The Journal of Antibiotics (2017) 70, 639–646

www.nature.com/ja

ORIGINAL ARTICLE

Molecular beacons to identify gifted microbes for genome mining

Richard H Baltz

Microbial genome mining is a promising technology that is revitalizing natural product discovery. It is now well documented that many bacteria with large genomes, particularly actinomycetes, encode many more secondary metabolites (SMs) than was previously known from their expressed secondary metabolomes. There are effective bioinformatics tools for counting the numbers and nature of SMs, and determining the total coding capacity from finished microbial genomes. However, these methods do not translate well to draft genomes, particularly for large SM gene clusters that contain nonribosomal peptide synthetase (NRPS) or type I polyketide synthase (PKS-I) mega-genes which are prone to fragmentation and misassembly. Small molecular beacons are required to assess the numbers and variety of NRPS, PKS-I and mixed NRPS/PKS-I pathways. In this report, I show that concatenated peptidyl carrier protein-thioesterase di-domains and acyl carrier protein-thioesterase di-domains can be used as multi-probes to survey finished or draft genomes to estimate the numbers of NRPS, PKS-I and mixed NRPS/PKS-I gene clusters to identify gifted actinomycetes.
The Journal of Antibiotics (2017) **70,** 639–646; doi:10.1038/ja.2017.1; published online 25 January 2017

INTRODUCTION

Microbial genome mining is an important emerging technology to discover new and novel natural products (NPs) for drug discovery.[1–4] The genome mining concept stems from the seminal observations of the Hopwood and Ōmura groups that the genome sequences of *Streptomyces coelicolor*[5] and *Streptomyces avermitilis*[6] appeared to encode about 10-fold more potential secondary metabolites (SMs) than were known from the expressed secondary metabolomes. These predictions have been confirmed experimentally,[7,8] and generalized to actinomycetes and other bacterial taxa with large genomes.[9–16]

NPs continue to be important sources of new and novel chemical scaffolds for drug discovery,[17,18] and actinomycetes, particularly *Streptomyces* species, continue to be the most productive sources.[4,19,20] To access the enormous untapped cryptic SM coding capacity of actinomycetes, it is critical to develop robust approaches to activate gene cluster expression. Many approaches, including isolation of mutants altered in transcription or translation, genetic manipulation of positive and negative regulation, heterologous expression in specialized hosts and others, have been described.[8,21–30]

For microbial genome mining to become a robust methodology to drive the discovery of new and novel NPs for drug discovery, it is important to have a set of bioinformatics tools that can predict which microorganisms are the most 'gifted' for SM production. antiSMASH 3.0[31] is particularly useful to identify the numbers and types of SMs encoded by microbes, and has been used to survey a wide range of bacteria and archaea for those microorganisms most gifted for SM production.[16] Among the most gifted are bacteria with genomes >8.0 Mb, including many actinomycetes. None-the-less, having a large genome is not sufficient to predict abundant coding capacity for SMs. Furthermore, antiSMASH 3.0 gives accurate accounting of SM coding capacity for finished genomes, but not for draft genomes which often have poorly assembled NRPS, PKS-I or mixed NRPS/PKS-I (NRPS-PKS-I refers to all three types) gene clusters because of the short reads by the most economical sequencing technologies, and the high sequence similarities within repeating functional domains in modular NRPS and PKS-I mega-genes.[32,33] Since the vast majority of bacterial genomes in public databases remain in draft form, it would be useful to have bioinformatics search methods to predict which microbes are the most gifted to target for complete genome sequencing and genome mining.

Historically, the most productive sources for NP-derived drugs have been biosynthetic pathways that employ NRPS-PKS-I mechanisms.[4] The majority of SM pathways employing these mechanisms in bacteria terminate assembly on mega-enzymes with TEs that release linear or cyclized molecules from terminal modules.[34] TE domains associated with NRPS-PKS-I mega-enzymes encoded by actinomycetes are usually preceded by PCPs or ACPs (see below). The peptidyl carrier protein-thioesterase (PCP-TE) and acyl carrier protein-thioesterase (ACP-TE) di-domains are relatively small, so their DNA sequences should be assembled correctly by any sequencing methodology. There are exceptions to this release strategy. For instance, the NRPS mega-enzymes that encode glycopeptides related to vancomycin and cephamycins have terminal modules with epimerase (E) domains between the PCP and TE,[35,36] and the hybrid PKS-I/NRPS pathways for rapamycin and FK506 cyclize the molecules by a different mechanism.[37,38] In spite of the exceptions, a survey of the number

CognoGen Biotechnology Consulting, Sarasota, FL, USA
Correspondence: Dr RH Baltz, CognoGen Biotechnology Consulting, 7636 Andora Drive, Sarasota, FL 34238, USA.
E-mail: rbaltz923@gmail.com
Received 20 October 2016; revised 12 December 2016; accepted 26 December 2016; published online 25 January 2017

Table 1 Sources of ACP-TE and PCP-TE di-domains

Microorganism	*Position*	*Protein/function*	*Accession #*	*CP-TE size (aa)*	*Reference*
ACP-TE multi-probe					
Streptomyces fradiae ATCC	1	TylGV	AAB66508.1	440	45
Saccharopolyspora erythraea NRRL 2338	2	EryAIII	CAA39538.1	394	9
Saccharopolyspora spinosa NRRL 18395	3	SpnF	AAG23262.1	407	43
Streptomyces sp. CK4412	4	TmcB	ABI94380.1	430	55
Streptomyces avermitilis	5	PKS	WP_010981852.1	411	6
PCP-TE multi-probe					
Streptomyces roseosporus NRRL 11379	1	DptD	AAX31559.1	371	47
Streptomyces griseus IFO 13350	2	NRPS	WP_012377959.1	378	11
Streptomyces avermitilis MA-4680	3	NRPS	WP_010982303.1	377	6
Amycolatopsis orientalis ATCC 19795	4	NRPS	WP_051173837.1	376	56
Streptomyces clavuligerus ATCC 27064	5	NRPS	WP_003953783.1	374	57

