SCIENCE CHINA
Life Sciences

SPECIAL TOPIC: Biosynthesis and regulation of secondary metabolites in microorganisms
• RESEARCH PAPER •
Progress of Projects Supported by NSFC

July 2013 Vol.56 No.7: 609–618
doi: 10.1007/s11427-013-4506-0

Assembly and features of secondary metabolite biosynthetic gene clusters in *Streptomyces ansochromogenes*

ZHONG XingYu[1,2], TIAN YuQing[1*], NIU GuoQing[1] & TAN HuaRong[1*]

[1]*State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*

Received April 25, 2013; accepted May 28, 2013

A draft genome sequence of *Streptomyces ansochromogenes* 7100 was generated using 454 sequencing technology. In combination with local BLAST searches and gap filling techniques, a comprehensive antiSMASH-based method was adopted to assemble the secondary metabolite biosynthetic gene clusters in the draft genome of *S. ansochromogenes*. A total of at least 35 putative gene clusters were identified and assembled. Transcriptional analysis showed that 20 of the 35 gene clusters were expressed in either or all of the three different media tested, whereas the other 15 gene clusters were silent in all three different media. This study provides a comprehensive method to identify and assemble secondary metabolite biosynthetic gene clusters in draft genomes of *Streptomyces*, and will significantly promote functional studies of these secondary metabolite biosynthetic gene clusters.

secondary metabolite, gene cluster, assemble, features, draft genome

Citation: Zhong X Y, Tian Y Q, Niu G Q, et al. Assembly and features of secondary metabolite biosynthetic gene clusters in *Streptomyces ansochromogenes*. Sci China Life Sci, 2013, 56: 609−618, doi: 10.1007/s11427-013-4506-0

Streptomyces are the most abundant source of secondary metabolites (SMs), including antibiotics and other bioactive compounds. Bioactive compounds were discovered mainly by using traditional trial-and-error methods. In recent years, great effort has been devoted to activate the cryptic SM biosynthetic gene clusters, which were revealed by genome mining of *Streptomyces*. In the model organism *Streptomyces coelicolor* A3(2), at least 29 putative gene clusters for SMs biosynthesis were detected [1], and 37 such gene clusters were found in the industrial organism *Streptomyces avermitilis* [2]. A large number of SM biosynthetic gene clusters were found to be distributed widely in other *Streptomyces* genomes and the majority of them are species specific [3]. Among the gene clusters detected, only a few were known to be responsible for the biosynthesis of specific SMs, while the vast majority is of unknown functions. Various genomic-based approaches have been devised and have led to the discovery of novel bioactive compounds [4–6].

The discovery of natural SM products via genomic-based approaches is largely dependent on the identification and annotation of SM biosynthetic gene clusters through *in silico* analysis. Several bioinformatic programs have been proved to be successful in the prediction of SM biosynthetic gene clusters with complete genome sequences [7–12]. Among them, antiSMASH is a comprehensive bioinformatic tool that can rapidly identify and annotate all types of known SM biosynthetic gene clusters with detailed information [12].

The number of complete genomes of *Streptomyces* is reported to be 12, which is less than 10% of the total number of *Streptomyces* genomes [13]. Most of these genomes are draft genomes. The number of complete *Streptomyces* ge-

*Corresponding author (email: tanhr@im.ac.cn; tianyq@im.ac.cn)

nomes is limited due to their high G+C content property, which results in shorter reads and much higher error rates than other bacteria whose G+C content is low. In addition, it is very difficult to completely assemble the reads generated by shotgun sequencing because of the relative large size of *Streptomyces* genomes (normally over 8 Mb). In contrast, draft genome sequences of *Streptomyces* can be obtained simply by assembling the reads generated from shotgun sequencing using automated genome assembly packages [14–16]. The major concern is that this approach results in a large number of contigs (typically from hundreds to thousands), and SM biosynthetic gene clusters might be split into several contigs. It is of great importance to devise methods to identify and assemble complete SM biosynthetic gene clusters based on draft genome sequences.

In this paper, a comprehensive antiSMASH-based method was employed to assemble the SM biosynthetic gene clusters in the draft genome of a nikkomycin producer, *S. ansochromogenes*. Transcriptional profiles of core genes selected in all SM biosynthetic gene clusters were examined in three different fermentation media and features of several gene clusters were presented. Our method can be applied to draft genome sequences of other actinomycetes or fungi and will facilitate the exploitation of these untapped resources.

1 Materials and methods

1.1 Culture of *S. ansochromogenes* 7100

S. ansochromogenes 7100 was cultured in three different fermentation media, SP, R5 and SMMS. SP medium was used previously in our laboratory to produce nikkomycin [17]. R5, a general rich medium, and supplemented minimal medium (SMMS) have both been used for antibiotics production in *Streptomyces* cultures [18]. Spores of *S. ansochromogenes* were inoculated into yeast extract-malt extract (YEME) medium and incubated at 28°C on a rotary shaker (220 r min^{-1}) for 48 h as the seed culture. One milliliter of seed culture was then transferred into 100 mL fermentation medium and cultured till *S. ansochromogenes* cells were harvested at different time intervals.

1.2 DNA isolation and manipulation

Isolation of genomic DNA from *S. ansochromogenes* was performed according to standard protocols [18]. PCR reactions were carried out using either Taq DNA polymerase (PUEX, BEST ALL-HEAL L.L.C, New York, USA) or KOD plus DNA polymerase (TOYOBO, Toyobo Co., Ltd, Osaka, Japan). When necessary, the PCR products were cloned into the *Eco*R V site of pBluescript KSII+ and sequenced using the M13F and M13R primers on the ABI 3730 platform (Life Technologies Corporation, California, USA) by Majorbio Corporation (Majorbio Pharm Technology Co., Ltd, Shanghai, China).

1.3 RNA isolation and reverse transcription-polymerase chain reaction (RT-PCR)

RNA was isolated from *S. ansochromogenes* grown in the three different media with different time-courses. Briefly, *S. ansochromogenes* cells were collected by centrifugation and lysed by grinding in liquid nitrogen with mortar and pestle. When the samples were ground into fine powder, 1 mL TRIzol reagent (Life Technologies Corporation) was added to the tubes. The samples were mixed well and centrifuged at 12000 r min^{-1} for 10 min. The supernatants were subjected to phenol/chloroform extractions, and RNAs were isolated with isopropanol precipitation.

