Supplemental Material

Title: Functional Genome Mining for Metabolites Encoded by Large Gene Clusters through Heterologous Expression of a Whole-Genome Bacterial Artificial Chromosome Library in *Streptomyces* spp.

Min Xu[1], Yemin Wang[1], Zhilong Zhao[2], Guixi Gao[1], Sheng-Xiong Huang[3], Qianjin Kang[1], Xinyi He[1], Shuangjun Lin[1], Xiuhua Pang[2], Zixin Deng[1], and Meifeng Tao[1*]

[*]To whom correspondence should be addressed. Email: tao_meifeng@sjtu.edu.cn

CONTENTS

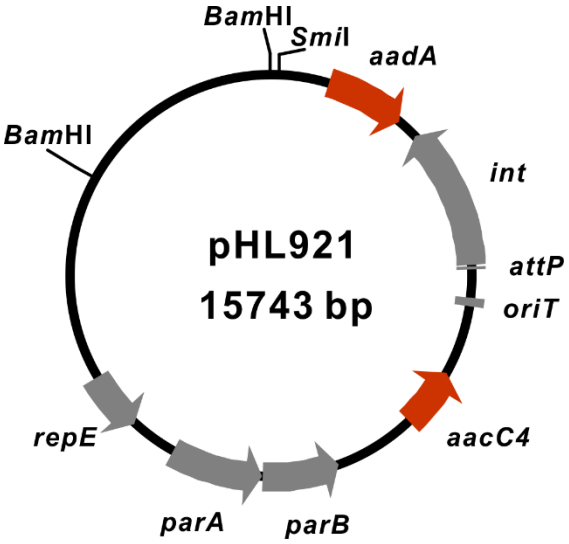


Figure S1. BAC vector pHL921. *repE*, *parA,* and *parB* provide stable replication at unit number in *E. coli*. *aacC4*, apramycin resistance. *aadA*, streptomycin/spectinomycin resistance. *oriT*, origin of transfer; *attP*, attachment site of phage *ΦC*31; *int*, *ΦC*31 integrase. The two *Bam*HI sites are for the cloning of large fragment genomic DNA partially digested with *Sau*3AI. The unique *Smi*I site (ATTT^AAAT) is very rare in high G+C *Streptomyces* DNA and was used to linearize BAC constructs for evaluation of their sizes using pulsed field gel electrophoresis (PFGE). DNA sequence: GenBank KP823602.

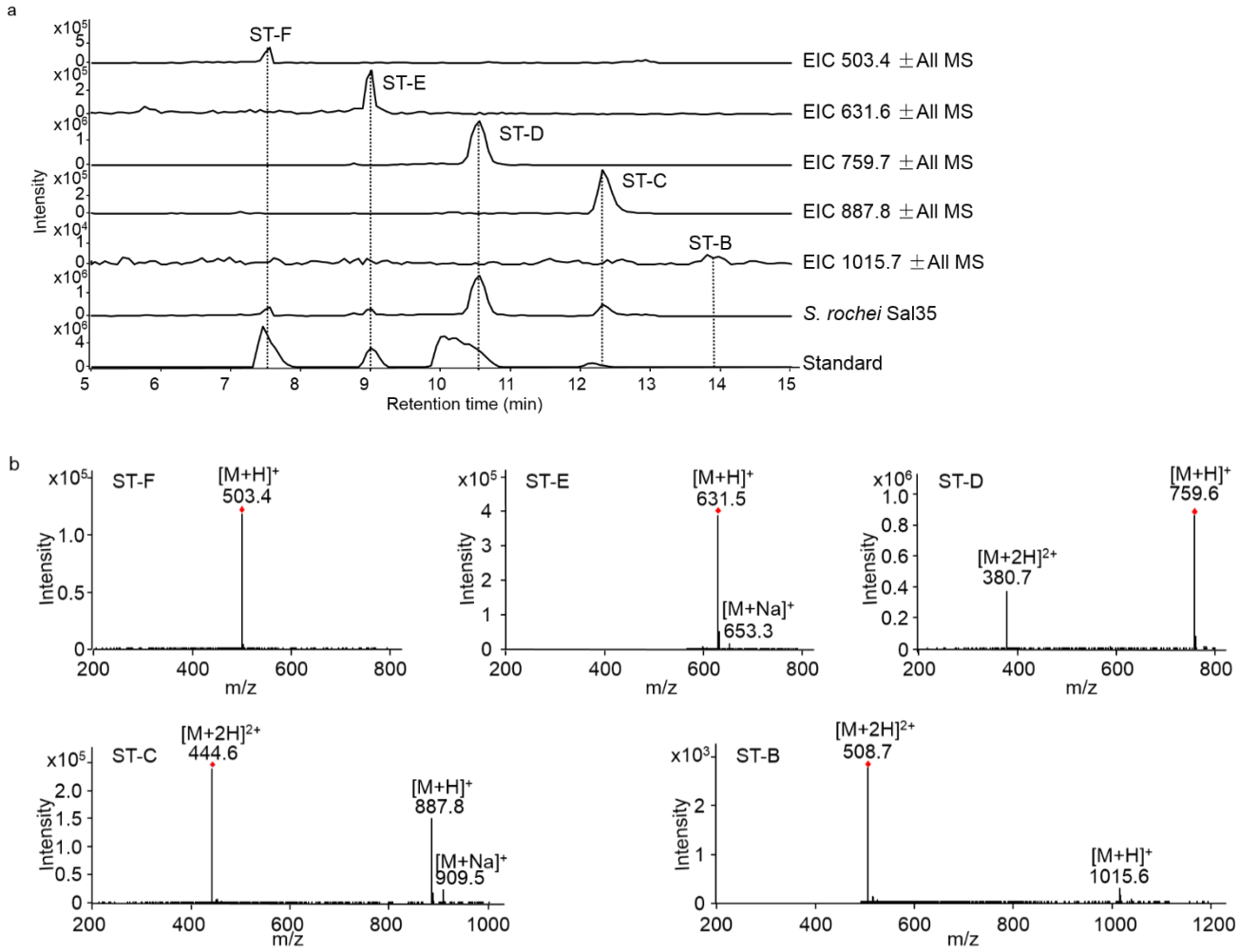


Figure S2. Streptothricins produced by *S. rochei* Sal35. (a), Extracted ion chromatography of streptothricins produced by *S. rochei* Sal35. ST-F, E, D, C and minute amounts of ST-B were detected in *S. rochei* Sal35 ferment. Note different scale on y-axis. (b), Mass spectra of streptothricins.