of PCP-TE and ACP-TE di-domains per genome might help identify the most gifted microbes. Gifted microbes also tend to encode multiple MbtH homologs, which serve as non-enzymatic chaperones for NRPS adenylation reactions,[39,40] and multiple phosphopantetheinyl transferase (PPTase) genes involved in converting apo-PCPs and apo-ACPs to active holo-enzymes.[16,23] In this report, I describe the use of concatenated PCP-TE and ACP-TE pentamers as beacons to estimate the numbers of NRPS and PKS-I gene clusters in finished and draft genomes from actinomycetes, and also survey the numbers of genes encoding MbtH and PPTases using concatenated multi-probes.[16,40] The results indicate that these four types of multi-probes targeting NRPS-PKS-I mega-enzymes serve as useful beacons to identify gifted microbes from finished or draft genome sequences of actinomycetes.

MATERIALS AND METHODS

Concatenated CP-TE di-domains

The sources of ACP-TE and PCP-TE di-domains are shown in Table 1. Five ACP-TEs and five PCP-TEs were concatenated to generate individual multi-probes (Supplementary Figures S1 and S2) for analyses of actinomycete genomes.

MbtH and PPTase multi-probes

The MbtH and PPTase multi-probes have been described elsewhere.[16,40]

Protein searches

Protein searches were carried out with multi-probes by BLASTp for MbtH homologs and DELTA BLASTp for PPTase, ACP-TE and PCP-TE homologs (http://blast.ncbi.nlm.nih.gov/Blast.cgi).[41,42] The multi-probes also pick up many ACPs and PCPs not linked to TEs, so only full-length hits were counted. Also, each hit was verified manually to correspond to the appropriate ACP or PCP annotation (PKS-I of NRPS); in some cases, ACP-TEs pick up PCP-TEs and vice versa, so false-positive 'hits' were not counted.

Bioinformatic searches for SMGCs

The numbers and sizes of NRPS, PKS-I and mixed NRPS/PKS-I gene clusters in finished and unfinished actinomycetre genomes were carried out by antiSMASH 3.0.[31]

RESULTS

Assembly of NRPS and PKS-I mega-genes in draft genome sequences

To illustrate the problem of assembly of large NRPS and PKS-I gene clusters in draft genomes, BLASTp analysis was carried out on unfinished genomes of the producers of spinosad (*Saccharopolyspora spinosa*), tylosin (*Streptomyces fradiae*) and daptomycin (*Streptomyces roseosporus*) as subjects using PKS-I and NRPS mega-enzymes from finished biosynthetic gene clusters as queries (Table 2). The PKS-I mega-enzyme involved in biosynthesis of spinosad has five subunits, SpnA–SpnE.[43] The draft genome of *S. spinosa*[44] has the *spnA* gene assembled correctly. However, *spnB* is truncated; *spnC* is missing; *spnD* is split into two unlinked segments, one fused to a heterologous sequence and the other fused to two heterologous sequences; and *spnE* is split into two unlinked segments.

The PKS-I mega-enzyme involved in tylosin biosynthesis in *S. fradiae* has five subunits, TylGI–TylGV.[45] The draft genome of *S. fradiae*[46] has a correctly assembled *tylGV*. However, *tylGI* is truncated; the complete *tylGII* gene is fused to a heterologous segment; *tylGIII* is truncated; and *tylGIV* is split into three truncated segments which span only two-thirds of the protein (Table 2).

The NRPS mega-enzyme involved in daptomycin biosynthesis in *S. roseosporus* has three subunits, DptA, DptBC and DptD,[47] which have been exploited in combinatorial biosynthesis.[48] The three subunit arrangement has also been observed in the highly related cryptic daptomycin-like pathway in the finished genome of *Saccharopolyspora viridis*[49] and the finished biosynthetic gene cluster for taromycin A in the marine *Saccharopolyspora* sp. CNQ-490.[50] Draft genomes of two strains of *S. roseosporus* are available on the NCBI website from a sequencing project by the Broad Institute. *S. roseosporus* NRRL 11379 (ATCC 31568; A21978.6) is the wild-type strain discovered by Eli Lilly and Company, and *S. roseosporus* NRRL 15998 (A21978.65) is a more productive derivative of NRRL 11379 derived by *N*-methyl-*N*′-nitro-*N*-nitrosoguanidine mutagenesis.[51] It is instructive that the draft genomes of these highly related strains have the daptomycin gene cluster assembled in two different ways. The wild-type strain (A21978.6) has *dptA*, *dptBC* and *dptD* correctly assembled and in proper order, but has an additional partial sequence of *dptBC*. A21978.65 has a correctly assembled *dptA* gene, but *dptBC* is split into two partial segments, each fused to two heterologous DNA fragments. The *dptD* gene is also split into two segments, one of which is fused to a heterologous DNA segment.

These examples demonstrate that draft genome sequences are not adequate to correctly assemble secondary metabolite gene clusters (SMGCs) that employ NRPS or PKS-I biosynthetic mechanisms, the hallmarks of the most productive sources for clinically useful drugs.[4] This major shortcoming limits the utility of antiSMASH 3.0[31] and