For RT-PCR analysis, cDNA was generated from 500 ng total RNAs by reverse transcription using the SuperScript® III Reverse Transcriptase kit (Life Technologies Corporation) following the manufacturer's protocol. To detect the expression state of each gene cluster, inward primers were designed inside the core genes of each gene cluster. The primers used for the PCRs are listed in Table 1. PCR reactions were carried out with the following cycle conditions: 95°C for 5 min; 95°C for 30 s; 62°C for 30 s; 72°C for 30 s; 28 cycles; final extension for 7 min. To exclude the possible contamination of genomic DNA, each RNA sample was treated by DNase I (Promega Corporation, Wisconsin, USA). Quality and quantity of the RNAs were examined by UV spectrophotometer and agarose gel electrophoresis, respectively. PCR reactions were carried out as described above except that the cycle number was increased to 35.

1.4 Genome sequencing and assembly

The genomic DNA of *S. ansochromogenes* 7100 was sequenced using the 454 GS-FLX system (Roche Diagnostics Co., Basel, Switzerland). The generated reads were assembled using the Newbler package [16], which was designed specifically for the 454 GS-series of pyrosequencing platforms (Roche Diagnostics Co.).

1.5 Annotation of secondary metabolite biosynthetic gene clusters

Open reading frames (ORFs) were predicted by the Glimmer package and annotated using the RAST server [19,20]. The annotated draft genome sequence file, which included information for both the contigs and the ORFs, was uploaded to the antiSMASH web server (http://antismash.secondarymetabolites.org/). All core genes responsible for the formation of the SM backbone and some accessory genes responsible for the tailoring of the SM backbone of SM gene clusters were identified and mapped onto the draft genome of *S. ansochromogenes*. The complete SM gene cluster sequences were obtained by local BLAST searches [21], PCR product sequencing and gap filling method (described below).

Table 1 Primers used in this study

Primers used for RT-PCR and cosmid selection			
Clusters	Primer sequences (5′–3′)	Clusters	Primer sequences (5′–3′)
pks1	*pks1*-F CACCGCCTCCGTCTACGAAGCACA *pks1*-R GGCGCATCCGGCATCCCTCCAT	*terp3*	*terp3*-F ACTACGCCTGCCTCCTCAAC *terp3*-R CAGCCACTCGGCCTTGTAC
pks2	*pks2*-F CGGGCAGATGTACCACGAC *pks2*-R CACCAGCGAGGAGGAGCAG	*terp4*	*terp4*-F ATGACGCCCACGACCTCTTC *terp4*-R ACCGCAGCATGTGCTCCTC
pks3	*pks3*-F CTGCGAGTTCGGCGGCTAC *pks3*-R TCGGCGGTACGACCTGCTT	*terp5*	*terp5*-F GGCCAGTTCCCGCTGATAG *terp5*-R GACGAGTCCGACGAGGTGA
pks4	*pks4*-F GGTGCTGGCCCTGATACGC *pks4*-R GTGCCCGAGGTTGGACTTGA	*terp6*	*terp6*-F CGGTCTTCGGTGCCTATCTC *terp6*-R CCGTCGGTCATCCAGTCGT
pks5	*pks5*-F AAGGACCCGTCGTCACCG *pks5*-R CCGTTCCAGCAGCAGCAT	*sid1*	*sid1*-F CGACTGCTTCTTCCGCTTCC *sid1*-R GCGAGGTCCACCATCTGCTT
pks6	*pks6*-F TGTGCAAGCCCTCGGTGTC *pks6*-R CACGTCGATGTCGCTGGTG	*sid2*	*sid2*-F CTGCTCAACTGCCTGCTGC *sid2*-R CCGGCTGTCGATCATCTCG
pks7	*pks7*-F AGGGCTCCGCCTTCTTCGT *pks7*-R CCGGTCGTTCTGCTTGGTG	*sid3*	*sid3*-F CCGATACCGTGACAACCAGG *sid3*-R GTGAGGAAGCGGCGGAAGG
pks8	*pks8*-F CGGTGGTCAACTCCCTCTTC *pks8*-R TCGGTCGTCCCAGTCGTAG	*buty1*	*buty1*-F TCCCGTTCGGTCACCAGTTC *buty1*-R ACGCAGACGCCGGTAGACG
nrps1	*nrps1*-F CATCGACAGCCAGGTGAAGC *nrps1*-R GTAGGCGGGCAGGTAGGAG	*buty2*	*buty2*-F GGCTTCCGCATCGAGGTTC *buty2*-R CCGACGGCGAGACAGGTGA
nrps2	*nrps2*-F CGGTGAGTGCATCGACATCC *nrps2*-R CCGTAGGCGTCCACGAGAAT	*buty3*	*buty3*-F GGGATGCGGATGGAGGTGA *buty3*-R GGGATGCGGATGGAGGTGA
nrps3	*nrps3*-F GGACACCTACGGGCTCACC *nrps3*-R CACAGCACCAGACGCAAGG	*buty4*	*buty4*-F AAGGAGGTGCTGGTGGACG *buty4*-R GGCGAGAACTCGGTGAAGC
nrps4	*nrps4*-F CTACCACGCCGAGACCACC *nrps4*-R CTGCTGAAGTCCCGCCAAG	*ecto1*	*ecto1*-F TGACCAACGAGGAGACTGGG *ecto1*-R CTTCGGTGAGCAGCGGGTA
nrps5	*nrps5*-F CGGTGGTGAACGCCGAGTA *nrps5*-R CCGAACAGGGTCAGCAGGA	*ecto2*	*ecto2*-F GATCGTCCGCAGCATCAGC *ecto2*-R CCAGTCTCAGGTCCTCGTAGTCG
pks-nrps1	*pks-nrps1*-F GGGGCCTGATCGGCTTCTT *pks-nrps1*-R GCACCTCGGCGTGTTCCAT	*mela1*	*mela1*-F ACAAGGGCCGACGGATACAG *mela1*-R GTAGTGGCTGACGACGCTGA
pks-nrps2	*pks-nrps2*-F CGGGAGAAGGTCGGCTACA *pks-nrps2*-R GGGTTGGGACGGGAGAAGT	*mela2*	*mela2*-F CGCCGTATCCAGGGTGTC *mela2*-R CGTCCGGTACGACGTGTAG
lanti	*lanti*–F CGAAACCCGAGGAAGCAGG *lanti*–R CAGGCGACGAGCAGGAAGG	*acar*	*acar*–F ACGTCGGGGTGGGGGTCAAGA *acar*–R GAGCCGGGCCAGTTCGTGTTCC
terp1	*terp1*-F GCCGCGACCGGATCTATCT *terp1*-R CCCACCTCACGCAGCAACT	*nik*	*nik*–F CGGCGCTGGAGGAACGTAC *nik*–R GGGTGTAGAGGCCGATGCT
terp2	*terp2*-F GCCTCACCTTCGCCCACTG *terp2*-R GCCGTCTGATTTCCTCCACC		
Outward primers used for assembly of gene clusters			
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C186F	GCCTGCTCCTGCCGAAGA	C357F	GGCTCTTGGGACTCCTGTGG
C784R	GGTGATCGACGGCGAAGC	C784F	ACGAGCCGCTCCTCCTGC
C1486F	CGACGAGGCGCAGGTGAA	C1486R	TCCATGCCGTCTCGTCCC
C1342F	AACGCCCACCTCGTCCTG	C769F	GGACGGTGCCGACCAGAG
C769R	GGCAACTCCCTGCTGCTCAC	C321R	CGAGCTGCTGACCGAGCC
C4F	AGGACGTGGGCGAACTGC	C4R	CCGAGTACCTCGCCACCTTC
Primers used for gap filling			
Primers used for gap filling of *nrps2* gene cluster			
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C334R	CGTCTACGACGACCTCTGG	C1039F	GACCAGATACTCGGTGCTGC
C454F	GACCGACACGTCGAACGC	C1039R	TGGTGCGGTCGGTGTTCC
C454R	GGCGACCGAGGAACTGC	C1394F	CGCCTTCAGCACCACCAC
C450R	CACCTCGTCCAGGAACTCG	C1394R	GCTGTTCCTCGCCTCCTG
C450F	GCACCACCTGGCAGACCC	C1236F	GAGCATCGCCACGACCAG
C85R	GTCCCAGCCGTCCTCCTC	C1236R	TACGGCATCCCGCAGTTC
C86F	TCGTCACCATCGCCACCT	C945F	CAGGACGCCGAGTTCACC
C945R	CTCCCTCTGCGGCGGTTC	C612R	CCATCCTGGCGGACGACT
C111R	CTGAGGGTGTCCAGGATGTG	C612F	GAGGCCGCCTACCTGCTC
Primers used for gap filling of *nrps4* gene cluster			
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C398R	GATCTGGGCCGAGGTGCTG	C708R	GCTCCAACGCCTGCTCCTG
C708F	CGCCAGTCCAGGTGGGTGA	C1274F	TCCTCGCTCTGGTCGTGCTG
C872F	GAACGGGTCGGGCAGGAAG	C1274R	TCCACCTCGCTCAACTTCG