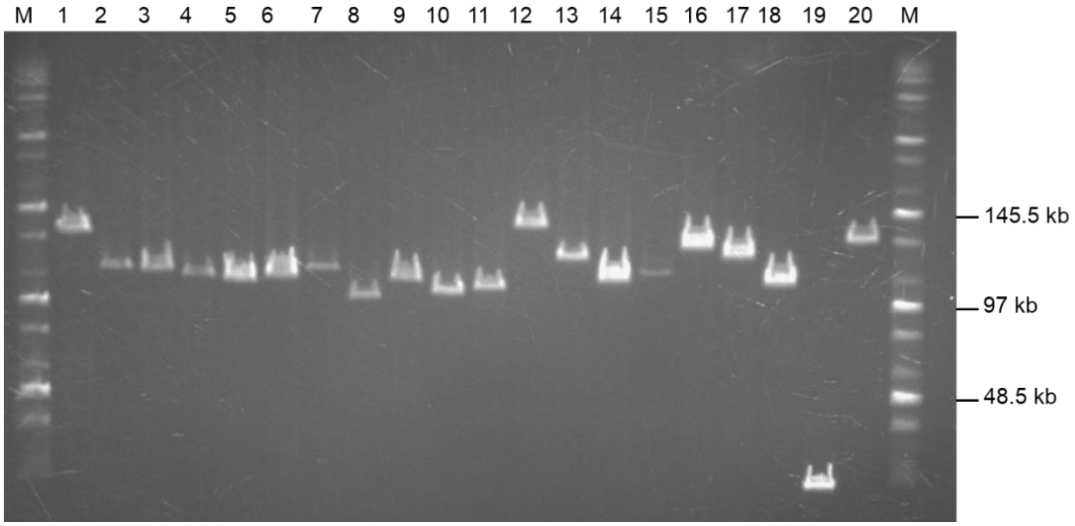


Figure S3. Determination of the size of BAC inserts by PFGE. *Smi*I digested BACs were evaluated by PFGE, an average insert size of ~100 kb (= 113 kb - 13 kb vector) and insertion rate of 95% (19/20) could be estimated from the PFGE gel. NEB MidRange IPFG Marker was used here and the PFGE condition was: switch time: 1-25 s, voltage: 6 V/cm; included angle, 120° ; 14 °C in 0.5 x TBE for 20 h.

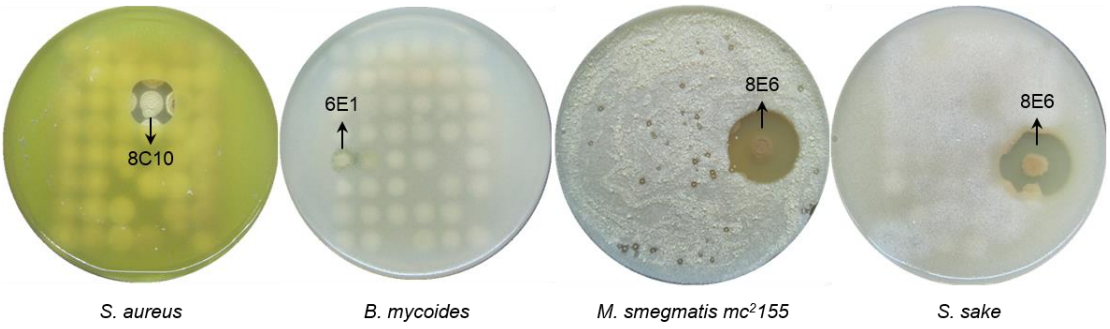


Figure S4. Output of the antimicrobial screening in LEXAS. Antibiosis observed after overlay with *S. aureus*, *B. mycoides*, *M. smegmatis* mc^2 155, and *S. sake* and incubated for 16-24 h. BAC clones bearing BGCs producing bioactive secondary metabolites could be directly visualized and screened by LEXAS.

