Breaking the silence: new strategies for discovering novel natural products
Hengqian Ren1,*, Bin Wang2,* and Huimin Zhao1,2,3

Natural products have been a prolific source of antibacterial and anticancer drugs for decades. One of the major challenges in natural product discovery is that the vast majority of natural product biosynthetic gene clusters (BGCs) have not been characterized, partially due to the fact that they are either transcriptionally silent or expressed at very low levels under standard laboratory conditions. Here we describe the strategies developed in recent years (mostly between 2014–2016) for activating silent BGCs. These strategies can be broadly divided into two categories: approaches in native hosts and approaches in heterologous hosts. In addition, we briefly discuss recent advances in developing new computational tools for identification and characterization of BGCs and high-throughput methods for detection of natural products.

Addresses
1 Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States
2 Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States
3 Departments of Chemistry and Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

Corresponding author: Zhao, Huimin (zhao5@illinois.edu)
* These authors contributed equally.

Current Opinion in Biotechnology 2017, 48:21–27
This review comes from a themed issue on Pharmaceutical biotechnology
Edited by Chu-Young Kim and Tiangang Liu

http://dx.doi.org/10.1016/j.copbio.2017.02.008
0958-1669/© 2017 Elsevier Ltd. All rights reserved.

Introduction
Natural products are a prolific source for drug discovery. Over the past 30 years, a large fraction of anticancer drugs and anti-infectives came from either natural products or derivatives of natural products [1]. Moreover, due to an increase in drug resistance, there is an ever-increasing need for new drugs [2,3]. However, since the Golden Age of natural product discovery in the 1950s–1960s, traditional methods have only yielded known natural products unfortunately, indicating a need for new methods that can discover novel natural products. Thanks to advances in DNA sequencing technologies and bioinformatics, researchers have become increasingly aware of the vast untapped biosynthetic capacity of microorganisms. Next generation sequencing can generate genome sequences in a fast and inexpensive way, while computational tools such as antibiotics & Secondary Metabolite Analysis SHeLL (antiSMASH) [4] can identify biosynthetic gene clusters (BGCs) from genome sequences in batches. Currently millions of putative BGCs have been identified, which have potential to produce a vast number of novel natural products [5*]. Moreover, in recent years, researchers have identified numerous BGCs from human microbiota. Because of potential microbiota–host interactions, these BGCs might be more likely to produce compounds related to human health and disease [6].

Despite the existence of numerous putative BGCs, only a tiny fraction of them have been characterized to date, and one of the key reasons is that in most cases BGCs are expressed poorly or not at all in their native hosts due to tight regulation under conventional laboratory culture conditions. Because of their tight regulation, most BGCs will not be activated unless special environmental cues exist. These BGCs are commonly known as ‘cryptic pathways.’ In order to activate those cryptic pathways, many innovative approaches have been developed and comprehensively reviewed elsewhere [5**,7].

A typical workflow for a modern natural product discovery strategy begins with identification of silent BGCs using bioinformatics tools, followed by activation of target BGCs using synthetic biology strategies and structural characterization of the corresponding natural products (Figure 1). In this review, we will provide a brief update on the approaches for activating cryptic pathways that were developed in the last two years. In addition, we will discuss recent progress in the development of computational tools and product detection methods for genome mining.