(*To be continued on the next page*)

(*Continued*)

Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C872R	GTCACCGAGCACGACTGGA	C1234R	AGGGTGTTGAGGGTGAGGC
C193F	CCTTGGCGTCCAGGGTGAG	C1234F	GCACTCCTACACCCAACTC
C933F	GCCTGGATGCTGAGGATGG	C1069R	TCCTACCTGCGTCAACTCC
C645F	CCTGCCGGTAGGCGGTTTC	C933R	CGACATCGGCTCCCAGGAC
C196F	GAGACGGTGGTGCCGAACA	C645R	ACCGCCAAGTAGCTGTTCC
C177F	GACAGCGTGGAGTGGGTGAA	C196R	CCGACCAGCAGGTGAAGAT
C177R	CGCCTCCTTCTCCTTCGACACCT	C550F	GGCACCACATAGCCGACGAGCAT
C550R	TGGACGAACGGCTCGGCTTC	C1175F	GGTCAGGCCCCGGAAGCAGT
C193R	GACCACGACGAACGGTGAGGAT	C1069F	GTCCAGCACCACGCCCATCG
C398F	GGGCAACGGAGCTGGAC	C1289R	CCAACTCCACCGCCCTGA
	Primers used for gap filling of *pks-nrps1* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C186F	GCCTGCTCCTGCCGAAGA	C784F	ACGAGCCGCTCCTCCTGC
C784R	GGTGATCGACGGCGAAGC	C357F	GGCTCTTGGGACTCCTGTGG
	Primers used for gap filling of *pks-nrps2* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C1342F	AACGCCCACCTCGTCCTG	C769R	GGCAACTCCCTGCTGCTCAC
C769F	GGACGGTGCCGACCAGAG	C407F	GAGGAACGGCAGCAACGC
	Primers used for gap filling of *pks2* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C211R	GGGCCGAGATCGACGAGTAG	C309F	CGCCTGCCTGGTCTTCCT
C211F	CCAGGTCACCGTCCACTC	C321S	AGCGTCAGGCTGTCCAGG
	Primers used for gap filling of *pks4* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C110R	CGGCTCCGTCAAGTCCAACC	C494F	GACTTCCAGGGCGAACAGGG
C494R	CGAGCACCCTGGAGGAACTG	C1150F	GCTCGATCAGGACGTAGCGG
C1150F	CACCTGGCAGACGGACGA	C1426F	GGGGTAGCGGCAGCTCAT
C38F	GGCCCGGATAACCCTCGTAC	C1426R	CGGCTGCTGGACCTGACC
C38R	GACGCACTCGGCACCACC	C1341F	ACAGCTCCAGTGCCTCGTCC
C459F	GGAGGCGGTGCGTTCGTC	C1341R	GGCGGCACCGTGTTCACC
C459R	CCGCTGACCACCGAGGAG	C1312F	CCTGAGGTGGGCGACGAG
C761R	CGGCCCACATCACGTCCC	C1312R	GCTGACCGAGGAACGGGACT
C152F	GAACAGCGGGCCGTAGGC	C761F	GAACAGCGGGCCGTAGGC
C152R	TCTCCTGGGAGGGCGTCA	C826F	CGGTACGCAGGGCGAGTT
C709R	GCGTCGGAGGTCGGGTAG	C826R	GCGGCTCCAGGAGGTCAT
C408F	TGCGACGCCTACCACCAG	C709S	GGCCGGAGGAGTTGAGGA
	Primers used for gap filling of *pks5* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C582F	GGCTTCCAGGCGGTCCAG	C971R	CTCCGTCGCCCTCACCCT
C582R	CCGGCGTGTCCTCCTTCG	C714F	TCGGTGGCGAGCAGGGAC
C611R	CGGTCAGGGAGGAGAACAGC	C714R	CGACCTGCCCACCTACCC
C611F	GCGCCCACCTCGACGAAC	C816F	CTCGGGATCGAAGGTGAGGC
C585F	GCTGCCCTTGAGGGAGTG	C816R	CCCTCCACCTCCACGAAC
C402F	CCGGGTCCACGGCAGTTC	C585R	GGCGAGGCGTTCACCCAC
C402R	GAGGCCCTGCTGCCGTTC	C689F	CGGAGCGGTAGCCGAGGAT
C689R	GCGTCTCCTCCTTCGGCATC	C890R	TCGGCGGAGTCCAGCACC
C399F	GCCCGTACCGTGTCGTCC	C890F	ACCGAACCGCTGCTCCTG
C313F	GTGGAAGGCGTGCGAGAC	C399R	CCGATGTGGACGTGGTGG
C110F	GGCGGTGACGGAGTCCAG	C313R	CGCTGGTGTCCACCCTCA
C31F	GAGCGGGTGGTTCGGTCTG	C971F	GCGGGCACGCTGACACTG
	Primers used for gap filling of *pks7* gene cluster		
Primers	Sequences (5′–3′)	Primers	Sequences (5′–3′)
C178F	GTCCGCTCCTCCTACCCG	C1345R	CACTCGCCGACTCCCTCT
C1345S	GGTCAACACCCGTCTCCAGG	C29F	ACCAAGACGGAGACGGAGGC