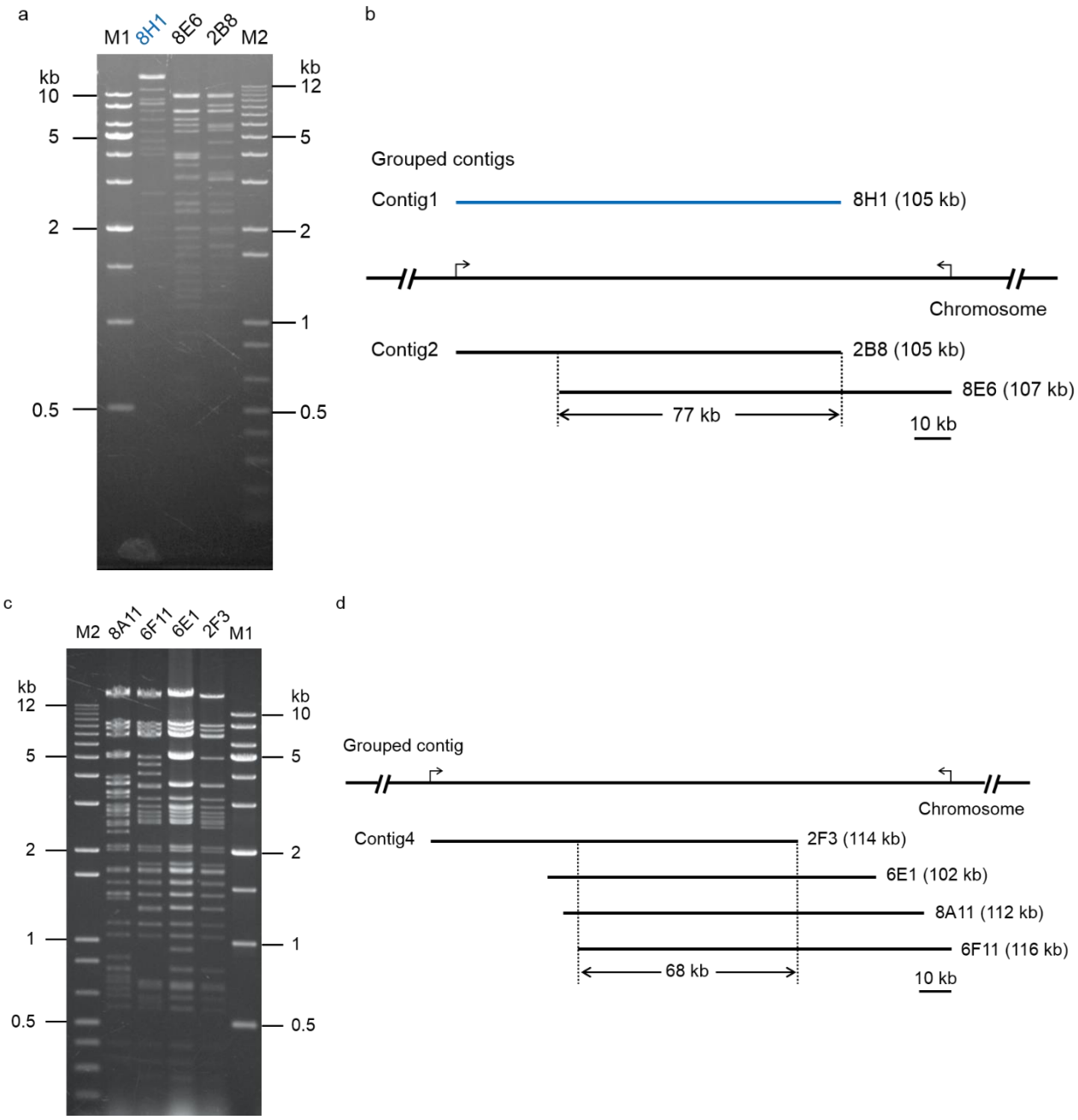


Figure S5. Contigs grouped by *Pvu*II restriction digestion. (a), *Pvu*II restriction digestion of 2B8, 8E6, and 8H1. 8H1 shared no similar bands with other positive BACs was named contig1, 2B8 and 8E6 formed contig2 and shared a ~77 kb overlap. (b), The relative arrangement of 2B8 and 8E6 on *S. rochei* sal35 chromosome. (c), *Pvu*II restriction digestion of 2F3, 6E1, 6F11, and 8A11. These 4 BAC clones shared a ~68 kb overlap and formed contig4. (d), The relative arrangement of contig4 BAC clones on *S. rochei* Sal35 chromosome. The black arrows on the chromosome indicated the boundaries of contig 2 (in b) or contig 4 (in d). Restriction digestion and contig mapping of contig3 were exhibited in the main text (Figure 4). DNA markers used here were 1 kb DNA ladder (Dongsheng Biotech, M1) and 1 kb plus ladder (Invitrogen, M2).

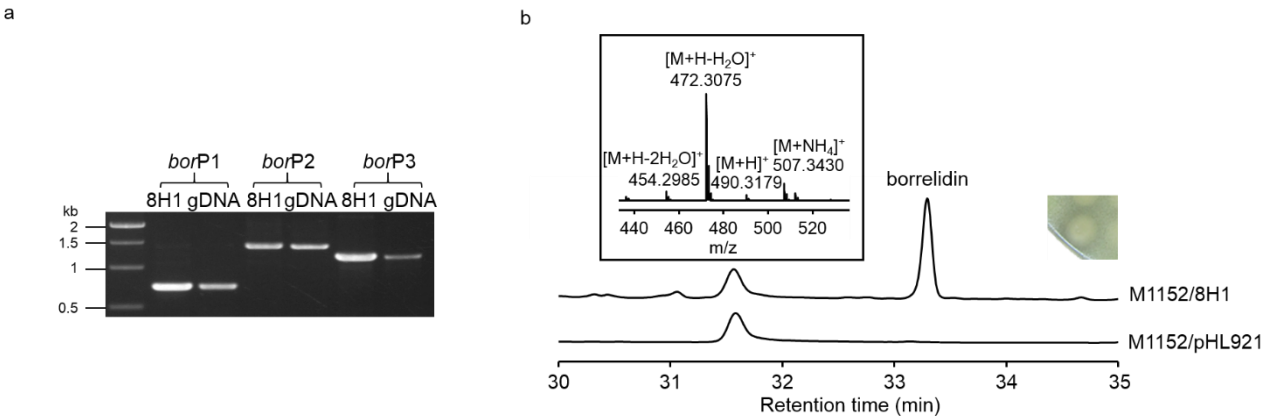
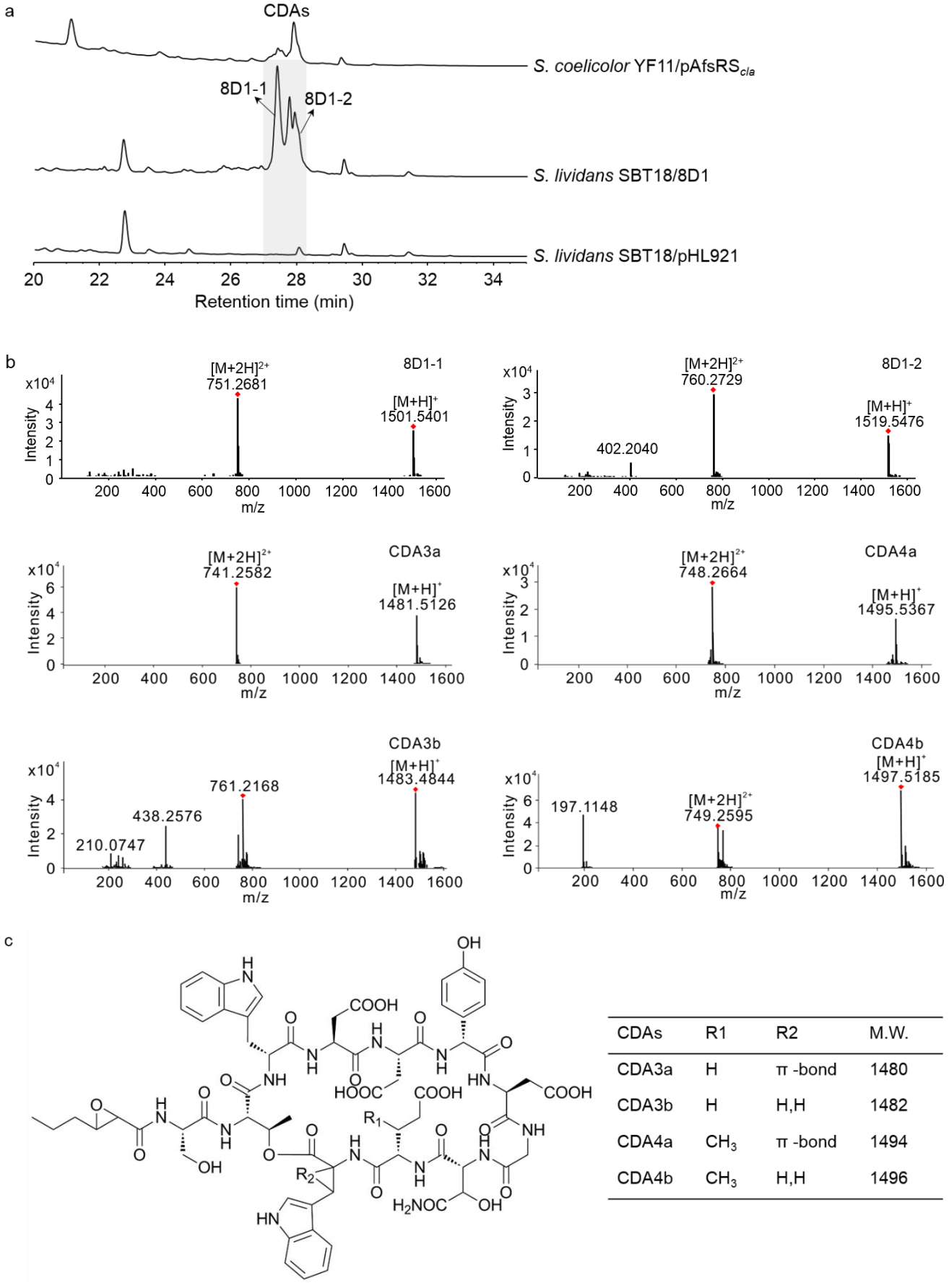