Computational tools for identification of BGCs
Based on the implementation strategies, computational tools for identification of BGCs mainly fall into two groups. The first group is rule-based tools, such as ClusterScan, NP.searcher, NaPDoS, SEARCHPKS, and SMURF for mining polyketide synthase/non-ribosomal peptide synthetase (PKS/NRPS), BAGEL for RiPPs, and antiSMASH [8]. These tools use predefined rulesets to identify signature genes that are associated with biosynthetic pathways. As a result they can predict known
classes of BGCs with high confidence and positive rates but cannot readily identify BGCs that use completely different enzymatic mechanisms. Notable among them is antiSMASH, which in its 3rd version can detect 44 different classes of BGCs, and, upon integration with ClusterFinder which uses a rule-independent algorithm further provides alternative options for automatic genome-scale predictions of BGCs of both known and novel classes [4,9]. The second group is rule-independent tools, such as ClusterFinder that predicts candidate BGCs via calculating the frequencies of Pfam domains that are commonly found in BGCs, thus those genomic regions rich in such Pfam domains hold more possibilities as candidates even in the total absence of signature genes [9]. By contrast, EvoMining takes advantage of phylogenomic analysis-based evolutionary logic to identify functionally diverged paralogs of primary metabolic enzymes, and presumes that the additional copies are repurposed for secondary metabolite biosynthesis [10]. These rule-independent tools can identify new BGC classes; but at this stage it is important to reduce false-positive rate such that new rules can emerge. A newly established website called secondary metabolite bioinformatics portal (SMBP) provides information on all aspects of secondary metabolite bioinformatics, enabling quick access to those computational tools and databases [8].

**Activation of silent BGCs in native hosts**

Actinomycetes, especially the genus *Streptomyces*, harbor dozens of BGCs per genome but control their expressions strictly [11]. Therefore, it remains an overwhelming challenge to activate these BGCs and identify the associated natural products if they are transcriptionally silent. Two main strategies have been developed to activate silent BGCs, either to unlock the suppression of BGC gene expression in the native hosts, or to directly bypass the regulation system by refactoring and reconstructing controlling elements in BGCs in heterologous hosts.

Genetic manipulation in native hosts takes advantages of the intact elements in BGCs, for example, precursor supply and end-product export, but it needs to break through the regulatory system first. Overexpressing positive regulator genes and deleting negative regulator genes are two main approaches used to activate transcription. For example, a new biaryl polyketide venemycin and two known ansatrienins in *Streptomyces venezuelae* and *Streptomyces* sp. XZQH13 were discovered when the large ATP-binding regulators of the LuxR family (LAL) transcriptional activator genes, *cemR* and *astG1* were constitutively expressed, respectively [12,13]. Moreover, two novel chattamycins were discovered when overexpressing an atypical response regulator (ARR) family regulator gene, *chal* [14]; and overexpression of a LuxR family activator gene *vamI* led to production of three new naphthalenic octaketide ansamycins [15]. For the jadomycin and guadinmycin biosynthetic pathways, however, no significant effects were observed when ARR family activator genes were overexpressed [16*]. This approach is proved to be pathway-specific, thus the efficacy varies case by case. However, due to its simplicity, this approach may be further developed for high throughput activation of silent BGCs on a strain level. In addition, overexpression of the homologs of the target activator genes may provide an alternative strategy as indicated in the activation of a mureidomycin BGC in *Streptomyces roseosporus* [17].

Another approach to activate cryptic BGCs is to artificially knock-in a strong promoter upstream of the target BGC to initiate the expression of the biosynthetic machinery, and thus synthesis of the corresponding natural product (Figure 2). Traditional promoter knock-in strategies based on homologous double-crossover recombination are often time-consuming and labor-intensive especially when applied to streptomycetes [18,19]. Recently we developed a highly efficient clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (Cas9) system for multiplex genome editing in streptomycetes, either targeting at precise gene deletion [20] or promoter knock-in [21*]. By applying this one-step strategy to five *Streptomyces* species, multiple silent BGCs belonging to different classes were activated, among which a novel pentangular polyketide from *Streptomyces viridochromogenes* was characterized [21***].

---

**Figure 1**

Overview of the workflow for a modern natural product discovery strategy. Silent BGCs are first identified from genomic sequences by bioinformatics tools and then activated by different synthetic biology strategies. Products are characterized by high-throughput methods and the data obtained can further guide the next round genome mining.
Workflow of the promoter knock-in strategy. Golden Gate and Gibson assembly are used to insert sgRNA (fusion of the crRNA and tracrRNA) and homology arms to a vector harboring a codon-optimized SpCas9 gene to yield knock-in plasmids which are then transformed into the native host. The codon-optimized SpCas9 cuts specific DNA sequence guided by sgRNA, and then inserts strong promoters into the chromosome via homologous recombination. Serial insertion of strong promoters upstream of each operon in a silent BGC can eventually result in a fully engineered gene cluster. Different combinations of engineered operons can also produce intermediates which provide important information on the biosynthetic mechanism. Fermentation followed by comparative analysis of metabolites by LC–MS or NMR enables the characterization of the corresponding natural products.

