAJ16472 | CYP105A | BAJ16472 | 100% | 414 |
| **MycG** | 397 | BAA03672 | CYP107B | BAA03672 | 100% | 397 |
| **Fma-P450** | 400 | ABD19974 | – | – | – |
| **Fr9R** | 482 | ACIC32704 | CYP136-like | WP,022984814.1 | 32% | 455 |
| **PntD** | 299 | AOD58592 | – | – | – |
| **PntD** | 299 | AOS8592 | – | – | – |
| **TamI** | 413 | ADCT9647.1 | CYP107B-like | ADCT9647.1 | 100% | 413 |
| **TsrR** | 482 | AGN11891 | CYP136-like | BAP15297 | 33% | 455 |

| **Catalyze epoxidation**                                      |
|---------------------------------------------------------------|
| **Ausn30** | 1005 | AAM54108 | CYP102F | AAM54108 | 100% | 1005 |
| **CYP105A** | 401 | AAS97447 | CYP107B | AAS97447 | 100% | 401 |
| **EpiM** | 419 | Q09KZ4 | CYP167A | Q09KZ4 | 100% | 419 |
| **EpoC** | 425 | AAB938496 | CYP107B-like | WP,020576451.1 | 42% | 406 |
| **GohO3** | 416 | AAB93670 | CYP105-like | AAB93670 | 100% | 416 |
| **HrdR** | 409 | AAP5338 | CYP105-like | AAP5338 | 100% | 409 |
| **OleP** | 397 | BA02272 | CYP107B | BA02272 | 100% | 397 |
| **PinD** | 397 | CADC29532 | CYP107E | CADC29532 | 100% | 397 |
| **PinG** | 398 | CADC29532 | CYP105H | CADC29532 | 100% | 398 |
| **TamI** | 413 | ADC79647 | CYP107B-like | ADC79647 | 100% | 413 |

| **Oxidation of methyl group**                                 |
|---------------------------------------------------------------|
| **AmphN** | 399 | AAK73509 | CYP105H | AAK73509 | 100% | 399 |
| **NysN** | 398 | AAF71771 | CYP105H | AAF71771 | 100% | 398 |
| **C-C coupling**                                              |
|---------------------------------------------------------------|
| **DynOrf19** | 403 | ACH47070 | CYP107M | WP,015621195.1 | 51% | 398 |
| **HmtS** | 397 | CBZ42153 | CYP113A | WP,030604461.1 | 52% | 398 |
| **OxyB** | 398 | AAL90878 | CYP165B | WP,030610681.1 | 100% | 398 |
| **OxyC** | 406 | AAL90878 | CYP165B | WP,030610681.1 | 100% | 406 |
| **C-N-C coupling**                                            |
|---------------------------------------------------------------|
| **DynE10** | 400 | ACH47071 | CYP107B | WP,029893036.1 | 53% | 400 |
| **SpcN** | 390 | AGL96571 | CYP244A | AGL96571 | 100% | 390 |

Catalyzes the nitration using NO and O2

- **TtxE** | 406 | ELIP66108.1 | CYP14505XTX | ELIP66108.1 | 100% | 406 |

Oxidative decarboxylation

- **HmtN** | 419 | 4E2P_A | CYP113A | 4E2P_A | 100% | 419 |
| **HmtT** | 419 | 4G0V_A | CYP113A | 4G0V_A | 100% | 418 |
| **Mei-Orf4** | 398 | ADC45514 | CYP107-like | ADC45514 | 100% | 398 |
| **StaP** | 426 | ABF94389 | CYP245A | ABF94389 | 100% | 426 |
| **SpC** | 427 | AGL96575 | CYP245A | AGL96575 | 100% | 427 |

References:
73. X. Liu/Synthetic and Systems Biotechnology 1 (2016) 95–108

---

<sup>a</sup> CYP names as annotated at website: https://cyped.biocatnet.de/.