Figure S6. Heterologous expression and identification of borrelidin. (a), PCR verification of *bor* pathway in 8H1. Specific 0.7 kb, 1.5 kb, and 1.2 kb PCR products targeting *borE*, *borA2*, and *borJ* were amplified and sequenced which indicated the presence of *bor* pathway in 8H1. (b) Comparative metabolic profiling of *S. coelicolor* M1152/8H1 and *S. coelicolor* M1152/pHL921. The borrelidin peak was noted and growth inhibition zone around 8H1 in LEXAS was also shown. Molecular mass determined for the borrelidin peak by HRESI-QTOF, and m/z of 490.3179 ($[M+H]^+$, calculated 490.3169, error 2.0 ppm) for borrelidin was shown as insert.



CDAs	R1	R2	M.W.
CDA3a	H	π -bond	1480
CDA3b	H	H,H	1482
CDA4a	CH_3	π -bond	1494
CDA4b	CH_3	H,H	1496

Figure S7. Detection and identification of CDAs produced by *S. lividans* SBT18/8D1. (a),

Detection of CDAs from extract of *S. lividans* SBT18/8D1. The minor peak in the vector control was not CDA. *S. coelicolor* YF11/pAfsRS$_{cla}$ propagated in SV2 medium (1) producing CDAs was set as control. (b), Mass spectra of 8D1-1, 8D1-2, CDA3a/b and CDA4a/b detected in the *S. lividans* SBT18/8D1 extract. (c), Chemical structures of CDA3a/b and CDA4a/b (2).

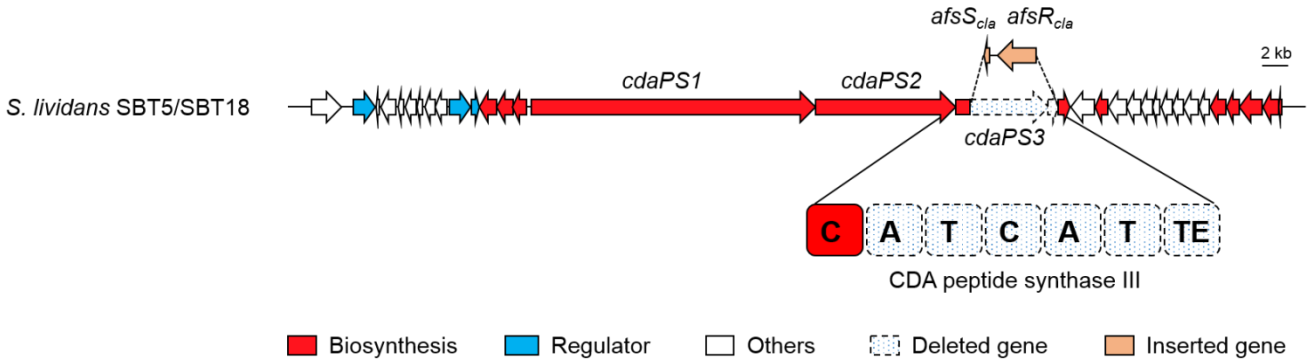


Figure S8. Disrupted *cda* BGC in *S. lividans* SBT5 and *S. lividans* SBT18. *cda* pathway was disrupted in *S. lividans* SBT5/SBT18 by deleting a 7,116 bp region covering *cdaPS3* and a downstream hydrolase gene, and two pleiotropic activators *afsR$_{cla}$* and *afsS$_{cla}$* were introduced at the deletion site (3). Dashed arrows, deleted genes in *S. lividans* SBT5/SBT18. Dashed boxes, A-T-C-A-T-TE domains of CdaPS3 that were deleted in *S. lividans* SBT5/SBT18.

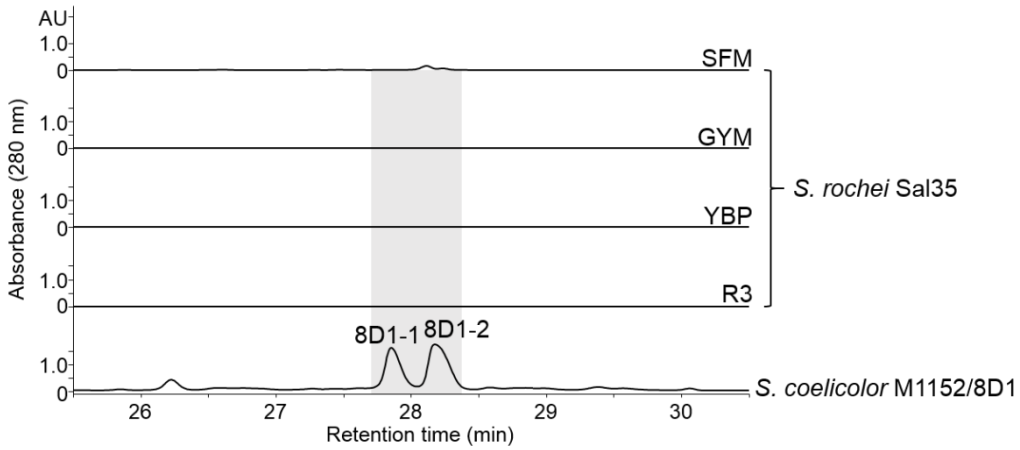


Figure S9. Detection of 8D1-1 and 8D1-2 in *S. rochei* Sal35 and *S. coelicolor* M1152/8D1 by HPLC. *S. rochei* Sal35 was fermented in SFM, GYM, YBP, and R3 media and *S. coelicolor* M1152/8D1 was fermented in R3 medium. 8D1-1 and 8D1-2 were not detected

in the *S. rochei* Sal35 extracts. The minor peak in *S. rochei* Sal35 SFM extract was not 8D1-2.