Considering the common multi-operon architecture in BGCs, this strategy may be used to iteratively knock-in promoters and activate the silent full-length BGCs, while different combinations of activated operons (by promoter insertion) can also provide insights into the biosynthetic mechanism (Figure 2). A system called CRISPRi was recently applied to Streptomyces coelicolor, which used a catalytically dead Cas9 (dCas9) to interfere with gene expression in a sequence-specific manner [22], and thus can be potentially applied to repress transcription of negative/positive regulatory genes for discovery of natural products from silent BGCs. Other tools such as the in vitro DNA editing system [23] and the transformation-associated recombination (TAR) method [24*] were also established based on the versatile CRISPR/Cas9 system, which further advances technological development in natural product discovery. Successful application of the CRISPR/Cas9 system in fungi has also been reported [25] and reviewed previously [26].

Other promising approaches to activate silent BGCs, such as one-strain-many-compounds (OSMAC) [27], co-culturing [28], ribosome engineering [29], and the use of elicitors [30,31], are all applicable to a wide range of hosts and may be scalable. However, the resulting host physiological disturbances are global, which make comparative metabolic profiling sometimes very difficult. Another promising approach named reporter-guided mutant selection (RGMS) employs genome-scale mutagenesis to create genetic diversity, and then directly selects desired target-activated mutants after double-reporter screening [16*]. This approach possesses a pathway-specific trait to either activate silent BGCs or improve production of weakly expressed BGCs. However, it also has built-in limitations, for example, requiring hosts to be kanamycin and hygromycin sensitive, as well as having no background yellow color.

**Activation of silent BGCs in heterologous hosts**

Emerging synthetic biology tools have enabled researchers to activate silent BGCs in heterologous hosts. The tight regulation of silent BGCs can be relieved by either direct cloning or refactoring using well-characterized promoters. Heterologous hosts possess significant growth advantages over most native hosts and therefore make activation of BGCs from unculturable organisms possible.

One of the most straightforward approaches to activate silent BGCs in a heterologous host is direct cloning. A major obstacle in this approach is the difficulty of handling large DNA fragments. Recently researchers have
developed multiple synthetic biology tools that leverage homologous recombination (HR) to specifically capture the gene cluster of interest. Transformation-associated recombination (TAR) in *Saccharomyces cerevisiae* has been utilized for cloning large genomic fragments [32]. Random genomic DNA fragments are co-transformed with an excess of the TAR cloning vector into *S. cerevisiae*. Since the cloning vector shares homologous regions to the target DNA sequences at both ends, once the cell gets both the cloning vector and the genomic fragment with the target DNA region, recombination will occur in high efficiency and result in the formation of a circular yeast artificial chromosome (YAC) molecule with the target sequence. This approach was successfully used to clone DNA fragments up to 250 kb from complex genomes [33]. The entire procedure takes approximately 2 weeks and generates positive clones with 1–5% yield. After further refinement for direct capture of gene clusters and discovery of taromycin A in *S. coelicolor* [34*], Moore and co-workers successfully extended this method to the heterologous expression of BGCs in *Bacillus subtilis* and *Escherichia coli* (Figure 3) [35,36]. Moreover, by coupling with a target-directed genome mining process, a distinct group of thiotetronic acid natural products was discovered [37].