<sup>b</sup> Not found in database.
motifs can be further used as queries for BLAST searches of the GenBank non-redundant protein database (www.ncbi.nlm.nih.gov/BLAST) to identify their closest homologues in other organisms and tentatively assign the CYP proteins to subfamilies. Briefly, >40% of amino acid sequence identity places a CYP in the same family and >55% places it in the same subfamily.11 Besides GenBank, there is a CYPED database which scours Genbank by BLAST-searching and retrieving the CYP sequences to put in their database (https://cyped.biocatnet.de/).12-15 This database contains the most bacterial CYPs (Table 1) that are involved in the biosynthesis of bioactive natural products (Fig. 1). There is also another online CYP database (http://dmlelson.uthsc.edu/CytochromeP450.html), which contains 1042 bacterial CYP genes.

2.3. The logic for CYP pattern-based genome mining

There is evidence that CYPs can be used for genome mining for pre-mRNA spliceosome inhibitors. First, all current pre-mRNA spliceosome inhibitors are epoxide-containing natural products (Fig. 1A) whose biosynthesis gene clusters contain genes encoding putative CYP oxygenases or epoxidases. It has been reported that the presence of epoxide group is important in conferring activity to FR analogues.116,117 Second, CYPs may not directly be involved in the formation of some epoxide groups,117,118 but CYPs are present in the relevant biosynthetic gene clusters (Fig. 2). For example, in a recent metagenome mining report, all six epoxyketone gene clusters contain CYPs,119 indicating that CYP-containing gene clusters may be a rich source for the biosynthesis of epoxide compounds. Third, even antitumor activities of CYP-catalyzed compounds are related to different mechanisms. Small molecule screenings identified that oxaspiro compounds (the farnesyltransferase inhibitor manumycin A analogues) are pre-mRNA splicing inhibitors.120 In addition, a number of compounds such as the CYP-related non-epoxide compounds staurosporine (kinase inhibitor), are novel inhibitors of spliceosome assembly.121 Fourth, homologous CYPs can be identified through genome mining by a 50–60% similarity except for the 80%–90% rule for PKS and nonribosomal peptide synthetase (NRPS) domain searching. For example, BLAST analysis demonstrated a 50% identity between CYPs PlbD and HerG from pladienolide and herboxidiene (both are pre-mRNA spliceosome inhibitors), respectively.122 In the biosynthesis of FD-891, the CYP GfsF showed 57% identity to the CYP monooxygenase Mflv2418.123 Finally, the CYP gene can be used as a signature gene for gene cluster cloning and gene identification.125 However, CYP pattern should be considered because CYPs and other tailoring enzymes such as transferases and epoxidases act together on the released PKS or NRPS megasynthases.120,122-123 (Fig. 3).

2.4. Discovery of thailanstatins from Burkholderia sp.

Recently, the genus Burkholderia has attracted the attention of several research groups to employ different genome mining strategies for the study of bioactive natural products.124 Among those products, FR is a general spliceosome inhibitor discovered in 1992 from Pseudomonas sp. No. 2663.125-127 now identified as Burkholderia sp. FERM BP3421.128 There are three oxygenase activities encoded in FR gene cluster: (a) the flavin-dependent monooxygenase (FMO) domain in the last module of fr9GH, (b) the CYP encoded by fr9R, and (c) the Fe(II)/Ike-glykolate-dependent dioxygenase encoded by fr9P. The hemiketal FR is mainly biosynthesized through the epoxidation at C3 by the FMO domain of Fr9GH, the hydroxylation at C4 by the Fr9R, the hydroxylation at C1 by Fr9P and then decarboxylation.124 First, BLAST analysis shows that Fr9R has a high identity (82%) to TsrR in the sequenced genome of B. thailandensis MSMB43 (currently being described as “Burkholderia hynptydooensis”).128-129 Second, Fr9K is a key 3-hydroxy-3-methylglutaryl-CoA synthase (HCS) homologue–enzyme catalyzing the transfer of CH2=COO— from acyl-S-acyl carrier protein (ACP) to a β-ketothioester polyketide intermediate. Fr9K has a high identity (87%) to one of the enzymes (TskK) predicted from the genome of B. thailandensis MSMB43. Further, a regulator gene expression studies for optimization of production media, and targeted isolation and purification efforts toward diene compounds (UV235 nm), led to the discovery of three compounds called thailanstatins129 and a group of similar or identical compounds called spliceostatins130,131 (Fig. 4). Particularly, thailanstatin A contains a carboxyl group, which not only makes it more stable than FR in PBS solution, but also leads to increased activities compared to FR when carboxyl groups were esterified (with enhanced membrane permeability).130