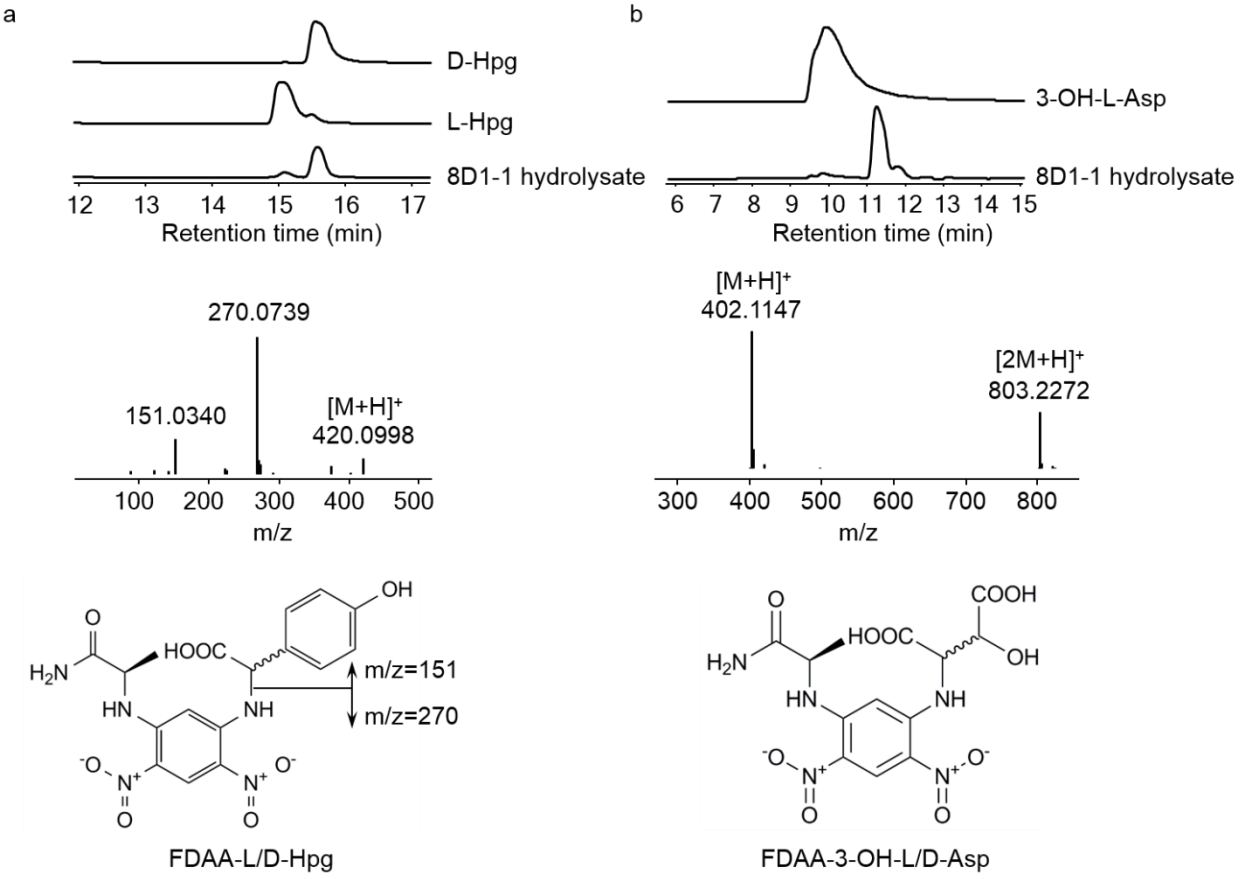


Figure S10. Chiral analysis of Hpg6 (a) and 3-OH-Asn9 (b) in 8D1-1. 8D1-1 acid hydrolysate was derivatized with FDAA and analyzed by LC-MS, and compared to derivatized standards (4). As the terminal amide bond of 3-OH-Asn would also be hydrolyzed during the acid hydrolyzation in 6N HCl to form 3-OH-Asp and 3-OH-D-Asp was not commercially available, so 3-OH-L-Asp was used as the standard to determine the configuration of 3-OH-Asn(1). Hpg6 and 3-OH-Asn9 were determined both in D-configuration. Chemical structures of FDAA-L/D-Hpg and FDAA-3-OH-L/D-Asp and the mass fragmentation of FDAA-L/D-Hpg were also shown.

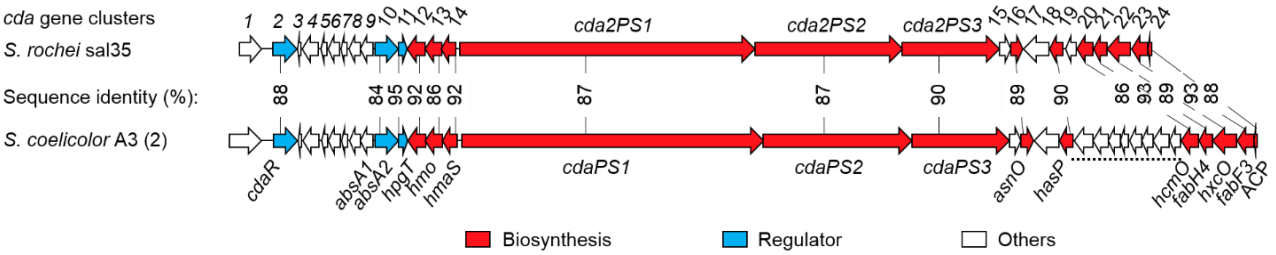


Figure S11. Comparison of gene organization in *cda2* and *cda* pathways from *S. rochei* Sal35 and *S. coelicolor* A3(2). The two pathways share identical gene organization except for a 7 genes region absent in *cda*2 pathway in *S. rochei* Sal35 and underlined with dashed line which is not required for CDA biosynthesis. DNA sequence identity is shown between the two gene clusters.