Table 2 Annotation of PKS and NRPS genes in finished clusters and draft genomes

Strain	*Cluster*	*Type*	*Source*	*Protein*[a]	*Accession #*	*Size (AA)*	*Coordinants (AA)*	*Reference*
S. spinosa NRRL 18395	Spn	PKS-I	FC	SpnA	AAG23264.1	2595	1–2595	43
			DG	SpnA	WP_010309393.1	2595	1–2595	44
			FC	SpnB	AAG23265.1	2152	1–2152	43
			DG	SpnB-p	WP_49887651.1	1306	847–2152	44
			FC	SpnC	AAG23266.1	3170	1–3170	43
			DG	SpnC-m	–	–	–	44
			FG	SpnD	AAG23263.1	4928	1–4928	43
			DG	SpnD-pf2	WP_010309374.1	1885	(4)–5–1873–(12)	44
			DG	SpnD-pf1	WP_49887650.1	2638	2150–4776–(12)	44
			FC	SpnE	AAG23262.1	5588	1–5588	43
			DG	SpnE-p	WP_010309366.1	945	1–945	44
			DG	SpnE-p	WP_049887649.1	3427[b]	2140–5588[b]	44
S. fradiae ATCC 19609	Tyl	PKS-I	FC	TylGI	AAB66504.1	4472	1–4472	45, NCBI
			DG	TylGI-p	KDS83975.1	1164	3309–4472	46
			FC	TylGII	AAB66505.1	1864	1–1864	45, NCBI
			DG	TylGII-f	KDS83974.1	1904	(40)–1–1864	46
			FC	TylGIII	AAB66506.1	3729	1–3729	45, NCBI
			DG	TylGIII-p	KDS83971.1	2925	805–3729	46
			FG	TylGIV	AAB66507.1	1611	1–1611	45, NCBI
			DG	TylGIV-p	KDS83972.1	276	1–276	46
			DG	TylGIV-p	KDS83874.1	141	322–462	46
			DG	TylGIV-p	KDS83991.1	657	955–1611	46
			FC	TylGV	AAB66508.1	1841	1–1841	45, NCBI
			DG	TylGV	KDS83992.1	1841	1–1841	46
S. roseosporus NRRL 11379	Dpt	NRPS	FC	DptA	AAX31557.1	5830	1–5830	47
NRRL 11379			DG	DptA	EWS90116.1	5830	1–5830	NCBI
NRRL 15998			DG	DptA	EFE72875.1	5830	1–5830	NCBI
NRRL 11379			FC	DptBC	AAX31558.1	7338	1–7338	47
NRRL 11379			DG	DptBC	EWS90115.1	7338	1–7338	NCBI
NRRL 11379			DG	DptBC-p	EWS96377.1	398	5217–5614	NCBI
NRRL 15998			DG	DptBC-pf2	EFE72850.1	2946	(392)–4784–7221–(117)	NCBI
NRRL 15998			DG	DptBC-pf2	EFE72873.1	1093	(29)–2611–3667–(66)	NCBI
NRRL 11398			FC	DptD	AAX31559.1	2379	1–2379	47
NRRL 11397			DG	DptD	EWS90114.1	2379	1–2379	NCBI
NRRL 15998			DG	DptD-pf1	EFE72872.1	1350	1–1307–(43)	NCBI
NRRL 15998			DG	DptD-p	EFE72871.1	1029	1351–2379	NCBI

Abbreviations: AA, amino acid; DG, draft genome; Dpt, daptomycin; FC, finished cluster; NCBI, National Center for Biotechnology Information; Spn, spinosad; Tyl, tylosin.
[a]Protein designations: XxxX, intact protein; XxxX-f, heterologous AAs () fused to intact protein; XxxX-m, missing protein; XxxX-p, partial protein; XxxX-pf1, partial protein fused to heterologous AAs at one end (); XxxX-pf2, partial protein fused to heterologous AAs at both ends ().
[b]WP_049887649.1 also has an internal deletion of 22 AA.

other bioinformatics tools that require high-quality finished genome sequences for reliable predictions,[16] as further demonstrated below.

Distribution of NRPS, PKS-I and mixed NRPS/PKS-I clusters in finished and draft genomes

To further exemplify the problem of incorrect assembly of large SMGCs containing NRPS-PKS-I mega-genes, I have compiled the numbers and sizes of clusters employing these biosynthetic mechanisms determined by antiSMASH 3.0 analyses of ten *Streptomyces* genomes, five finished and five drafts (Table 3). The finished genomes encode 7–26 NRPS-PKS-I clusters ranging from 41.4 to 246.2 kb, and averaging 70.5 to 94.7 kb. Of the 71 total clusters, 41 were >60 kb, and none were <40 kb. In contrast, among the 69 clusters from the unfinished genomes, which ranged from 4.2 to 79.7 kb, and averaged 26.5 to 50.5 kb, only 5 were >60 kb and 40 were <40 kb. It is apparent that draft genomes have many fragmented, and therefore incorrectly assembled NRPS-PKS-I gene clusters.

Construction of ACP-TE and PCP-TE multi-probes

Many NRPS-PKS-I biosynthetic pathways encode terminal modules containing ACP-TE or PCP-TE di-domains to release the nascent peptides, polyketides or mixed peptide-polyketides from the mega-enzymes as linear or cyclized intermediates or final products. A survey of 37 important NPs produced by actinomycetes and biosynthesized by these mechanisms indicated that 25 (68%) have terminal modules with APC-TE or PCP-TE di-domains (Supplementary Figure S3). Individual ACP-TEs and PCP-TEs are relatively small, ranging from 394 to 440 amino acids and 371 to 378 amino acids, respectively (Table 1). As such, their coding domains should remain together in genome assemblies regardless of the sequencing technologies employed. The ACP-TE pentamer was constructed from well-characterized di-domains from PKS subunits from tylosin, erythromycin, spinosad and tautomycetin pathways and an uncharacterized PKS from *S. avermitilis*. The PCP-TE pentamer was constructed from the third subunit of the daptomycin NRPS cluster (DptD) and four

Table 3 Distribution of NRPS, PKS-I and mixed NRPS/PKS-I cluster sizes in finished and draft *Streptomyces* genomes