2.5. Kutzneria species are potential resources of CYP-catalyzed compounds

In the Actinobacteria phylum, the genus Kutzneria is a minor branch of the Pseudonocardiaceae family, currently containing eight species (Fig. 5). Only aculeaximycin and kunzinerids are known to be produced by Kutzneria albida DSM 43870 and Kutzneria sp. 744, respectively.112,113 Genome mining of Kutzneria species suggests that they could produce compounds similar to the previously reported pre-mRNA spliceosome inhibitors and a group of interesting CYP catalyzed compounds. For example, the pladienolide biosynthetic gene cluster has only four genes (including the pldB CYP gene, and the putative epoxidase encoding gene pldD) besides the PKS genes.129 Inspired by the successful discovery of thailanstatins, pladienolide B CYP monooxygenase-based bioinformatics analysis was conducted and revealed that the Kutzneria sp. 744 genome contains the most homologous (78%) enzyme KUTG_06291 (EWM16617) when CYP PlbD sequence was used as input for homology search of the GeneBank (Fig. 6). Therefore, the genus Kutzneria was postulated to produce compounds similar to pladienolide. However, there are still a lot of gaps in the sequenced genome of Kutzneria sp. 744.

Kutzneria albida DSM 43870 is the only Kutzneria strain with a complete sequenced genome. So far, the K. albida genome (9.87 Mb) is among the largest actinobacterial genomes sequenced.131 Thus, Kutzneria albida DSM 43870 was selected for further phylogenetetic analysis of CYPs in parallel with a library of 44 CYPs originate from biosynthetic gene clusters of 32 bioactive natural products.71-97,108-109,114 The Kutzneria sp. 744 genome contains the most homologous (78%) enzyme KUTG_06291 (>99% identity (87%) to one of the enzymes (TstK) predicted from the genome of B. thailandensis MSMB43. Further, a regulator gene expression studies for optimization of production media, and targeted isolation and purification efforts toward diene compounds (UV235 nm), led to the discovery of three compounds called thailanstatins and a group of similar or identical compounds called spliceostatins (Fig. 4). Particularly, thailanstatin A contains a carboxyl group, which not only makes it more stable than FR in PBS solution, but also leads to increased activities compared to FR when carboxyl groups were esterified (with enhanced membrane permeability).130
Fig. 1. P450 related microbial natural compounds. Except for staurosporine, all the selected compounds are epoxide-containing compounds. Other epoxide compounds without known biosynthetic pathways, such as antitumor compound trapoxin, are not listed. Panel A: Natural products with known spliceosome inhibitory activities; Panel B: Natural products with other activities such as tubulin-binding compounds epothilone A, telomerase inhibitor griseorhodin A, RNA polymerase inhibitor tirandamycin and proteasome-inhibitor epoxomicin, etc.
ysis suggests that tirandamycin-like compounds may be produced through GC19 based on the 66% identity between KALB_3945 and TamI.

Unfortunately, the 27% identity between KALB_5792 (in GC33) and AziB1 (CYP in the azinomycin B biosynthetic gene cluster\(^{98}\)), and the 19% identity between KALB_5804 (in GC33) and EpoK (CYP in the epothilone biosynthesis gene cluster\(^{99}\)) make it unpredictable for the natural products biosynthesized by GC33.

Except for Kutzneria albida DSM 43870\(^{100}\), another top five strains, Actinoplanes sp. N902-109 (NC_021191.1), Micromonospora aurantiaca ATCC 27029 (CP002162.1), Streptomyces violaceusngeri Tu 4113 (CP002994.1), Streptomyces bingchengensis BCW-1 (CP0020471) and Streptomyces sp. PVA 94-07 (CM002273.1), were prioritized through CYP pattern-based genome mining of pladienolide biosynthetic pathway. These strains have the potential CYP pattern pathways to produce unknown compounds (Tables S2–S6).

### 3. Conceivable methods to translate biosynthetic pathways to natural products

This section covers the methods (Fig. 7) or tools that have been used or have the potential to be applied to translate biosynthetic pathways to natural products.