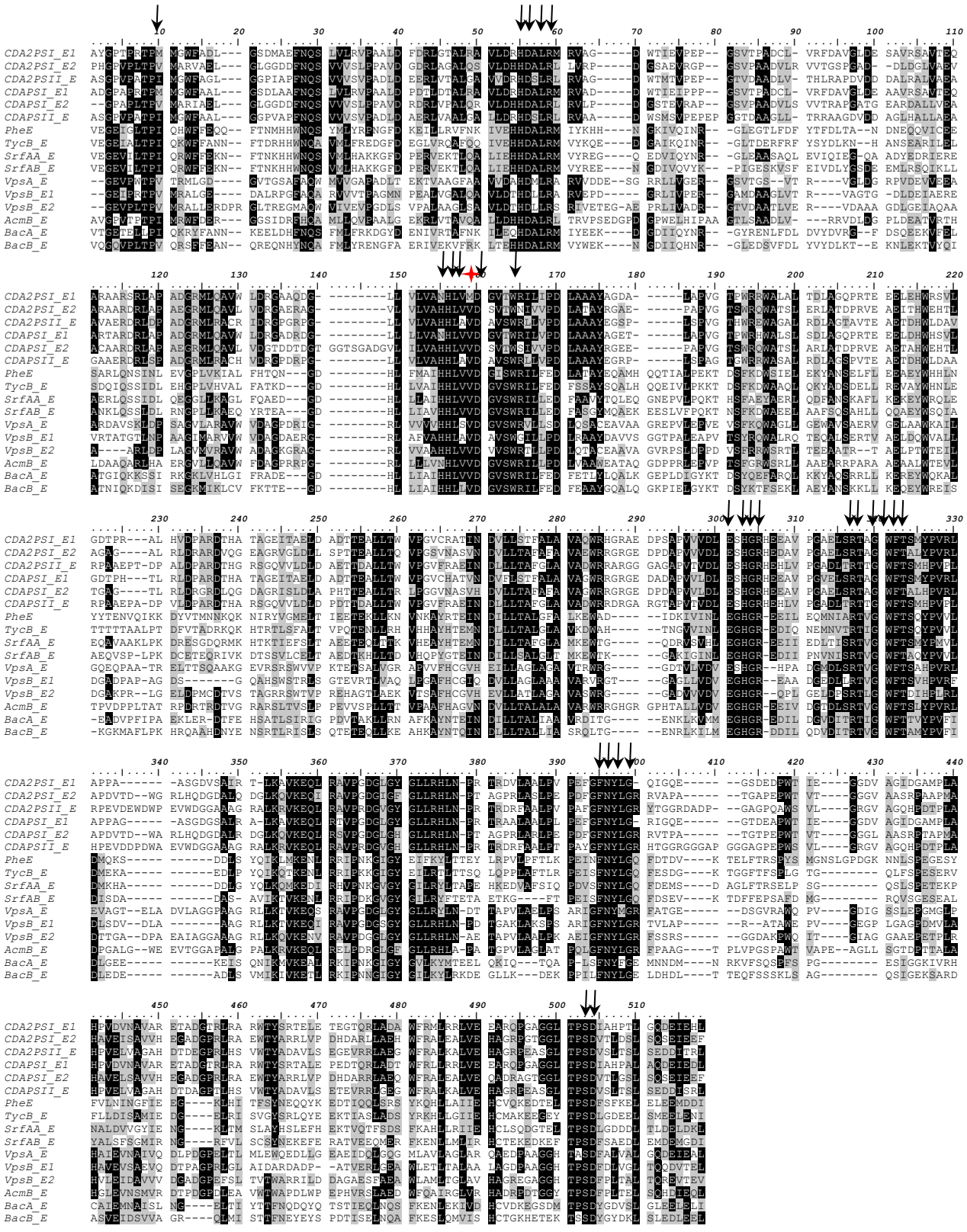
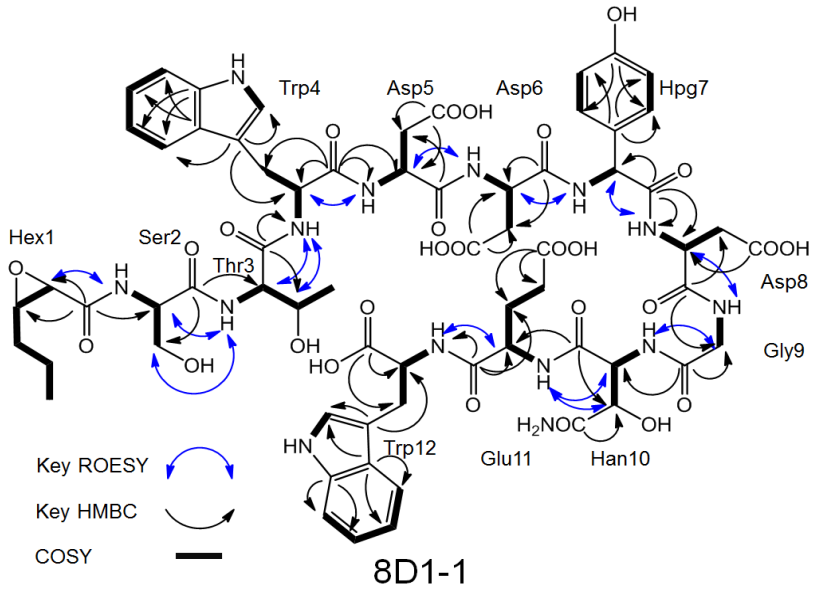


Figure S12. Homology alignment of epimerization domains. The point mutation of conserved Val139 to Met139 was noted with the red asterisk in Cda2PS1 E1 domain. This point mutation along with other point mutations should be responsible for the inactivation of Cda2PS1 E1 domain which resulted in the release of the linear CDAs from the Cda2 NRPS

assembly line． PheE was selected from gramicidin S biosynthesis pathway，TycB_E, SrfAA_E and SrfAB_E, VpsA_E, VpsB_E1, and VpsB_E2, AcmB_E, BacA_E, and BacB_E were from tyrocidine, surfactin, vancomycin, actinomycin, and bacitracin biosynthesis pathway. The conserved amino acids in epimerization domains were labelled by black arrows (5).

Table S1. ^{13}C (201 MHz) and ^{1}H NMR (800 MHz) data of 8D1-1 in deuterated DMSO (DMSO-*d6*).