	Cluster sizes in finished genomes (kb)					*Cluster sizes in draft genomes (kb)*				
Cluster	*S. rap*	*S. gri*	*S. ave*	*S. alb*	*S. amb*	*S. kan*	*S. wed*	*S. hal*	*S. cha*	*S. fra*
1	246.2	112.4	129.6	155.2	169.6	59.0	54.6	126.0	55.9	79.7
2	182.1	107.8	108.2	105.7	176.1	57.9	49.7	88.5	54.6	60.6
3	172.1	102.4	104.8	74.1	68.7	57.5	45.0	62.8	48.8	47.6
4	138.3	102.2	104.0	67.3	61.3	52.2	40.4	54.1	48.3	11.4
5	134.4	96.3	80.1	63.1	51.0	50.5	33.0	50.9	44.0	10.1
6	123.9	85.6	61.6	58.7	47.4	45.9	32.9	50.0	42.0	9.7
7	118.1	77.4	56.5	50.3	41.4	43.3	31.2	49.5	30.9	
8	115.8	72.8	52.1	49.4		39.3	19.3	48.5	29.3	
9	111.4	65.3	51.6	44.3		36.4	14.9	45.1		
10	101.8	61.9	50.1			34.3	12.3	36.2		
11	94.5	54.8	49.8			33.8	5.4	34.3		
12	91.1	52.5	47.3			31.9	5.3	31.1		
13	84.1	51.4	46.2			31.2	4.2	30.0		
14	80.9	49.5	44.6			30.1				
15	78.8	45.8				29.4				
16	76.9					29.0				
17	57.8					25.5				
18	54.4					24.0				
19	53.4					23.8				
20	53.3					22.1				
21	53.0					19.1				
22	51.5					16.8				
23	51.0					14.7				
24	49.2					13.3				
25	44.0					12.5				
26	43.8					9.4				
27						7.5				
28						5.7				
29						5.3				
Sum	2461.8	1138.1	986.5	668.1	615.5	861.4	348.2	657.0	353.8	219.1
Ave	94.7	75.9	70.5	74.2	87.9	29.7	26.8	50.5	44.2	36.5

Abbreviations: S. alb, *Streptomyces albus* J1074; S. amb, *Streptomyces ambofaciens*; S. ave, *Streptomyces avermitilis*; S. cha, *Streptomyces chartreusis*; S. fra, *Streptomyces fradiae*; S. gri, *Streptomyces griseus*; S. hal, *Streptomyces halstedii*; S. kan, *Streptomyces kanamycinicus*; S. rap, *Streptomyces rapamycinicus*; S. wed, *Streptomyces wedmorensis*.
Strain numbers are shown in Table 5.

uncharacterized NRPS subunits from *Streptomyces griseus*, *S. avermitilis*, *Amycolatopsis orientalis* and *Streptomyces clavuligerus*. The use of five diverse CP-TE di-domains in each case, coupled with the use of DELTA BLASTp,[48] gives high likelihood that individual ACP-TE and PCP-TE di-domains in finished and draft genomes will be counted as surrogates for the number of NRPS-PKS-I gene clusters, independent of poor assembly and fragmentation of large pathways into smaller erroneous SMGCs in antiSMASH 3.0 searches of draft genomes. It is instructive that the ACP-TE multi-probe readily picks up the terminal ACP-TEs of SpnE and TylGV, and the PCP-TE multi-probe picks up the terminal PCP-TE of DptD in draft genomes, even though the tylosin, spinosad and daptomycin gene clusters were assembled incorrectly (Table 1). The multi-probes count each pathway one time, regardless of the other mistakes in gene cluster assembly.

Survey of PCP-TE and ACP-TE di-domains in finished actinomycete genomes

To establish the correlation between the numbers of CP-TEs and NRPS-PKS-I gene clusters, 25 finished actinomycete genomes ranging from 3.64 to 12.7 Mb were surveyed for the numbers CP-TEs by DELTA BLASTp with the two multi-probes, and for NRPS-PKS-I gene clusters by antiSMASH 3.0 (Table 4). In addition, the numbers of MbtH homologs and PPTases encoded by these strains[23,40] are also shown. The numbers of NRPS-PKS-I gene clusters generally increased in proportion to geneome size, ranging from 1 in *Thermobifida fusca* (genome, 3.64 Mb) to 26 in *Streptomyces rapamycinicus* (genome, 12.7 Mb) (Table 4; Figure 1). The numbers of CP-TEs also increased with genome size (Table 4), and the ratios of CP-TEs/NRPS-PKS-I clusters ranged from 0.46 to 1.0, with a mean of 0.65 and a slope of ~0.7 (Figure 2a), consistent with data from well-characterized SMCs (Supplementary Figure S3). It is noteworthy that the four most 'gifted' G actinomycetes (*Kutzneria albida*, *Streptomyces violaceusniger*, *Streptomyces bingchenggensis* and *Streptomyces rapamycinicus*), which devote 2.3–3.1 Mb of their very large genomes (9.9–12.7 Mb) to SM biosynthesis and encode 43–53 SMGCs,[16] encode 19–26 NRPS-PKS-I clusters and 13-21 CP-TEs. These stand out as highly gifted by counting NRPS-PKS-I clusters or CP-TEs.

It is noteworthy that the average numbers of MbtH and PPTase genes for the 25 strains were 4.0 and 3.5, respectively, and the most gifted strains generally encode higher total numbers (Table 4), as demonstrated previously.[16,23] The two least gifted strains, *Thermobifida fusca* YX and *Pseudonocardia dioxanivorans* CB1190, encode only single MbtH and PPTase proteins, and a single NRPS-PKS-I cluster with one CP-TE di-domain.