#### 3.1. Direct cloning and heterologous expression of biosynthetic gene clusters

Heterologous expression was primarily used to avoid the lack of genetic tools and regulatory complexity in native hosts via the use of strains that are more amenable to engineering. Direct cloning, genetic engineering and heterologous expression of microbial natural product biosynthesis pathways were reviewed in 2013.\(^{101}\) The heterologous expression methods can also be borrowed from those for the discovery of small molecules from metagenomics cosmid library.\(^{102,103}\) The approaches to cloning targeted gene clusters directly from genomic DNA include RecE-mediated homologous recombination,\(^{104}\) oriT-directed capture,\(^{105}\) transformation-associated recombination (TAR)\(^{106,107}\) to facilitate functional expression experiments and phageφBT1 integrase-mediated site-specific recombination.\(^{108}\) Among those, TAR cloning in Saccharomyces cerevisiae has been extensively applied to capture and express large biosynthetic gene clusters from environmental DNA samples.\(^{109,110}\) Recently, the TAR direct cloning approach was used to clone targeted whole gene cluster from rare actinomycetes and Gram-negative bacteria. By using TAR, the targeted gene cluster was cloned into plasmid pCAP01 by homologous recombination,\(^{111}\) orTDirectected capture,\(^{112}\) transformation-associated recombination (TAR)\(^{113,114}\) to facilitate functional expression experiments and phageφBT1 integrase-mediated site-specific recombination.\(^{115}\) Among those, TAR cloning in Saccharomyces cerevisiae has been extensively applied to capture and express large biosynthetic gene clusters from environmental DNA samples.\(^{116,117}\) Since the development of a TAR-based genetic platform allows for heterologous production of lipopeptides in different hosts, efforts have been focused on the development of the high throughput TAR capture method to express the pathway.\(^{118}\) Successful production of the desired products often requires an optimal relationship of timing and flux between primary and secondary cellular metabolism. Besides the use of genetic engineering for the direct cloning of a whole gene cluster, there has been considerable interest in the development of engineered bacterial strains for efficient heterologous production of secondary metabolites.\(^{119,120}\) For example, the deletion of a 1.4 Mb segment from the left subtelomeric region of the 9.02 Mb Streptomyces avermitilis genome.
resulted in the generation of two large-deletion mutants. By using these large-deletion mutants, i.e., *S. avermitilis* SUKA17 or 22, twenty of the entire biosynthetic gene clusters for secondary metabolites, including aminoglycosides, nonribosomal peptides, polyketides and terpenes were successfully expressed. It is worthy to note that the engineered hosts are not only useful for the production of exogenous secondary metabolites, but they also facilitate scale-up production and preparation of promising natural products due to their “clean” background.

3.2. Manipulation of the biosynthetic pathways

The recent approaches on the activation and up-regulation of microbial biosynthetic pathways for the discovery of natural products have been reviewed. Except for these methods, there are four types of research reports that may explain how to translate biosynthetic gene clusters to novel natural products by using synthetic biology methods. First, optimized bioactive compounds can be produced by engineering the biosynthetic operon and reconstituting the biosynthetic pathway. The group led by Müller and Brönstrup successfully deleted the gene in hydroxymalonyl-CoA biosynthetic operon of the bengamide biosynthetic gene cluster and inserted a promoter on the expression construct pBen32. Heterologous expression of the modified biosynthetic gene cluster led to the discovery of more potent compounds. Second, new compounds with increased activities are produced by inactivation of CYP pattern genes. Bills’ team works on the manipulation of pneumocandin biosynthetic pathway. They focused on the inactivation of three genes (GLP450-1, GLP450-2, and GLOXY1) and generated 13 different pneumocandin analogues that lack one, two, three, or four hydroxyl groups on 4R, 5R-dihydroxy-ornithine and 3S, 4S-dihydroxy-homotyrosine of the parent hexapeptide. Among them, seven analogues are previously unreported. Third, natural product analogues can be produced by creating promoter-driven tailoring enzyme constructs. Brady’s group created an *ermE* promoter cas-

Fig. 3. Pattern-based genome mining of homologous gene clusters of spliceosome inhibitors which contain CYPs. The biosynthetic gene clusters of thailanstatin (Accession no. KJ461964.1) and FR901464 (Accession no. JX307851.1), pladienolide (Accession no. AB435553.1) and herboxidiene (Accession no. JN671974.1) as well as staurosporine (Accession no. AB088119.3) were used for the search of homologous gene clusters by using antiSMASH 3.0.