8D1-1

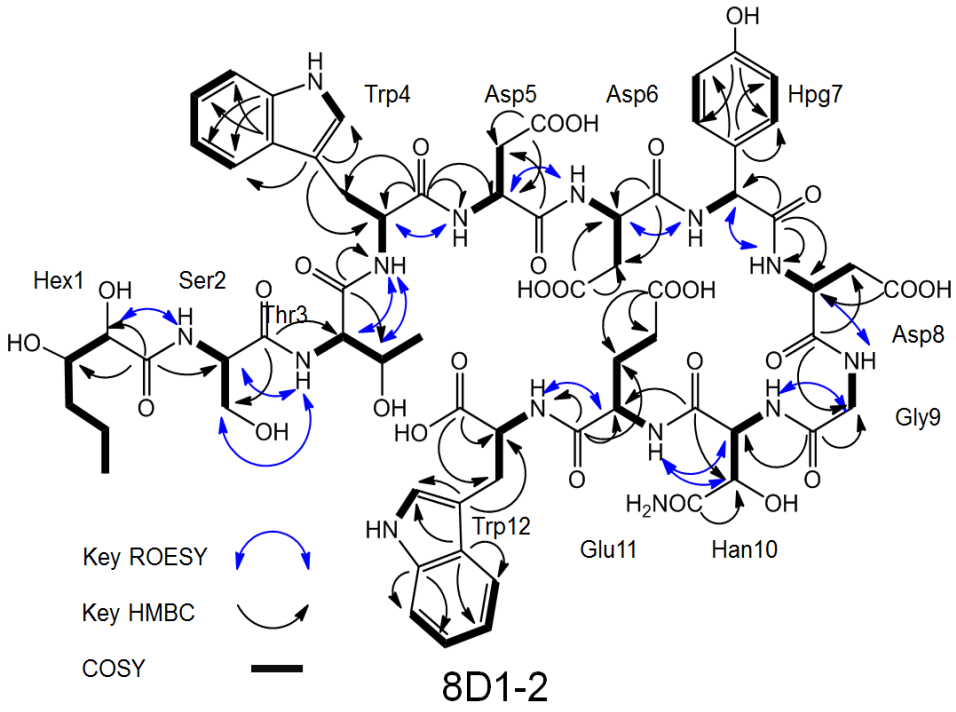
Position	δ_C	δ_H (Mult, J_{H-H})	Position	δ_C	δ_H (Mult, J_{H-H})
Hex1					
1	168.4		Trp4		
2	54.6	3.32 (d, 1.5)	3'	110.3	
3	58.1	3.05 (m)	4'	118.8	7.6 (d, 7.8)
4	33.3	1.45 (m), 1.58 (m)	5'	118.6	6.97 (t, 7.8)
5	19.1	1.41 (m)	6'	121.3	7.05 (t, 7.8)
6	14.1	0.91 (t, 7.3)	7'	111.7	7.31 (d, 7.8)
Ser2			8'	136.5	
1	170.3		9'	127.6	
2	55	4.44 (m)	Asp5		
3	61.9	3.58 (dd, 10.8, 6.0), 3.66 (dd, 10.8, 6.0)	1	170.5	
α-amine		7.98 (br)	2	49.9	4.63 (m)
Thr3			3	36.7	2.39 (m), 2.68 (m)
1	170.4		4	172.4	
2	59	4.11 (d, 4)	α-amine		8.31 (br)
3	66.8	3.93 (m)	Asp6		
4	19.9	0.86 (d, 6.2)	1	171.2	
α-amine		7.93 (br)	2	50.1	4.61 (m)
Trp4			3	36.7	2.56 (m), 2.63 (m)
1	171.8		4	172.4	
2	54.2	4.48 (m)	α-amine		8.21 (br)
3	28.1	2.9 (dd, 13.8, 9.0), 3.16 (dd, 15.1, 4.7)	Hpg7		
α-amine		7.99 (br)	1	170.5	
1'-amine		10.74 (s)	2	56.2	5.36 (d, 7.3)

2'	124.4	7.13 (s)	3 (1')	128.9	

Continued

Position	δ_C	δ_H (Mult, J_{H-H})	Position	δ_C	δ_H (Mult, J_{H-H})
Hpg7			Glu11		
2'	128.7	7.16 (d, 8.3)	1	171.3	
3'	115.4	6.68 (d, 8.3)	2	52.3	4.28 (br)
4'	157.2		3	27.6	1.69 (m), 1.95 (m)
5'	115.4	6.68 (d, 8.3)	4	30.3	2.19 (m), 2.23 (m)
6'	128.7	7.16 (d, 8.3)	5	174.7	
α-amine		8.21 (br)	α-amine		7.61 (br)
Asp8			Trp12		
1	171.8		1	173.8	
2	50.3	4.57 (m)	2	53.5	4.45 (m)
3	37.3	2.48 (m), 2.61 (m)	3	27.6	3.04 (m), 3.18 (dd, 14.6, 5.3)
4	172.4		α-amine		8.19 (br)
α-amine		8.48 (br)	1'amine		10.80 (s)
Gly9			2'	124.2	7.18 (s)
1	169.3		3'	110.1	
2	42.9	3.72 (brd, 12), 3.85 (brd, 12)	4'	118.6	7.52 (d, 7.8)
α-amine		8.31 (br)	5'	118.8	6.99 (t, 7.8)
Han10			6'	121.4	7.06 (t, 7.8)
1	168.9		7'	111.9	7.33 (d, 7.8)
2	56.6	4.62 (m)	8'	136.5	
3	72.5	4.11 (d, 4)	9'	127.6	
4	174				
α-amine		8.08 (br)			

Table S2. ^{13}C (201 MHz) and ^{1}H NMR (800 MHz) data of 8D1-2 in DMSO-*d6*.



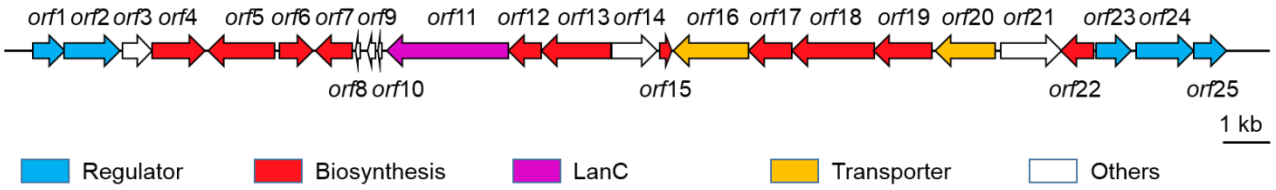
Position	δ_C	δ_H (Mult, J_{H-H})
Hex1		
1	170.9	
2	75.2	4.23 (br)
3	65.4	4.27 (brd, 10.4)
4	33.8	1.6 (m), 1.69 (m)
5	19.8	1.28 (m), 1.47 (m)
6	13.7	0.83 (t, 7.3)
Ser2		
1	170.3	
2	54.7	4.46 (m)
3	62.1	3.55 (m), 3.69 (m)
α-amine		7.91 (brd, 7.6)
Thr3		
1	170.3	
2	58.8	4.15 (brdd, 2.8, 8.0)
3	66.8	3.95 (m)
4	19.9	0.86 (d, 6)
α-amine		7.91 (brd, 7.6)
Trp4		
1	171.9	
2	54.2	4.5 (m)
3	28.2	2.9 (m), 3.15 (m)
α-amine		7.95 (d, 7.3)

Position	δ_C	δ_H (Mult, J_{H-H})
Trp4		
1'-amine		10.75 (br)
2'	124.4	7.15 (s)
3'	110.2	
4'	118.9	7.60 (d, 7.8)
5'	118.6	6.97 (t, 7.8)
6'	121.3	7.05 (t, 7.8)
7'	111.7	7.30 (d, 7.8)
8'	136.5	
9'	127.6	
Asp5		
1	171.2	
2	49.8	4.63 (m)
3	36.6	2.41 (m), 2.68 (m)
4	172.2	
α-amine		8.36 (d, 7.2)
Asp6		
1	170.4	
2	50.1	4.62 (m)
3	36.6	2.55 (m), 2.65 (m)
4	172.2	
α-amine		8.22 (br)