Table 4 CP-TEs and NRPS-PKS-I gene clusters in finished actinomycete genomes

Strain	*Genome size (MB)*[a]	*MbtH genes*[b]	*PPTase genes*[b]	*NRPS-PKS-I clusters*[c]	*CP-TEs*[d]	*CP-TEs per cluster*
Thermobifida fusca YX	3.64	1	1	1	1	1.0
Saccharomonospora viridis P101	4.31	2	3	2	2	1.0
Salinispora tropica CNB-440	5.18	3	5	8	6	0.75
Nocardia farcinica IFM 10152	6.02	2	2	9	6	0.67
Streptomyces albus J1074	6.84	3	3	9	5	0.56
Micromonospora aurantiaca ATCC 27029	7.03	2	4	6	4	0.67
Pseudonocardia dioxanivorans CB1190	7.10	1	1	1	1	1.0
Saccharopolyspora erythraea NRRL 2338	8.21	3	3	13	7	0.54
Actinosynnema mirum DSM 43827^T	8.25	7	2	17	8	0.46
Streptomyces collinus Tü 365	8.27	4	6	11	6	0.54
Streptomyces ambofaciens ATCC 23877	8.39	2	4	7	5	0.71
Streptomyces griseus NBRC 13350	8.55	8	3	15	8	0.53
Streptomyces coelicolor A3(2)	8.67	2	3	8	5	0.63
Actinoplanes missouriensis ATCC 14538^T	8.77	2	2	4	2	0.50
Amycolatopsis orientalis HCCB10007	8.95	4	4	13	13	1.0
Streptomyces avermitilis MA-4680	9.03	4	6	13	6	0.46
Saccharothrix espanaensis DSM 44229^T	9.36	6	1	14	8	0.57
Kutzneria albida DSM 43870^T	9.88	9	4	21	15	0.71
Streptomyces hygroscopicus 5008	10.15	5	5	13	11	0.85
Amycolatopsis mediterranei U32	10.24	6	3	17	8	0.47
Streptosporangium roseum DSM 43021^T	10.34	8	4	13	13	1.00
Streptomyces violaceusniger Tu4113	10.66	3	4	19	13	0.68
Streptomyces bingchenggensis BCW-1	11.94	4	7	26	21	0.81
Streptomyces rapamycinicus NRRL 5491	12.70	7	3	25	17	0.68
Average	8.44	4.0	3.5	11.9	8.0	0.65

[a]Data from Baltz[16] and NCBI.
[b]Data from BLASTp and DELTA BLASTp with MbtH[40] and PPTase[16] multi-probes, respectively.
[c]NRPS, PKS-I and mixed NRPS/PKS-I clusters determined from antiSMASH 3.0 analyses.[31]
[d]Data from DELTA BLASTp anlayses with PCP-TE and ACP-TE multi-probes.

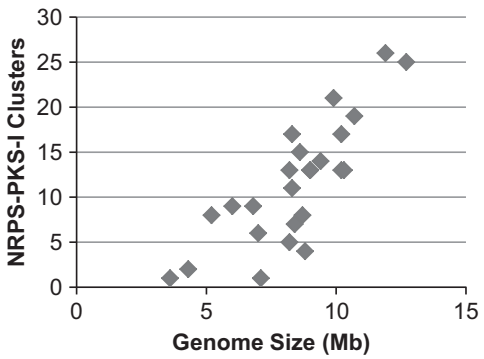

Figure 1 NRPS-PKS-I clusters as a function of actinomycete genome size. Data from Table 4.

Survey of PCP-TE and ACP-TE di-domains in draft actinomycete genomes

Less than 10% of large actinomycete genomes in public databases are finished (Baltz, unpublished). To further examine the use of the CP-TE multi-probes to search for gifted actinomycetes among draft genomes, 30 strains from two populations were chosen for anti-SMASH 3.0 and multi-probe analyses. The first population included 15 draft genomes from producers of important secondary SMs,[4] and the other 15 were chosen from a group of 369 unspeciated *Streptomyces* isolates (https://www.ncbi.nlm.nih.gov/genome/?term=streptomyces) with genomes >8.4 Mb (Table 5). The first set had an average of 14.7 NRPS-PKS-I 'clusters', and 5.1 CP-TEs per strain, giving a ratio of CP-TEs/NRPS-PKS-I clusters of 0.35, or approximately one-half that observed with finished genomes. The 0.35 ratio suggests that this set of draft genomes has many fragmented NPRS-PKS-I clusters, and the number of clusters is overestimated by antiSMASH 3.0 by ~2-fold. The second set had an average of 11.6 NRPS-PKS-I clusters and 2.8 CP-TEs, giving a ratio of CP-TEs/NRPS-PKS-I clusters of 0.24, suggesting that the numbers of NPRS-PKS-I clusters are overestimated by >2-fold. Figure 2b shows a scatter plot of these data. Note that there is much more scatter in the plot of CP-TEs versus NRPS-PKS-I clusters with draft genomes than with finished genomes (Figure 2a). This is likely due to the variable quality in assemblies of NRPS-PKS-I clusters is draft genomes. The regression lines in Figures 2a and b are drawn to reflect the average ratios calculated in Table 5. In spite of the fragmentation of actual NRPS-PKS-I clusters in draft genomes, the six most gifted *Streptomyces* sp., encoding 8–12 CP-TEs, were easily identified by multi-probe analyses. These strains also encode 1–6 MbtH and 2–3 PPTase proteins, and could be candidates for complete genome sequencing.

DISCUSSION

Microbial genome mining is a promising approach to discover new and novel NPs for drug delopment.[3,4,16] It is now clear that microbes with large genomes encode the largest numbers of SMs, and that the uncultured microbial majority, which generally have small genomes, are nearly devoid of SM pathways encoding drug-like NPs.[16] Historically, the majority of important NP drugs were biosynthesized by NRPS, PKS-I or mixed NRPS/ PKS-I mechanisms, mostly by actinomycetes.[4] It has been demonstrated that among the actiomycetes, there are some strains that are gifted or highly gifted for SM biosynthesis, encoding 20–53 SMs, and dedicating 0.8–3.1 Mb of DNA