---

[101] X. Liu/Synthetic and Systems Biotechnology 1 (2016) 95–108
Fig. 4. Genome mining protocol of *Burkholderia thailandensis* MSMB43 for thailanstatins/spliceostatins. Starting from the two key biosynthetic enzymes CYP Fr9R and 3-hydroxy-3-methylglutaryl-CoA synthase (HCS) Fr9K that catalyze the formation of the hemiketal hydroxyl group and the epoxide group of FR901464, the homologous enzymes and biosynthetic gene cluster of thailanstatin were predicted. The Reverse Transcriptional (RT)-PCR help to identify the optimum medium for the expression of regulatory gene *tstA*. The novel compounds were finally obtained through microbial fermentation, natural product isolation and structural elucidation by tracking of the characteristic UV absorption for the diene motif at UV235 nm.

Fig. 5. Phylogenetic tree for taxa of the genus *Kutzneria* and other antitumor compound-producing species (*Streptomyces* and *Burkholderia*). The tree was calculated from complete 16S rRNA gene sequences using the neighbor joining method, illustrating the genus *Kutzneria* position relative to other selected species. All species are type strains except for the strains *Kutzneria* sp. 744, *Burkholderia* sp., MSMB43 and MSMB121. The 16S RNA sequence of the pladienolide producer *Streptomyces platensis* Mer-11107 is not available, but DNA hybridization suggests that *Streptomyces platensis* Mer-11107 has 87% similarity to the type strain CGMCC4.1975. Percentages at nodes represent levels of bootstrap support from 1000 resampled datasets; Scale bar, 0.02 nucleotide substitutions per site.
Fig. 6. Strategy for the discovery of novel compounds from Kutzneria species. By starting with the structure of pladienolide B, and moving through phylogenetic analysis of the streptomycetes-related cytochrome P450s, Kutzneria species were selected for further analysis of P450-related polyketides. The genome sequencing of the closest strain of Kutzneria sp. 744 was finished on February, 2014 without detailed annotation and there are still a lot of gaps waiting to be closed.

Fig. 7. Schematic representation of methods to translate biosynthetic pathways to novel natural products.
sette, which was introduced to the upstream of the first ORF of the biosynthetic gene cluster of interest. Further introduction of the cosmid containing the constitutively expressed tailoring operon into S. toyocaensis ΔStal resulted in the production of three new glycopeptides. Fourth, stable compounds with increased titer are produced by engineering the CYP pathway. Researchers in Pfizer engineered a pAE-PF29 vector which enabled the overexpression of Fr9R encoded by CYP fr9R in Burkholderia sp. FERM BP-3421, and led to the enhanced production of stable thailanstatin A from splicestatin C.

3.3. Synthetic biology tool kits

Except for the four examples of genetic manipulations, synthetic biology tools to edit bacterial genomes have been reviewed by other researchers. The Clustered Regularly Interspaced Short Palindromic Repeats and Cas proteins (CRISPR-Cas) systems are composed of a powerful and broadly applicable set of tools to manipulate Streptomyces genomes. There are three research groups that worked on the gene deletion of Streptomyces species by using different CRISPR-Cas toolkits. The group led by Zhao developed a temperature-sensitive pCRISPomyces system, which is applicable to the genome editing of Streptomyces lividans 66, Streptomyces viridochromogenes DSM 40736 and Streptomyces albicus J1074. Lee and Webber’s group focused on the gene deletion efficiency of actinorhodin biosynthetic gene cluster in Streptomyces coelicolor A3 by an engineered CRISPR-Cas system with improved efficiency. Sun’s team worked on the discovery of a novel CRISPR-Cas system in combination with the counterselection system Coda(sm), and the D314A mutant of cytosine deaminase, to delete the actinorhodin biosynthetic gene cluster in Streptomyces coelicolor M145 genome.

4. The advantages and challenges of the above mentioned strategies

The advantages and challenges to generate a bioactive natural product library by CYP pattern-based genome mining as compared to the traditional strategies such as microbial fermentation with different media and chromatographic fractionation for different polarities can be summarized as follows.