The putative chemical structure of the product derived from each gene cluster was predicted using antiSMASH tools. The modules of type I PKSs and NRPSs were predicted and analyzed using the PKS-NRPS analysis package [22].

1.6 Assembly of gene clusters through PCR product sequencing and gap filling

The outward primers (Table 1) used for assembly of SM biosynthetic gene cluster were designed based on the end sequence of the contigs containing the core or accessory genes. The PCR reactions were performed with all possible outward primer matches. For the gap filling of the assembled gene clusters, primers (Table 1) from the two ends of ordered contigs were designed. Genomic DNA from *S. ansochromogenes* 7100 was used as the template. PCR products were purified by agarose gel electrophoresis and then sequenced by Majorbio Corporation.

1.7 Construction of cosmid library and screening

The genomic library of *S. ansochromogenes* 7100 was constructed using a SuperCos I Cosmid Vector Kit (Agilent Technologies, California, USA), according to the manufacturer's instructions. Primers *nrps1*-F/R and *pks-nrps2*-F/R (Table 1) were used for screening the cosmid library by PCR. The positive clones were identified in 96-well plates. The positive cosmids were extracted and end-sequenced with T_3 and T_7 primers.

2 Results

2.1 General features of the *S. ansochromogenes* draft genome

Genome sequencing of *S. ansochromogenes* 7100 generated 2321457 reads with an average length of 387 bp. The overlapping reads were assembled into 1299 random contigs with an average length of 7229 bp (N50 value=15893 bp). The largest contig has a length of 74262 bp, while the smallest one has a length of 500 bp. The total size of the assembly was approximate 9.0 Mb, with an average G+C content of 72.2%. A total of 9000 putative protein coding sequences (CDSs) were identified in the draft genome.

2.2 Assembly and identification of secondary metabolite biosynthetic gene clusters

To analyze the SM biosynthetic gene clusters, a comprehensive antiSMASH-based method was adopted to assemble the SM biosynthetic gene clusters in *S. ansochromogenes*. A schematic diagram illustrating the assembly of the SM biosynthetic gene clusters was summarized in Figures 1 and 2. Sixty-two antiSMASH identified gene clusters, including nikkomycin biosynthetic gene cluster, were obtained. This result included complete and split gene clusters. The core genes or accessory genes were identified in the clusters. The contigs, which contained either the core genes or accessory genes, were picked out manually. BLAST searches were done for the core and accessory genes and the search results were grouped into two categories: the first category contained genes that shared high identities with sequences deposited in the NCBI non-redundant nucleotide databases (>50%); and the second category contained genes that shared low identities (<50%). For the first category, homologous gene clusters were searched in the NCBI non-redundant nucleotide databases using HMMER-BLAST [23] with the core genes and accessory genes as the query sequences. The sequences of the homologous gene clusters were used to perform local BLAST searches against the nucleotide database created from *S. ansochromogenes* draft genome sequence. Contigs related to an individual homologous gene cluster were picked out, and the order and orientation of these contigs were determined based on the synteny between the SM biosynthetic gene clusters in *S. ansochromogenes* and homologous clusters in other organisms. Most of the SM biosynthetic gene clusters (31 of 35) were assembled using this method (Figure 1). For the second category, the complete sequences of two gene clusters were obtained based on the sequence generated from PCR reactions with possible matches of outward primers designed in two ends of the contigs (Figure 2). The complete sequences of two additional gene clusters were obtained based on sequences from the end-sequencing of the cosmids (Figure 2). Finally, a total of at least 35 gene clusters were identified in the draft genome of *S. ansochromogenes* and general information of these gene clusters were summarized in Table 2.

2.3 Features of secondary metabolite biosynthetic gene clusters

The 35 identified gene clusters were predicted to be involved in the biosynthesis of polyketides (PKs), nonribosomal peptides (NRPs), siderophores, melanins, terpenes, nucleoside, butyrolactones, ecotines, acarbose, lantipeptide and polyketide-nonribosomal peptides (PK-NRPs) hybrids. The combined length of these gene clusters was estimated to be about 540 kb, accounting for about 6% of the *S. ansochromogenes* genome. Considering the abundance of bioactive PKs and NRPs in *Streptomyces*, here we present the general features of several PK, NRP and PK-NRP biosynthetic gene clusters.