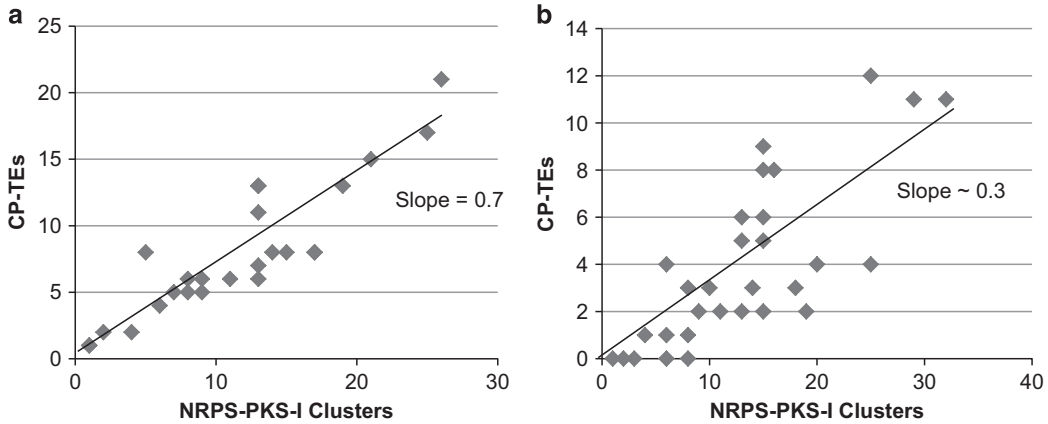

Figure 2 Relationship between CP-TEs and NRPS-PKS-I clusters in finished (**a**) and draft (**b**) actinomycete genomes.

Table 5 CP-TEs and NRPS-PKS-I gene clusters in draft genomes of actinomycetes

Microorganism	*Size (Mb)*[a]	*MbtH genes*[b]	*PPTase genes*[b]	*NRPS-PKS-I clusters*[c]	*CP-TEs*[d]	*CP-TEs/cluster*
Actinomycete species with draft genomes						
Streptomyces fradiae ATCC 19609	7.67	2	2	6	4	0.60
Streptomyces tsukubaensis NRRL 18488	7.67	0	2	25	4	0.16
Streptomyces halstedii NRRL ISP-5068	7.74	4	2	13	6	0.46
Streptomyces roseosporus NRRL 11379	7.85	6	4	13	5	0.38
Streptomyces roseosporus NRRL 15998	7.82	6	4	15	5	0.33
Saccharopolyspora spinosa NRRL 18395	8.58	2	2	15	9	0.60
Streptomyces ghanaensis ATCC 14672	8.60	3	2	8	3	0.38
Streptomyces natalensis ATCC 27448	8.65	2	1	9	2	0.22
Streptomyces viridochromogenes DSM 40736	8.65	9	3	11	2	0.18
Streptomyces anulatus ATCC 1523	8.76	1	4	16	8	0.50
Streptomyces chartreusis NRRL 12338	9.1	0	1	8	0	0.0
Streptomyces wedmorensis NRRL 3426	9.38	6	3	13	2	0.15
Streptomyces peucetius NRRL WC-3868	9.53	4	2	25	12	0.48
Streptomyces kanamyceticus NRRL B-2535	9.78	6	2	29	11	0.38
Dactylosporangium aurantiacum NRRL B-8018	11.45	7	3	14	3	0.21
Average	8.18	3.9	2.5	14.7	5.1	0.35
Streptomyces sp. with draft genomes						
Streptomyces sp. TP-0356	8.41	0	2	8	1	0.13
Streptomyces sp. C	8.46	1	3	6	0	0.0
Streptomyces sp. CNY243	8.49	3	3	15	2	0.13
Streptomyces sp. MspMP-M5	8.60	5	3	19	2	0.11
Streptomyces sp. BoleA5	8.65	1	1	3	0	0.0
Streptomyces sp. PsTaAH-124	8.81	1	2	18	3	0.17
Streptomyces sp. AcH505	9.0	1	1	6	1	0.17
Streptomyces sp. HmicA12	9.16	3	2	2	0	0.0
Streptomyces sp. AA4	9.18	5	1	10	3	0.3
Streptomyces sp. 351MFTsu5.1	9.33	0	2	4	1	0.25
Streptomyces sp. WM6378	9.49	5	2	15	6	0.4
Streptomyces sp. 150FB	9.8	4	2	15	8	0.53
Streptomyces sp. NRRL F5122	10.16	1	1	1	0	0.0
Streptomyces sp. A558	10.58	9	2	20	4	0.2
Streptomyces sp. RD22 (RTd22)	11.19	4	3	32	11	0.34
Average	9.29	2.9	2.0	11.6	2.8	0.24

[a]Data from NCBI.
[b]Data from BLASTp and DELTA BLASTp with MbtH[40] and PPTase[16] multi-probes, respectively.
[c]NRPS, PKS-I and mixed NRPS/PKS-I clusters determined from antiSMASH 3.0 analyses.[31]
[d]Data from DELTA BLASTp anlayses with PCP-TE and ACP-TE multi-probes.

coding capacity to SMGCs. These gifted strains also encode multiple NRPS-PKS clusters.[16]

The analysis of gifted status among a wide range of microbes with finished genome sequences was carried out by using the standard antiSMASH 3.0 algorithm.[16,31] However, the quality and reliability of antiSMASH 3.0 analysis is limited by the quality of genome seqeuences analyzed. In this report, I demonstrate that the large SMGCs encoding daptomycin, spinosad and tylosin were assembled incorrectly in draft genome sequences, and that draft genomes generally tend to have fragmented assemblies of NRPS-PKS-I gene clusters, resulting over-estimation of cluster numbers by antiSMASH 3.0. This unfortunate outcome of the current acceptance of draft genome quality for publication and deposition of genome sequences makes it difficult to mine genomic data for SMGCs encoding drug-like molecules. To address this shortcoming, it is necessary to first sort through genomic data with small DNA sequences that can serve as beacons for desired drug-like pathways. In the current report, multi-probes directed at di-domains containing TE functionality were evaluated to identify gifted actinomycetes that encode multiple NRPS, PKS-I and mixed NRPS/PKS-I-derived SMs. The approach derives from the observation that the majority of NRPS and PKS-I mega-enzymes have terminal modules containing single TE domains preceded by small PCP or ACP domains, respectively. Exceptions include NRPSs that terminate with D-amino acids (for example, vanocmycin and cephamycin which have PCP-E-TE tri-domains)[35,36] and the PKS-I/NRPSs rapamycin and KF506, which terminate assembly by a different mechanism.[37,38] NRPSs that insert terminal D-amino acids could be identified by an extending the PCP-TE multi-probe to include PCP-E-TE tri-domains from the vancomycin and cephamycin pathways. Rapamycin/FK506-like SMGCs can be identified by BLASTp searching with pathway-specific probes (for example, cyclodeaminase).[37,38]