4.1. Advantages

- CYP is a class of most important tailoring enzymes that can be explored and exploited for the structural diversification of bioactive natural products.
- Genome mining is an efficient method. The combination of PCR screening and genome mining will prioritize the cryptic gene clusters of interest for annotation and heterologous expression. This is more efficient when compared to the traditional conception of one strain many active compounds (OSMAC).
- Genome mining can be used for large scale and high throughput method to accelerate drug discovery, as exemplified by a four-year case study on genome mining of 10,000 actinomycetes for 19 novel phosphonic acid natural products.
- Both heterologous expression and activation of cryptic pathways can generate novel compounds that have unique biological activities but are difficult to access using synthetic chemistry.
- CRISPR-cas9 enabled genome editing tools are very efficient. First, compared to the 5% (2/38) efficiency in the traditional in-frame gene deletion method, the efficiency of the CRISPR-Cas9 systems can be 100% without unwanted side- or off-target effects. Second, CRISPR-Cas9 systems can overcome the employment of the counter selectable marker for the selection of double crossover mutants. Third, CRISPR-Cas9 systems can avoid the scar sequence left at the target site and realize multiple gene deletions. Fourth, CRISPR/Cas system provides new modularity, which can target the site of interest by inserting of a short spacer into a CRISPR array/sgRNA construct. The insertion can be achieved with high throughput using modern DNA assembly techniques. CRISPR/Cas9 genome editing can shorten the two-month period of previous gene deletion method to the current 1–2 week time course.

4.2. Main challenges

- How to prioritize the characterization of orphan biosynthetic gene clusters, and how to rapidly connect genes to biosynthesized small molecules will come with increase in DNA-sequenced data.
- Pathway manipulations in native hosts are subjected to the genetic complexity of the bacteria. For example, researchers have developed a procedure which takes about 10 days for the chromosomal knock out of Gram-negative bacteria B. pseudomallei by using the phes-gat cassette-assisted by plasmid pKaG2. However, this method is not applicable to the gene knockout of another strain called B. gladioli, which requires another pCR4Blunt-TOPO-based vector for the recombination assisted by plasmid pKD46.
- Up to now, Cas9 targeting is still limited by protospacer adja-
cent motif (PAM) sequences. Also, targeting efficiency is site-dependent.
- Although the development of genetic manipulation tools could greatly enhance the chances toward discovery and production of natural products, a major challenge in the process of microbial genome mining is to produce compounds in high titers.
- Bacterial genomes are publically available but the accessibility of strains of interest is limited due to the regulations of international or domestic properties. Decreasing DNA synthesis costs and advances in DNA assembly could help to solve the issue with limited material access. A larger fraction of strains that are isolated in research labs worldwide will be the future challenge.

5. Conclusion

The current review provides strategies on the discovery of a group of epoxide-containing natural products (Fig. 8), highlights (1) CYP and its pattern enzyme-based genome mining as a guidance for the generation of a diversified natural product library, (2) direct cloning and heterologous expression of biosynthetic pathways and genomic manipulation methods and tools will translate the selected biosynthetic gene clusters to the bioactive natural product library. It is worthy to note that tailoring reactions play important roles on diversifications of scaffolds such as polyketide, peptide and hybrid polyketide-peptide backbones, and often tailoring enzyme patterns such as CYPs, ligases, Cyclases, ketoreductases, transferases, and oxygenases can be used for genome mining. This advanced genome mining will avoid the de-replications of biosynthetic gene clusters to quickly identify the annotated biosynthetic gene clusters from the vast pool of sequenced bacterial genomes. It is predicted that we will be sequencing genomes for pennies as nearly as 2020 (https://youtu.be/j88APstUcp4). Synthetic biology tools pioneered by different researchers will be continuously developed toward a high-throughput potential due to the increasing numbers of sequenced bacterial genomes. This review also points out the potential mechanisms and diversity of CYPs in the microbial biosynthesis of natural product antitumor agents. The Kutzneria strains, as rare actinobacterial species, can be explored and exploited for CYP catalyzed compounds, and can be used as a rich resource for the diversification of microbial CYPs.
Biological resource centres. Underpinning the future of life science and biotechnology.

Chembiochem
Du L, King JB, Morrow BH, Shen JK, Miller AN, Cichewicz RH. Quinn RJ, Carroll AR, Pham NB, Baron P, Palframan ME, Suraweera L, et al. Diversifying natural product libraries for high-throughput screening and rapid drug discovery. J Nat Prod 2008; 71:1095–8.