Eight PK biosynthetic gene clusters included type I (*pks1*, *pks2*, *pks4* and *pks5*), type II (*pks3* and *pks7*) and type III (*pks6* and *pks8*) gene clusters. Based on sequence identities and organization of the ORFs in the gene clusters, the homology searches suggested that the *pks1*, *pks6* and *pks7* gene clusters are supposed to be responsible for the biosyn-

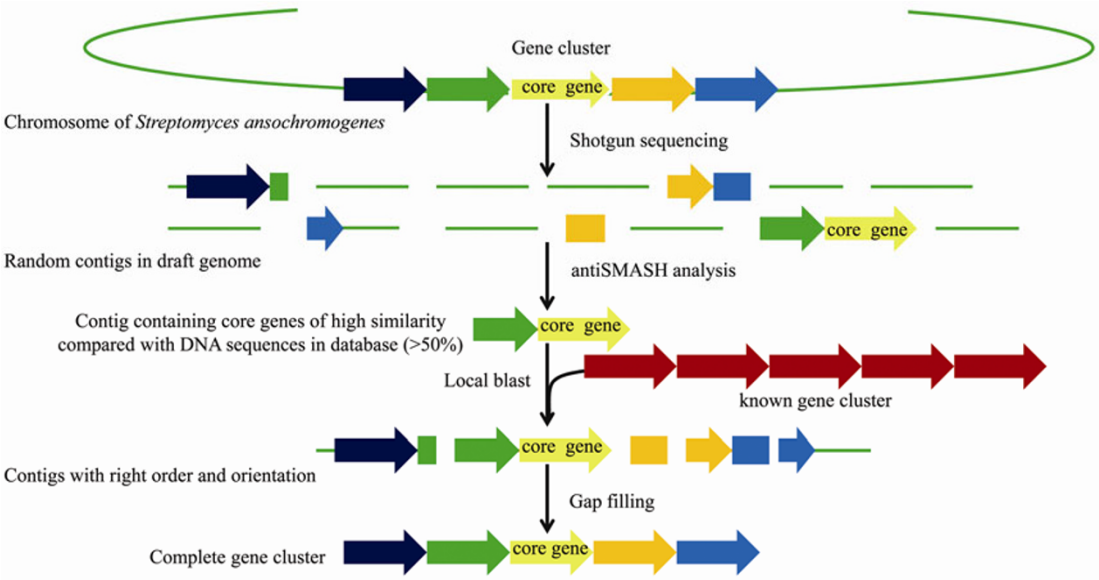

Figure 1 Assembly of SM gene clusters with high similarity to known gene clusters. Core genes in the *S. ansochromogenes* genome with high similarity to DNA sequences in the databases were identified by antiSMASH analysis. Gene clusters were assembled using a local BLAST method and reference gene clusters.

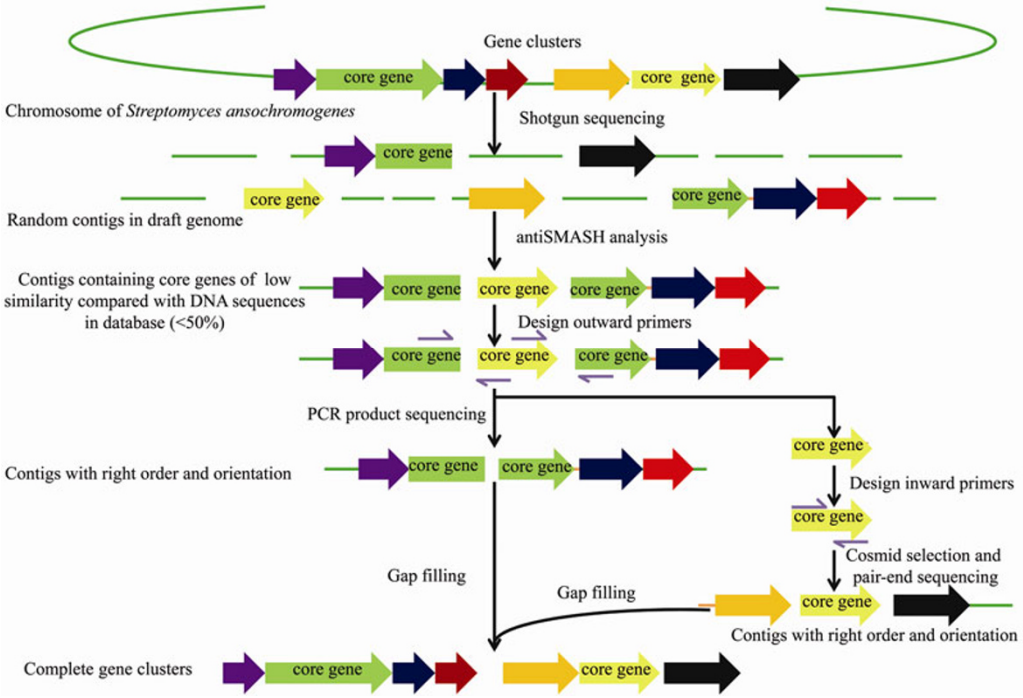

Figure 2 Assembly of SM gene clusters with low similarity to known gene clusters. Core genes in the *S. ansochromogenes* genome with low similarity to DNA sequences in the databases were identified by antiSMASH analysis. Gene clusters were assembled using either PCR product sequencing or cosmid pair-end sequencing.

thesis of angolamycin, flaviolin and oviedomycin, respectively. The *pks4* and *pks5* gene clusters had homologous gene clusters in *Streptomyces hygroscopicus* subsp. *jingangensis* 5008, their corresponding products remain unknown. The *pks4* gene cluster contained two large *pks* genes, two noncanonical *nrps* genes, one tailoring enzyme gene, and a LuxR family regulatory gene. The *pks5* gene cluster contained four *pks* mega synthetase genes with no tailoring enzyme genes and a two-component system regulatory gene was situated on the left side of the *pks5* gene cluster. The *pks2* and *pks8* gene clusters shared low sequence similarities with SM biosynthetic gene clusters deposited in the NCBI