The sizes of the individual PCP-TE and ACP-TE di-domains assembled in the multi-probes are small, ranging from 371 to 440 amino acids, and their coding sequences are likely to be assembled correctly by any DNA sequencing method employed. Thus CP-TE multi-probes can be used as molecular beacons to count NRPS-PKS-I clusters, even if the actual gene clusters are fragmented and mis-assembled. The ratio of CP-TEs/NRPS-PKS-I clusters in 25 finished actinomycete genomes was ~0.7. Thus a count of 7 CP-TEs by DELTA BLASTp analysis predicts ~10 NRPS-PKS-I clusters. The ratio of CP-TEs/NRPS-PKS-I clusters in 30 draft genomes was ~0.3, but with much scatter, consistent with gene cluster splitting and variable quality in assembly. None-the-less, the six most gifted strains were easily identified by the CP-TE multi-probes. Those six strains encode 8–12 CP-TEs, which predict ~11–17 NRPS-PKS-I clusters.

Only 15 large genomes among the 369 unspeciated *Streptomyces* draft genomes available on the NCBI website were surveyed. The CP-TE multi-probe analysis could be used to identify the top 10–20% of gifted streptomycetes in this group and also to screen much larger libraries of proprietary draft actinomycete genomes. The most gifted strains could be targeted for complete genome sequencing, which is now feasible and relatively inexpensive by using a combination of Illumina and PacBio sequencing.[16,52–54] By adding additional CP-TEs from other taxons, the CP-TE multi-probe approach can be adapted to survey draft genomes of other bacteria with large genomes, such as species of the Myxobacteria, *Burkholderia* and *Photorhabdus*, all of which have gifted members that encode multiple SMGCs.[16]

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

I thank the editors of the *Journal of Antibiotics* for inviting me to write an article for this special issue dedicated to Professor Satoshi Ōmura to help celebrate the honor of being awarded the Nobel Prize in medicine for the discovery of avermectin. I thank Professor Ōmura and his colleagues for completing the genome sequence of *Streptomyces avermitilis*, the avermectin producer, which stands as a model 'gifted' microorganism that has provided so much for so many.

1 Baltz, R. H. Renaissance in antibacterial discovery from actinomycetes. *Curr. Opin. Pharmacol.* **8**, 557–563 (2008).
2 Zerikly, M. & Challis, G. L. Strategies for the discovery of new natural products by genome mining. *ChemBioChem* **10**, 625–633 (2009).
3 Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.* **41**, 175–184 (2014).
4 Katz, L. & Baltz, R. H. Natural product discovery: past, present and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
5 Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
6 Ikeda, H. *et al.* Complete genome sequence of and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* **21**, 526–531 (2003).
7 Challis, G. L. Exploitation of the *Streptomyces coelicolor* A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* **41**, 219–232 (2014).
8 Ikeda, H., Shin-ya, K. & Ōmura, S. Genome mining of the *Streptomyces avermitilis* genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. *J. Ind. Microbiol. Biotechnol.* **41**, 233–250 (2014).
9 Oliynyk, M. *et al.* Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL 2338. *Nat. Biotechnol.* **25**, 447–453 (2007).
10 Udwary, D. W. *et al.* Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc. Nat. Acad. Sci. USA* **104**, 10376–10382 (2007).
11 Ohnishi, Y. *et al.* Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **190**, 4050–4060 (2008).
12 Baranasic, D. *et al.* Draft genome sequence of *Streptomyces rapamycinicus* strain NRRL 5491, the producer of the immunosuppressant rapamycin. *Genome Announc.* **1**, e00581–13 (2013).
13 Aigle, B. *et al.* Genome mining of *Streptomyces ambofaciens*. *J. Ind. Microbiol. Biotechnol.* **41**, 251–264 (2014).
14 Zaburannyi, N., Rabyk, M., Ostach, B., Federenko, V. & Luzhetskyy, A. Insights into naturally minimized *Streptomyces albus* J1074 genome. *BMC Genomics* **15**, 97 (2014).
15 Iftime, D. *et al.* Identification and activation of novel biosynthetic gene clusters by genome mining in the kirromycin producer Tü 365. *J. Ind. Microbiol. Biotechnol.* **43**, 277–291 (2016).
16 Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.* (e-pub ahead of print 12 August 2016; doi:10.1007/s10295-016-1815-x).
17 Demain, A. L. Importance of microbial natural products and the need to revitalize their discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 185–201 (2014).
18 Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).
19 Bérdy, J. Bioactive microbial metabolites. *J. Antibiot.* **58**, 1–26 (2005).
20 Barka, E. A. *et al.* Taxonomy, physiology, and natural products of *Actinobacteria*. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2015).
21 Baltz, R. H. *Streptomyces* and *Saccharopolyspora* hosts for heterologous expression of secondary metabolite gene clusters. *J. Ind. Microbiol. Biotechnol.* **37**, 759–772 (2010).
22 Baltz, R. H. Strain improvement in actinomycetes in the postgenomic era. *J. Ind. Microbiol. Biotechnol.* **38**, 657–666 (2011).
23 Baltz, R. H. Genetic manipulation of secondary metabolite biosynthesis for improved production in *Streptomyces* and other actinomycetes. *J. Ind. Microbiol. Biotechnol.* **43**, 343–370 (2016).
24 Ochi, K., Tanaka, Y. & Tojo, S. Activating the expression of bacterial cryptic genes by *rpoB* mutations in RNA polymerase or by rare earth elements. *J. Ind. Microbiol. Biotechnol.* **41**, 403–414 (2014).
25 Tanaka, Y. *et al.* Antibiotic overproduction by *rpsL* and *rsmG* mutants of various actinomycetes. *Appl. Environ. Microbiol.* **75**, 4919–4922 (2009).
26 Ochi, K. & Hosaka, T. New strategies for drug discovery: activation of silent or weakly expressed microbial gene clusters. *Appl. Microbiol. Biotechnol.* **97**, 87–98 (2013).
27 Komatsu, M. *et al.* Engineered *Streptomyces avermitilis* host for heterologous expression of biosynthetic gene cluster for secondary metabolites. *ACS Synth. Biol.* **2**, 384–396 (2013).