Continued

Position	δ_C	δ_H (Mult, J_{H-H})	Position	δ_C	δ_H (Mult, J_{H-H})
Hpg7			Glu11		
1	170.3		1	171.5	
2	56.1	5.36 (d, 7.3)	2	52.2	4.3 (m)
3 (1')	128.9		3	27.8	1.69 (m), 1.94 (m)
2'	128.6	7.14 (d, 8.0)	4	30.3	2.19 (m), 2.24 (m)
3'	115.4	6.67 (d, 8.0)	5	174.7	
4'	157.2		α-amine		7.65 (br)
5'	115.4	6.67 (d, 8.0)	Trp12		
6'	128.6	7.14 (d, 8.0)	1	173.7	
α-amine		8.17 (br)	2	53.5	4.45 (m)
Asp8			3	27.5	3.04 (m), 3.18 (m)
1	171.3		α-amine		8.22 (br)
2	50.1	4.60 (m)	1'-amine		10.79(br)
3	36.8	2.46 (m), 2.65 (m)	2'	124.2	7.17 (s)
4	172.2		3'	110.1	
α-amine		8.51 (br)	4'	118.6	7.52 (d, 7.8)
Gly9			5'	118.9	6.99 (t, 7.8)
1	169.2		6'	121.4	7.06 (t, 7.8)
2	42.8	3.74 (brdd, 4.5, 16), 3.83 (brdd, 4.5, 16)	7'	111.9	7.34 (d, 7.8)
α-amine		8.22 (br)	8'	136.5	
Han10			9'	127.6	
1	168.8				
2	56.3	4.63 (m)			
3	72.4	4.11 (brd, 3.7)			
4	174				
α-amine		7.98 (br)			

Table S3. Predicted functions of the proteins encoded by the cryptic BGC in contig 4.



orf	Product size (aa)	Homolog (source), accession no.	Identity/ similarity (%)	Conserved domain, accession no.
*orf*1	224	DNA binding response regulator [*Streptomyces* sp. NRRL F-4835], WP_030976652.1	99/99	CitB, COG2197
*orf*2	394	Two-component sensor histidine kinase [*Streptomyces* sp. NRRL F-5650], WP_051851823.1	98/98	HisKA_3, pfam07730
*orf*3	206	hypothetical protein [*Streptomyces* sp. NRRL F-4835], WP_051890672.1	99/99	-
*orf*4	370	3-oxoacyl-ACP synthase [*Streptomyces* sp. NRRL F-4835], WP_030976658.1	100/100	KAS_I_II, cd00834
*orf*5	474	acyl CoA ligase [*Streptomyces aizunensis*], AAX98201.1	44/59	CaiC, COG0318
*orf*6	235	Beta-ketoacyl-ACP reductase [*Streptomyces olivochromogenes*], KUN43576.1	64/75	FabG, PRK05557
*orf*7	255	enoyl-[acyl-carrier-protein] reductase [*Streptomyces* sp. NRRL F-4835], WP_030976664.1	100/100	FabI, COG0623
*orf*8	29	hypothetical protein	-	-
*orf*9	55	hypothetical protein	-	-
*orf*10	29	hypothetical protein	-	-
*orf*11	858	serine/threonine protein kinase [*Nocardia* sp. NRRL S-836], KOV80059.1	49/62	LanC_Ser/Thrkinase, cd04791
*orf*12	234	GNAT family N-acetyltransferase [*Streptomyces* sp. NRRL F-4835], WP_051890673.1	100/100	RimL, COG1670
*orf*13	488	beta-ketoacyl-[acyl-carrier-protein] synthase II [*Streptomyces* sp. NRRL F-5650], WP_051851828.1	88/100	FabF, TIGR03150
*orf*14	316	hypothetical protein [*Streptomyces*], WP_030976672.1	100/100	-
*orf*15	82	acyl carrier protein [*Streptomyces*], WP_030976674.1	100/100	AcpP, PRK00982

Continued

*orf*16	528	transporter, major facilitator family protein [*Streptomyces rimosus*], WP_003981566.1	53/64	MFS_1, pfam07690
*orf*17	302	3-oxoacyl-ACP reductase [*Streptomyces*], WP_030976662.1	100/100	BKR_SDR_c, cd05333
*orf*18	575	acetolactate synthase large subunit [*Streptomyces* sp. NRRL F-5650], WP_031036547.1	99/99	Acolac_lg, TIGR00118
*orf*19	413	aldehyde dehydrogenase [*Streptomyces* sp. NRRL F-5650], WP_051851819.1	99/99	AdhE, COG1012
*orf*20	424	hypothetical protein [*Streptomyces* sp. NRRL F-4835], WP_051890678.1	98/98	2A0121, TIGR00900
*orf*21	430	transcriptional regulator [*Streptomyces* sp. NRRL F-4835], WP_030976691.1	99/99	HTH_XRE, cd00093
*orf*22	228	4'-phosphopantetheinyl transferase [*Streptomyces*], WP_003986505.1	46/58	ACPS, pfam01648
*orf*23	248	two-component system response regulator [*Streptomyces neyagawaensis*], WP_055537867.1	41/52	CitB, COG2197
*orf*24	401	two-component sensor histidine kinase [*Streptomyces* sp. NRRL F-4835], WP_030976698.1	100/100	COG4585, COG4585
*orf*25	231	DNA-binding response regulator [*Streptomyces* sp. NRRL F-4835], WP_051890679.1	100/100	CitB, COG2197

Functions of each protein were annotated by BlastP (http://blast.ncbi.nlm.nih.gov/Blast.cgi). *orf*4, *orf*5, *orf*6, *orf*7, *orf*12, *orf*13, *orf*15, *orf*17, *orf*18, *orf*19, and *orf*22 were PKS related genes, and *orf*11 was a LanC like Ser/Thr kinase gene. We could not tell the structure of the secondary metabolites encoded by this BGC from its sequence. Genes organization of the BGC was also shown upon Table S5.

References

1. **Kempter C, Kaiser D, Haag S, Nicholson G, Gnau V, Walk T, Gierling KH, Decker H, Zähner H, Jung G.** 1997. CDA: Calcium‐Dependent Peptide Antibiotics from *Streptomyces coelicolor* A3 (2) Containing Unusual Residues. Angewandte Chemie International Edition in English **36:**498-501.
2. **Hojati Z, Milne C, Harvey B, Gordon L, Borg M, Flett F, Wilkinson B, Sidebottom PJ, Rudd BA, Hayes MA, Smith CP, Micklefield J.** 2002. Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*. Chem Biol **9:**1175-1187.
3. **Bai T, Yu Y, Xu Z, Tao M.** 2014. Construction of *Streptomyces lividans* SBT5 as an efficient heterologous expression host. Journal of Huazhong Agricultural University **33:**1-6.
4. **Bhushan R, Bruckner H.** 2004. Marfey's reagent for chiral amino acid analysis: a review. Amino Acids **27:**231-247.
5. **Stachelhaus T, Walsh CT.** 2000. Mutational analysis of the epimerization domain in the initiation module PheATE of gramicidin S synthetase. Biochemistry **39:**5775-5787.