Hsu Y, Potts MB, Colosimo D, Herrera-Herrera ML, Legako AG, Youssufuddin M, et al. Discopyrroles A–D: isolation, structure determination, and synthesis of potent migration inhibitors from Bacillus hunanensis. J Am Chem Soc 2013; 135:13387–92.

Du L, King JB, Morrow BH, Shen JK, Miller AN, Cichewicz RH. Diarylpyrrolenitrolide metabolite obtained from a Prunus thurifera isolate procured using an unconventional cultivation approach. J Nat Prod 2012; 75:1819–23.

Du L, Robles AJ, King JB, Powell DR, Miller AN, Moberry SL, et al. Crowdsourcing natural products discovery to access uncharted dimensions of fungal metabolite diversity. Angew Chem Int Ed Engl 2014; 53:804–9.

Quinn RJ, Carroll AR, Pham NB, Baron P, Palfraen ME, Suraweara L, et al. Developing a drug-like natural product library. J Nat Prod 2008; 71:464–8.

Kim MK, Kim S, Park SJ, Lee H, Kim TS, Jung SY, et al. Abstract 1702: High-throughput cell-based screening of drug library identifies albendazole as a sensitizer with combination of bortezomib for treatment multiple myeloma. Cancer Res 2014; 74:1702. http://www.selleckchem.com/.

Starosta SA, Volynets GP, Lukhashov SS, Gorbatiuk OB, Golub AG, Bdzhola VG, et al. Identification of apoptosis signal-regulating kinase 1 (ASK1) inhibitors among the derivatives of benzothiazol-2-yl-3-hydroxy-5-phenyl-1,5-dihydro-pyrrol-2-one. Bioorg Med Chem 2015; 23:2489–97. http://www.otavachemicals.com.

Venkataraman SK, Moores A, Stone A, Hurst A, Mikell JR, Moraes RM, et al. Rays and planar atoms: generation of botanical extract libraries. Pflanz Med 2009; 75:PA36. https://chromadex.com/Default.aspx.

Camp DJ, Davis RA, Campitelli M, Ebdon J, Quinn RJ. Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod 2012; 75:72–81.

Gaudenzio SP, Pereira F. Dereplication: racing to speed up the natural products discovery process. Nat Prod Rep 2015; 32(6):779–810.

Friedenhausen A, Derrien C, Casmann E, An MS. MS library on an ion-trap instrument for efficient dereplication of natural products. Different fragmentation patterns for [M + H]+ and [M + Na]+ ions. J Nat Prod 2005; 68:385–91.

Yang JY, Sanchez LM, Rath CM, Liu XT, Boudreau PD, Bruns N, et al. Molecular networking as a dereplication strategy. J Nat Prod 2013; 76:1686–99.

Bian J, Song F, Zhang L. Strategies on the construction of high-quality microbial natural product library – a review. Acta Microbiol Sin 2008; 48:1132–7.

Carter GT. Natural products and Pharma 2011: strategic changes spur new opportunities. Nat Prod Rep 2011; 28:1783–9.

Walsh CT. Combinatorial biosynthesis of antibiotics: challenges and opportunities. Chembiochem 2002; 3:125–34.
coli represents a promising natural product exploratory platform. *ACS Synth Biol* 2015;4:414–20.

142. Yamanaka K, Reynolds KA, Kersten RD, Ryan KS, Gonzalez DJ, Nizet V, et al. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* 2014;111:1957–62.

143. Kouprina N, Larionov V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nat Proto* 2008;3:371–7.

144. Du D, Wang L, Tian Y, Liu H, Niu G. Genome engineering and direct cloning of antibiotic gene clusters via phage varphiBT1 integrate-mediated site-specific recombination in *Streptomyces*. *Sci Rep* 2015;5:8740.

145. Feng Z, Kalifidas D, Brady SF. Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proc Natl Acad Sci U S A* 2011;108:12629–34.

146. Feng Z, Kim JH, Brady SF. Flusotatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *J Am Chem Soc* 2010;132:11902–3.

147. Kim JH, Feng Z, Bauer JD, Kalifidas D, Calle PY, Brady SF. Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* 2010;93:833–44.