Table 2 Secondary metabolite biosynthetic gene clusters in *S. ansochromogenes*

Cluster designation	Actual (*) or predicted products	Type	Approximate size (kb)	Accession numbers
pks1	Angolamycin	Type I PKS	90	KF170321
pks2	Unknown	Type I PKS	50	KF170322
pks3	Unknown	Type II PKS	30	KF170323
pks4	Unknown	Type I PKS	32	KF170324
pks5	Unknown	Type I PKS	40	KF170325
pks6	Unknown	Type III PKS	15	KF170326
pks7	Oviedomycin	Type II PKS	26	KF170327
pks8	Unknown	Type III PKS	13	KF170328
nrps1	Unknown	NRPS	37	KF170329
nrps2	Unknown	NRPS	47	KF170330
nrps3	Unknown	NRPS	38	KF170331
nrps4	Unknown	NRPS	50	KF170332
nrps5	Unknown	NRPS	35	KF170333
pks-nrps1	Unknown	PKS-NRPS	22	KF170334
pks-nrps2	Unknown	PKS-NRPS	25	KF170335
lanti	Lantipeptide	Lantipeptide	8	KF170336
terp1	Squalene/Phytoene	Terpene	16	KF170337
terp2	Unknown	Terpene	13	KF170338
terp3	Germacrene	Terpene	7	KF170339
terp4	Polyprenyl	Terpene	10	KF170340
terp5	Unknown	Terpene	12	KF170341
terp6	Unknown	Terpene	15	KF170342
sid1	Desferrioxamine	NRPS-independent	22	KF170343
sid2	Unknown	NRPS-independent	38	KF170344
sid3	siderophore	NRPS-independent	28	KF170345
buty1	Gamma-factor	Lactone	3	KF170346
buty2	A-factor	Lactone	4.5	KF170347
buty3	Gamma-factor	Lactone	2.5	KF170348
buty4	A-factor	Lactone	2	KF170349
ecto1	Ectoine	Ectoine	4	KF170350
ecto2	Ectoine	Ectoine	4	KF170351
mela1	Melanin	Ectoine	5	KF170352
mela2	Melanin	Ectoine	5	KF170353
acar	Acarbose	Oligosaccharides	25	KF170354
nik	Nikkomycin*	Nucleoside peptide	30	KF170355

non-redundant nucleotide databases. Two activator genes and one TetR family regulatory gene were situated on the left side of *pks2* gene cluster. The *pks8* gene cluster contained only one chalcone synthetase-like gene, which is about 2–4 kb away from other genes predicted to be involved in tailoring steps. The domain organization and proposed monomers of the type I PKSs are summarized in Table 3.

Five gene clusters (*nrps1–5*) were proposed to be responsible for the biosynthesis of NRP products. A homologous gene cluster of *nrps2* was found in the genome of *Streptomyces* sp. e14. A $FADH_2$-dependent halogenase coding gene was situated inside the *nrps2* gene cluster. Two absolute conserved motifs, GXGXXG in the N-terminal and WXWXIP in the C-terminal, were identified in the $FADH_2$-dependent halogenase [24]. The *nrps4* and *nprs5* gene clusters had homologs in *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008. Two 20 kb *nrps* genes, which are rare in other NRP biosynthetic pathways, were involved in biosynthesis of the backbone of the product derived from *nprs4* gene cluster. A regulatory gene encoding a PAC/PAS-like protein that senses changes of redox potential, light intensity, oxygen, small ligands, and the overall energy level of a cell, was situated on the left side of the *nrps4* gene cluster [25].

During NRP biosynthesis, the adenylaion (A) domain of each module of NRPSs might select the cognate amino acid from the pool of available substrates. Previous studies revealed that the similarities between the A domains activating the same amino acid were significantly high and there are defined rules for the structural basis of substrate recognition by the A domains of NRPSs. The functional domains residing in each NRPS in the genome of *S. ansocrhomogenes* were searched using PKS-NRPS analysis tool [22]. The domain arrangements in each module and the amino acid substrates that are recognized by the A domains are summarized in Table 4.

PK-NRP hybrid compounds were assigned to the *pks-nrps1* and *pks-nrps2* gene clusters. The *pks-nrps1* gene cluster had both a giant gene, which contained modules be-

Table 3 Domain organization and deduced monomers in type I PKSs

Polypeptides	Module	Domain organization[a)]	Predicted monomer
Pks1-1	Module1	KS-AT-ACP	Methylmalonyl-CoA
	Module2	KS-AT-KR-ACP	Methylmalonyl-CoA
	Module3	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks1-2	Module1	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks1-3	Module1	KS-AT-KR-ACP	Methylmalonyl-CoA
	Module2	KS-AT-DH-KR-ER-ACP	Malonyl-CoA
Pks1-4	Module1	KS-AT-KR-ACP	Methylmalonyl-CoA
Pks1-5	Module1	KS-AT-KR-ACP	Malonyl-CoA
	Module2	TE	
Pks4-1	Module1	ACP	Malonyl-CoA
	Module2	KS-AT-DH-KR-ACP	Malonyl-CoA
	Module3	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks4-2	Module1	KS-AT-DH-KR-ACP	Methylmalonyl-CoA
	Module2	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks5-1	Module1	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks5-2	Mdoule1	KS-AT-KR-ACP	Malonyl-CoA
Pks5-3	Mdoule1	KS-AT-DH-KR-ACP	Malonyl-CoA
	Mdoule2	KS-AT-DH-KR-ACP	Methylmalonyl-CoA
	Mdoule3	KS-AT-DH-KR-ACP	Malonyl-CoA
Pks5-4	Module1	KS-AT-ACP	Malonyl-CoA
	Module2	TE	

a) KS, ketosynthase domain; AT, acyltransferase domain; ACP, acylcarrier domain; DH, dehydratase domain; KR, ketoreductase domain; TE, thioesterase domain; ER, enoylreductase domain.

Table 4 Prediction of the amino acid residues that determine adenylation domain specificity, and the amino acid substrates and domain organization of NRPSs

Polypeptides	Residues in adenylation domain[a)]								Substrates[b)]	Domain organization[c)]
Nrps1-1	D	F	W	S	V	G	M	V	Thr	A-PCP
Nrps1-2	D	A	F	W	F	G	G	T	Val	A-PCP-C
Nrps1-3	D	V	W	H	F	S	L	V	Ser	A-PCP-N-PCP-TE
Nrps2-1	D	V	P	K	V	G	E	V	–	A-PCP
	D	V	F	C	V	A	M	T	–	A-PCP
	D	I	W	E	V	T	A	D	–	C-A-PCP
Nrps2-2	D	A	G	A	I	G	M	V	–	C-A-PCP-E-C
Nrps2-3	D	A	W	Q	A	A	T	V	–	A-PCP
	D	L	P	K	V	A	E	V	–	A-PCP-C
Nrps3-1	D	V	W	H	L	S	L	I	Ser	C-A-PCP
Nrps3-2	D	A	W	Q	C	A	T	I	–	C-A-PCP
	D	V	D	E	N	G	N	V	–	C-A-PCP
Nrps3-3	D	F	W	N	V	G	M	V	Thr	C-A-PCP
	D	L	T	K	X	X	E	V	–	N-A-PCP-TE
Nrps4-1	D	A	L	L	I	G	S	I	–	A-N-PCP
	D	A	F	S	V	A	I	V	–	C-A-PCP-E
	D	A	Y	W	W	G	G	T	Val	C-A-PCP-E
	D	V	F	S	V	A	I	V	Trp	C-A-PCP-E
	D	A	L	L	I	G	S	V	–	C-A-PCP
Nrps4-2	D	T	ND	D	M	G	F	V	–	C-A-PCP-E
	D	F	W	N	V	G	M	V	Thr	C-A-PCP
	D	T	W	N	L	G	M	V	Thr	C-A-PCP-E
	D	A	F	W	W	G	G	T	Val	C-A-PCP
	D	A	L	L	I	G	S	V	Trp	C-A-PCP-E
Nrps5-1	ND	ND	ND	ND	ND	ND	ND	ND	–	PCP
	D	I	W	Q	ND	S	T	A	–	A-PCP-TE