In addition to *S. cerevisiae*, *E. coli* can also be used for direct cloning of BGCs when assisted by phage recombination systems, such as the lambda Red system and the ET recombination system (Figure 3) [38,39]. By using the full length RecET in *E. coli*, Fu et al. discovered increased HR efficiency between two linear DNA molecules [40]. They then cloned all 10 megasynthetase gene clusters from *Photobacterium luminescens* and identified novel metabolites lumimycin A and lumimide A/B. Similar to TAR, the application of RecET has afforded the heterologous production of disorazol and salinomycin in *Myxococcus xanthus* and *S. coelicolor* respectively [41,42]. In addition, a cryptic NRPS gene cluster from *Paenibacillus lacae* was cloned and activated by heterologous expression in *E. coli*, which produced a novel compound, sevadicin [43]. Recently this method was optimized by coupling with the lambda Red system for convenient modifications of the plasmid [44].

One of the major problems associated with the above-mentioned direct cloning methods is the low yield of positive clones, which is caused by the nonspecific targeting of random genomic fragments. Based on the fact that the CRISPR/Cas9 system can cleave at almost any site, Jiang and co-workers developed a Cas9-assisted targeting of chromosome segments (CATCH) method for direct cloning of BGCs (Figure 3) [45,46*]. The target BGC is cleaved from intact genomic DNA by Cas9 in *E. coli* and then ligated with a cloning vector through Gibson assembly. The whole procedure takes approximately eight hours over several days and yields positive clones varying from 20% (for 100 kb inserts) to 60% (for 50 kb inserts).

There are other direct cloning methods that rely on site-specific recombination reactions. For example, Hu et al. developed a Cre/loxP-based method [47]. The loxP sites are first integrated into the boundary regions of the target BGC with elements that are necessary for plasmid replication. Then the Cre recombinase is expressed *in vivo* and the whole region flanked by loxP is circularized as a plasmid. The resulting plasmid is isolated and transformed into *E. coli*. Besides using recombination *in vivo*, the circularized BGC can also be generated by recombination catalyzed by phiBT1 integrase *in vitro* in a very similar manner [48].

In addition to direct cloning, refactoring is another strategy to activate silent BGCs in heterologous hosts. Refactoring usually involves replacement of original regulatory elements and sometimes can be coupled with direct cloning. For example, the taromycin gene cluster was not functional immediately after direct cloning, but was
activated after knockout of the negative regulator encoded in the gene cluster [34**]. The more commonly used refactoring strategy is to replace the native promoters by the well-characterized constitutive promoters. One way to accomplish this is through HR in \textit{S. cerevisiae} [49]. The target BGC is first cloned into a shuttle vector and then co-transformed with promoter cassettes into \textit{S. cerevisiae}. Since each promoter cassette has the homologous sequences to their corresponding insertion sites in the target BGC as well as different selection markers, the promoters will be inserted into the designated sites by HR and positive clones can be selected in the appropriate medium. By using this method, an indolotryptoline BGC was activated, which led to the discovery of lazaramides A and B. This strategy was further optimized by using CRISPR/Cas9, which generates DNA double-strand breaks at the promoter insertion sites to enable one-step insertion of multiple promoters with a single selection marker or no marker and therefore increases the HR efficiency [24*]. Another strategy to refactor BGCs is through DNA assembly, which builds up the whole construct including the biosynthetic genes, promoters and other elements from linear DNA fragments. Previously we developed a plug-and-play refactoring workflow based on DNA assembly and successfully activated a silent polycyclic tetramate macrolactam BGC [50,51]. For more details about this strategy, readers are referred to recent reviews [52,53].