148. Martinez A, Kolvek SJ, Yip CL, Hopke J, Brown KA, MacNeil IA, et al. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental DNA libraries and detecting heterologous natural products in multiple expression hosts. *Appl Environ Microbiol* 2004;70:2452–63.

149. Feng Z, Wang L, Rajski SR, Xu Z, Coevert-LeGal MF, Shen B. Engineered production of iso-migrastatin in heterologous *Streptomyces* hosts. *Bioorg Med Chem* 2009;17:2147–53.

150. Pfeifer B, Hu Z, Licari P, Khosla C. Process and metabolic strategies for improved production of *Escherichia coli*-derived 6-deoxyerythronolide B. *Appl Environ Microbiol* 2002;68:3287–92.

151. Komatsu M, Uchiyama T, Omura S, Cane DE, Ikeda H. Genome-minimization Streptomyces host for the heterologous expression of secondary metabolism. *Proc Natl Acad Sci U S A* 2010;107:2646–51.

152. Komatsu M, Komatsu K, Koiwai H, Yamada Y, Kozone I, Izumikawa M, et al. Engineered *Streptomyces avermitilis* host for heterologous expression of biosynthetic gene cluster for secondary metabolites. *ACS Synth Biol* 2013;2:384–96.

153. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat Rev Microbiol* 2015;13:509–23.

154. Zarins-Tutt JS, Barberi TT, Gao H, Mears-Sprag A, Zhang L, Newman DJ, et al. Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Nat Prod Rep* 2015;33:54–72.

155. Wenzel SC, Hoffmann H, Zhang J, Debussche L, Haag-Richter S, Kurz M, et al. Production of the bengamide class of marine natural products in *Myxobacteria*: biosynthesis and structure-activity relationships. *Angew Chem Int Ed Engl* 2015;54(1):15560–4.

156. Li Y, Chen L, Yue Q, Liu X, An Z, Bills GF. Genetic manipulation of the pneumocandin biosynthetic pathway for generation of analogues and evaluation of their antifungal activity. *ACS Chem Biol* 2015;10:1702–10.

157. Eustaquio AS, Chang LP, Steele GL, O’Donnell CJ, Koehn FE. Biosynthetic engineering and fermentation media development leads to gram-scale production of spliceostatin natural products in *Burkholderia* sp. *Metab Eng* 2015;33:67–75.

158. Kim E, Moore BS, Yoon YJ. Reinvestigating natural product combinatorial biosynthesis with synthetic biology. *Nat Chem Biol* 2015;11:649–59.

159. Cobb RE, Wang Y, Zhao H. High-efficiency multiplex genome editing of *Streptomyces* species using an engineered CRISPR/Cas system. *ACS Synth Biol* 2015;4(6):723–8.

160. Tong Y, Charusanti P, Zhang L, Weber T, Lee SY. CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS Synth Biol* 2014;3(9):1020–9.

161. Zeng H, Wen S, Xu W, He Z, Zhai G, Liu Y, et al. Highly efficient editing of the actinorhodin polyketide chain length factor gene in *Streptomyces coelicolor* M145 using CRISPR/Cas9-GoldA(sm) combined system. *Appl Microbiol Biotechnol* 2015;99:10575–85.

162. Tang X, Li J, Millan-Aguinaga N, Zhang JJ, O’Neill EC, Ugalde JA, et al. Identification of thieteronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* 2015;10:2841–9.

163. Kang Y, Norris MH, Wilcox BA, Tuanphok A, Kees PS, Hoang TT. Knockout and pullout recombineering for naturally transformable *Burkholderia thailandensis* and *Burkholderia pseudomallei*. *Nat Protoc* 2011;6:1085–104.

164. Moebius N, Ross C, Scherlach K, Rohn B, Roth M, Hertweck C. Biosynthesis of the respiratory toxin bongkrekic acid in the pathogenic bacterium *Burkholderia gladioli*. *Chem Biol* 2012;19:1164–74.

165. Lee SY, Kim HU, Park JH, Park JM, Kim TY. Metabolic engineering of microorganisms: general strategies and drug production. *Drug Discov Today* 2009;14:78–88.

166. Overmann J. Significance and future role of microbial resource centers. *Syst Appl Microbiol* 2015;38:258–65.