a) Amino acid residues are indicated using their single letter code; ND, consensus amino acid not detected. b) –, amino acid substrate is unknown. c) C, condensation domain; PCP, peptide carrier protein domain; A, adenylation domain; E, epimerization domain; N, domain undetermined.

longing to *pks* and *nrps*, and *pks*, *nrps* genes alone. The *pks-nrps2* gene cluster, on the other hand, contained only independent *pks* and *nrps* genes for the backbone formation of PK-NRP hybrid compound. The *pks-nrps1* gene cluster possessed a noncanonical *nrps* gene that did not harbor condensation (C) domains, but contained the adenylation (A) domains and the peptidyl carrier protein (PCP) domains. A two-component regulatory gene and a PaaX family regulatory gene were located in the left and right sides of the *pks-nrps2* gene cluster, respectively.

2.4 Analysis of gene clusters by RT-PCR

To examine the expression profiles of the gene clusters, *S. ansochromogenes* was incubated in three different fermentation media (see Materials and methods). RT-PCR analysis was carried out with primers of selected core gene for each gene cluster (Table 1). Fifteen gene clusters were silent in the three different media (Figure 3A), while 17 gene clusters were transcribed in all three media (Figure 3C). The other three gene clusters were transcribed in either one or two of the three media (Figure 3B). These results demonstrated that SM biosynthetic gene clusters have different expression pattern in different fermentation media.

3 Discussion

The discovery of new antibiotics is urgently required to combat the alarming rise in the emergence of resistant bacteria and the abuse of antibiotics. A large number of SM biosynthetic gene clusters revealed by genome sequencing represent potential new sources of antibiotics. The identification and annotation of these SM biosynthetic gene clusters are essential for successful mining of these, as yet, unexplored resources. Several bioinformatics packages have proven to be capable of identifying SM biosynthetic gene clusters with complete genome sequences [7–12]. Given that more than 90% of the available *Streptomyces* genomes are draft sequence [13], it is necessary to devise methods to identify and assemble SM biosynthetic gene clusters using the draft genome sequences of *Streptomyces*.

In this paper, we devised a comprehensive antiSMASH-based method to identify and assemble SM biosynthetic gene clusters from a draft genome sequence of *S. ansochromogenes*. The core and accessory genes of SM biosynthetic gene clusters were first identified using antiSMASH. This step provided a basic knowledge of the putative gene clusters. To further investigate the gene clusters of interest, complete nucleotide sequences are needed. Because the genes responsible for SMs biosynthesis are typically clustered together in the genome, contigs containing gene clusters of interest were organized according to the results obtained using HMMER-BLAST [23]. Gap sequences between contigs were filled either by PCR or by end-sequencing of cosmids. The method that we have described here can also be applied to the draft genomes of other microorgan-

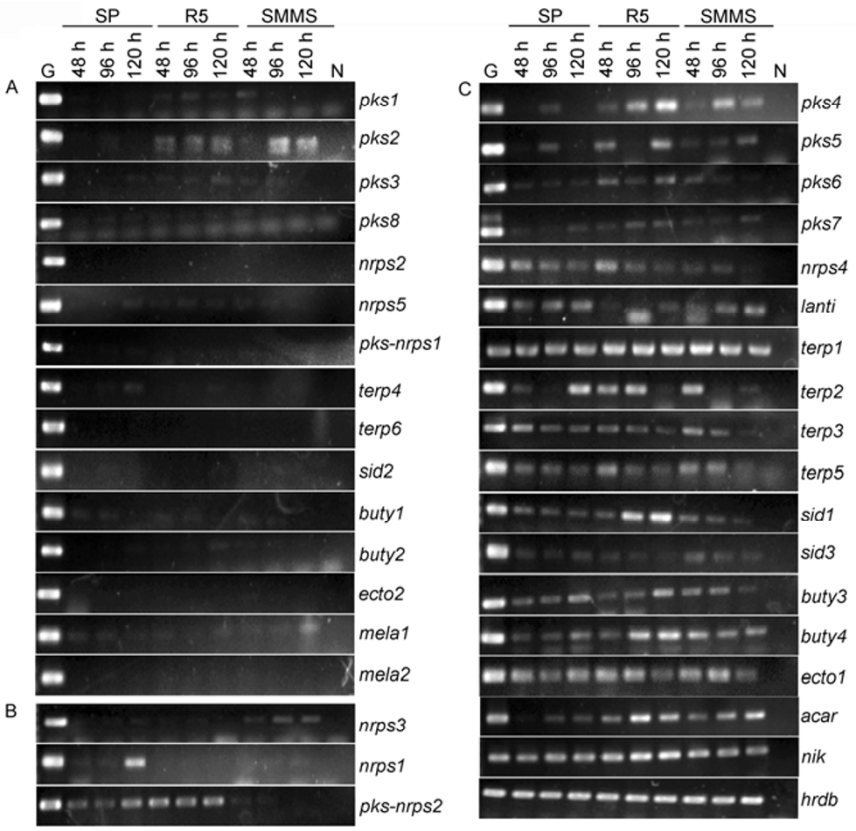

Figure 3 Analysis of gene clusters in the *S. ansochromogenes* genome by RT-PCR. A, Gene clusters that were silent in all three culture media. B, Gene clusters that were expressed in one or two of the three culture media. C, Gene clusters that were expressed in all three culture media. cDNA templates were synthesized from RNA extracted from *S. ansochromogenes* as detailed in Materials and methods. The constitutively expressed gene *hrdB,* which encodes the major sigma factor, was used as the positive control (Figure 3C). The media (SP, R5, SMMS) and incubation times are indicated at the top of each panel. *S. ansochromogenes* genomic DNA (G) and ddH_2O (N) were used as positive and negative controls for each primer pair. The name of each gene cluster is indicated to the right of each gel.

isms like fungi.