**High-throughput product detection**

Besides activation of silent BGCs, another challenge for discovery of natural products through genome mining is how to detect the product of the corresponding BGC, specifically in a high-throughput manner. Because of advances in the understanding of biosynthetic mechanisms of natural products and mass spectrometry, many computational tools have been developed to connect the BGC to its product based on MS/MS data. For example, an automated genomes-to-natural products platform (GNP) was recently developed to predict, combinatorialize and identify the product of modular gene clusters (PKS and NRPS) from LC–MS/MS of crude extracts [54**]. In order to match the polyketides and nonribosomal peptides to their corresponding BGCs, they further developed generalized retro-biosynthetic assembly prediction engine (GRAPE) and global alignment for natural products cheminformatics (GARLIC) [55]. Besides processing the BGCs and MS/MS spectra individually, a new strategy was developed to group BGCs and MS/MS spectrums into gene cluster families (GCFs) and molecular families (MFs) by similarity [56]. The relationships between GCFs and MFs can be utilized to aid in genome mining of unsequenced organisms. A plethora of MS/MS-based computational tools such as NRPQuest and RiPPquest have been developed, which were reviewed elsewhere [57,58]. Although most of these computational tools are limited to the analysis of peptide natural products, it can be expected that such MS/MS-based approaches will be extended to other natural products in the future. Recently, the Global Natural Products Social Molecular Networking (GNPS) was set up, which encourages researchers around the world to share their MS data online and can therefore further accelerate the identification of novel natural products from MS data [59]. Besides MS-based methods, improvements in NMR probes, experiments, data mining and databases have also made NMR based high-throughput detection of natural products possible, which was reviewed elsewhere [60,61].

**Conclusions and future perspectives**

Multiple new approaches for activating silent BGCs were developed in the last two years. These new approaches were accomplished with varying degrees of success, which provide room for further refinement and improvement. Approaches in the native hosts can avoid precursor availability issues or protein folding related issues, but are limited to culturable microorganisms. Most genomics driven approaches require the native hosts to be genetically trackable. In contrast, greater DNA manipulation freedom can be obtained when the target silent BGC is activated in a heterologous host, but currently the choice of the appropriate host organism still largely depends on previous studies, which may or may not afford the most appropriate host. Therefore, in addition to discovering more novel natural products, more knowledge about secondary metabolism is also needed in order to help researchers optimize the existing approaches and develop more general and scalable approaches.

Another major challenge in natural product discovery is to increase the throughput. Currently the characterization of BGCs cannot keep pace with the rate of BGC identification. Most of the approaches for activating silent BGCs described in this review require extensive genetic manipulations. However, with emerging DNA manipulation tools and advances in automated synthetic biology, it appears that this issue could be solved in the near future. In addition, many MS-based methods have been developed to date. Although most of those methods are limited to peptide natural products, they provide a new avenue for high-throughput MS-based product detection. We envision a new natural product discovery paradigm in which the BGC identification algorithm will be coupled with high-throughput pathway activation and product detection methods. Such paradigm will lead to discovery of numerous novel natural products with important biological activities.

**Acknowledgements**

We gratefully acknowledge financial support from the National Institutes of Health (GM077596) and the Visiting Investigator Programme of Agency for Science, Technology and Research of Singapore.
References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:
  ● of special interest
  ●● of outstanding interest


This is a comprehensive review on various approaches for activating silent BGCs.


This work alleviated the reporter method for activation of silent gene clusters and used a double-reporter system, xyle-neo, to lower the false positive rate dramatically. In principle, the target cluster activated mutants can be directly selected after the double-reporter screening.


24. Kang HS, Charlop-Powers Z, Brady SF: Multiplexed CRISPR/Cas9- and TAR-mediated promoter engineering of natural product biosynthetic gene clusters in yeast. ACS Synth Biol 2016, 5:1002-1010. This work combined CRISPR-Cas9 and TAR to enable efficient large gene cluster assembly and multiplexed promoter engineering in a single round. It also provides a mCRISSTAR python package and Webapp for transcriptional activation of complex natural product BGCs.


Activating silent biosynthetic gene clusters Ren, Wang and Zhao 27


This is the first application of TAR in natural product discovery.


This paper describes a direct cloning method based on the in vitro cleavage of a target BGC from genomic DNA using CRISPR/Cas9.


This paper describes a HR-based method for replacing promoters of a silent BGC on the shuttle vector.


The authors of this paper established a genomes-to-natural products (GNP) platform for prediction, prioritizing and identifying the products of modular gene clusters (PKSs and NRPSs) from LC–MS/MS data of crude extracts.