28 Gomez-Escribano, J. P. & Bibb, M. J. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb. Biotechnol.* **4**, 207–215 (2011).
29 Gomez-Escribano, J. P. & Bibb, M. J. Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: from genome mining to manipulation of biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* **41**, 425–431 (2014).
30 Zhu, H., Sandiford, S. K. & van Wezel, G. P. Triggers and cues that activate antibiotic production by actinomycetes. *J. Ind. Microbiol. Biotechnol.* **41**, 371–386 (2014).
31 Weber, T. *et al.* antiSMASH 3.0–a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237–W243 (2015).
32 Diminic, J. *et al.* Evolutionary concepts in natural products discovery: what actinomycetes have taught us. *J. Ind. Microbiol. Biotechnol.* **41**, 211–217 (2014).
33 Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
34 Horsman, M. E., Hari, T. P. & Boddy, C. N. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? *Nat. Prod. Rep.* **33**, 183–202 (2016).
35 Pootoolal, J. *et al.* Assembling the glycopeptide antibiotic scaffold: the biosynthesis of A47934 from *Streptomyces toyocaensis* NRRL 15009. *Proc. Nat. Acad. Sci. USA* **99**, 8962–8967 (2002).
36 Tahlan, K., Moore, M. A. & Jensen, S. E. δ-(L-α-aminoadipyl)-L-cysteinyl-D-valine synthetase (ACVS): discovery and perspectives. *J. Ind. Microbiol. Biotechnol.* (e-pub ahead of print 20 October 2016; doi:10.1007/s10295-016-1850-7).
37 Ban, Y. H., Park, S. R. & Yoon, Y. J. The biosynthetic pathway for FK506 and its engineering: from past achievements to future prospects. *J. Ind. Microbiol. Biotechnol.* **43**, 389–400 (2016).
38 Yoo, Y. J., Kim, H., Park, S. R. & Yoon, Y. J. An overview of rapamycin: from discovery to future perspectives. *J. Ind. Microbiol. Biotechnol.* (e-pub ahead of print 09 September 2016; doi:10.1007/s10295-016-1834-7).
39 Baltz, R. H. Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery. *J. Ind. Microbiol. Biotechnol.* **38**, 1747–1760 (2011).
40 Baltz, R. H. MbtH homology codes to identify gifted microbes for genome mining. *J. Ind. Microbiol. Biotechnol.* **41**, 357–369 (2014).
41 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
42 Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).
43 Waldron, C. *et al.* Cloning and analysis of the spinosad biosynthetic gene cluster of *Saccharopolyspora spinosa*. *Chem. Biol.* **8**, 487–499 (2001).
44 Pan, Y. *et al.* Genome sequence of the spinosyns-producing bacterium *Saccharopolyspora spinosa* NRRL 18395. *J. Bacteriol.* **193**, 3150–3151 (2011).
45 Cundliffe, E. Control of tylosin biosynthesis in *Streptomyces fradiae*. *J. Microbiol. Biotechnol.* **18**, 1485–1491 (2008).
46 Bekker, O. B. *et al.* Draft genome sequence of *Streptomyces fradiae* ATCC 19609, a strain highly sensitive to antibiotics. *Genome Announc.* **2**, e01247–14 (2014).
47 Miao, V. *et al.* Daptomycin biosynthesis in *Streptomyces roseosporus*: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151**, 1507–1523 (2005).
48 Baltz, R. H. Combinatorial biosynthesis of cyclic lipopeptide antibiotics: a model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways. *ACS Synth. Biol.* **3**, 748–758 (2014).
49 Baltz, R. H. Genomics and the ancient origins of the daptomycin biosynthetic gene cluster. *J. Antibiot.* **63**, 506–511 (2010).
50 Yamanaka, K. *et al.* Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc. Nat. Acad. Sci. USA* **111**, 1957–1962 (2014).
51 McHenney, M. A., Hosted, T. J., Dehoff, B. S., Rosteck, P. R. & Baltz, R. H. Molecular cloning and physical mapping of the daptomycin gene cluster from *Streptomyces roseosporus*. *J. Bacteriol.* **180**, 143–151 (1998).
52 Gomez-Escribano, J. P. *et al.* The *Streptomyces leeuwenhoekii* genome: *de novo* sequencing and assembly in single contigs of the chromosome, circular plasmid pSLE1 and linear plasmid pSLE2. *BMC Genomics* **16**, 485 (2015).
53 Gomez-Escribano, J. P., Alt, S. & Bibb, M. J. Next generation sequencing of actinobactereia for the discovery of novel natural products. *Mar. Drugs* **14**, 78 (2016).
54 Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
55 Choi, S. S., Nah, H. J., Pyeon, H. R. & Kim, E. S. Biosynthesis, regulation, and engineering of a linear polyketide tautomycetin, a novel immunosuppressant in *Streptomyces* sp. CK4412. *J. Ind. Microbiol. Biotechnol.* (e-pub ahead of print 12 October 2016; doi:10.1007/s10295-016-1847-2).
56 Jeong, H. *et al.* Genome sequence of the vancomycin-producing *Amycolatopsis orientalis* subsp. *orientalis* strain KCTC 9412T. *Genome Announc.* **1**, e00408–e00413 (2013).
57 Song, J. Y. *et al.* Draft genome of *Streptomyces clavuligerus* NRRL 3585, a producer of diverse secondary metabolites. *J. Bacteriol.* **192**, 6317–6318 (2010).

Supplementary Information accompanies the paper on The Journal of Antibiotics website (http://www.nature.com/ja)