Comparative genome analysis has demonstrated that any group of genomes contained both shared (core) and unique (auxiliary) SM biosynthetic gene clusters [26]. Compounds derived from the unique SM biosynthetic gene clusters are the metabolites that are specific to one strain. Eight gene clusters (*pks2*, *pks6*, *pks8*, *nrps1*, *nrps3*, *pks-nrps1*, *pks-nrps2*, *lanti*) shared low sequence identities (<40%) with known gene clusters in other strains. These gene clusters are most likely to be species specific and might produce compounds with novel chemical structures and biological activities. For the gene clusters (*nrps1*, *nrps3*, *pks6*, *pks-nrps2*, *lanti*) that were expressed in one of the three media tested, gene knock-out combined with the comparative metabolic profile between mutants and wild type can be used to identify their corresponding products. For the silent gene clusters (*pks2*, *pks8*, *pks-nrps1*) that were not expressed in any of the three media, various methods could be used to activate them [27], and subsequently to dissect the metabolic pathway and flux [28].

Surprisingly, four gene clusters (*buty1*, *buty2*, *buty3*, *buty4*) were predicted to encode enzymes for the formation of butyrolactone, which is the core structure of signaling molecules called bacterial hormones. In general, only one or two gene clusters are responsible for the formation of butyrolactone in *Streptomyces* [3]. The signal molecules have been reported to play important roles in regulation of antibiotic production and differentiation in *Streptomyces* [29]. The discovery of the existence of four butyrolactone biosynthetic gene clusters in *S. ansochromogenes* implied that there is a complex regulatory network involved in signaling molecules in *S. ansochromogenes*. RT-PCR analysis revealed that the *buty1* and *buty2* gene clusters were silent in three different medium, suggesting that some signaling molecules with novel structure probably existed in *S. ansochromogenes*.

This work was supported by grants from the Ministry of Science and Technology of China (2013CB734001), the National Natural Science Foundation of China (31270110, 31030003), and the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-J-6).

1 Bentley S, Chater K, Cerdeno-Tarraga A M, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2). Nature, 2002, 417: 141–147
2 Ōmura S, Ikeda H, Ishikawa J, et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci USA, 2001, 98: 12215–12220
3 Nett M, Ikeda H, Moore B S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat Prod Rep, 2009, 26: 1362–1384
4 Song L, Barona-Gomez F, Corre C, et al. Type III polyketide synthase β-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by *Streptomyces coelicolor* genome mining. J Am Chem Soc, 2006, 128: 14754–14755
5 Bok J W, Hoffmeister D, Maggio-Hall L A, et al. Genomic mining for *Aspergillus* natural products. Chem Bio, 2006, 13: 31–37
6 Gross H, Stockwell V O, Henkels M D, et al. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Bio, 2007, 14: 53–63
7 Starcevic A, Zucko J, Simunkovic J, et al. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. Nucleic Acids Res, 2008, 36: 6882–6892
8 Weber T, Rausch C, Lopez P, et al. CLUSEAN: a computer-based frame-work for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol, 2009, 140: 13–17
9 Li M, Ung P, Zajkowski J, et al. Automated genome mining for natural products. BMC Bioinformatics, 2009, 10: 185
10 Anand S, Prasad M, Yadav G, et al. SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res, 2010, 38: W487–W496
11 Khaldi N, Seifuddin F T, Turner G, et al. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol, 2010, 47: 736–741
12 Medema M H, Blin K, Cimermancic P, et al. AntiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res, 2011, 39: W339–W346
13 Pagani I, Liolios K, Jansson J, et al. The genomes online database (gold) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res, 2012, 40: D571–D579
14 Simpson J T, Wong K, Jackman S D, et al. ABySS: a parallel assembler for short read sequence data. Genome Res, 2009, 19: 1117–1123
15 Zerbino D R, Birney E. Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. Genome Res, 2008, 18: 821–829
16 Chaisson M J, Pevzner P A. Short read fragment assembly of bacterial genomes. Genome Res, 2008, 18: 324–330
17 Li Y, Ling H, Li W, et al. Improvement of nikkomycin production by enhanced copy of *sanU* and *sanV* in *Streptomyces ansochromogenes* and characterization of a novel glutamate mutase encoded by *sanU* and *sanV*. Metab Eng, 2005, 7: 165–173
18 Kieser T, Bibb M J, Buttner M J, et al. Practical *Streptomyces* Genetics. Norwich: The John Innes Foundation, 2000
19 Aziz R, Bartels D, Best A, et al. The rast server: rapid annotations using subsystems technology. BMC Genomics, 2008, 9: 75
20 Delcher A L, Harmon D, Kasif S, et al. Improved microbial gene identification with glimmer. Nucleic Acids Res, 1999, 27: 4636–4641
21 Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215: 403–410
22 Ansari M Z, Yadav G, Gokhale R S, et al. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res, 2004, 32: W405–W413
23 Eddy S R. Profile hidden markov models. Bioinformatics, 1998, 14: 755–763
24 Van Pée K H, Patallo E P. Flavin-dependent halogenases involved in secondary metabolism in bacteria. Appl Microbiol Biotechnol, 2006, 70: 631–641
25 Taylor B L, Zhulin I B. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol R, 1999, 63: 479–506
26 Walsh C T, Fischbach M A. Natural products version 2.0: connecting genes to molecules. J Am Chem Soc, 2010, 132: 2469–2493
27 Liu G, Chater K F, Chandra G, et al. Molecular regulation of antibiotic biosynthesis in *Streptomyces*. Microbiol Mol Biol R, 2013, 77: 112–143
28 Lai S, Zhang Y, Liu S, et al. Metabolic engineering and flux analysis of *Corynebacterium glutamicum* for L-serine production. Sci China Life Sci, 2012, 55: 283–290
29 Horinouchi S, Beppu T. A-factor as a microbial hormone that controls cellular differentiation and secondary metabolism in *Streptomyces griseus*. Mol Microbiol, 1994, 12: 859–864