SIMB
Society for Industrial Microbiology and Biotechnology

Habitat-specific type I polyketide synthases in soils and street sediments

Patrick Hill · Jörn Piel · Stéphane Aris-Brosou · Václav Krištůfek · Christopher N. Boddy · Lubbert Dijkhuizen

Received: 10 June 2013 / Accepted: 9 November 2013 / Published online: 16 November 2013

Abstract Actinomycetes produce many pharmaceutically useful compounds through type I polyketide biosynthetic pathways. Soil has traditionally been an important source for these actinomycete-derived pharmaceuticals. As the rate of antibiotic discovery has decreased and the incidence of antibiotic resistance has increased, researchers have looked for alternatives to soil for bioprospecting. Street sediment, where actinomycetes make up a larger fraction of the bacterial population than in soil, is one such alternative environment. To determine if these differences in actinomycetal community structure are reflected in type I polyketide synthases (PKSI) distribution, environmental DNA from soils and street sediments was characterized by sequencing amplicons of PKSI-specific PCR primers. Amplicons covered two domains: the last 80 amino acids of the ketosynthase (KS) domain and the first 240 amino acids of the acyltransferase (AT) domain. One hundred and ninety clones from ten contrasting soils from six regions and nine street sediments from six cities were sequenced. Twenty-five clones from two earthworm-affected samples were also sequenced. UniFrac lineage-specific analysis identified two clades that clustered with actinomycetal GenBank matches that were street sediment-specific, one similar to the PKSI segment of the mycobactin siderophore involved in mycobacterial virulence. A clade of soil-specific sequences clustered with GenBank matches from the ambruticin and jerangolid pathways of *Sorangium cellulosum*. All three of these clades were found in sites >700 km apart. Street sediments are enriched in actinomycetal PKSIs. Non-actinomycetal PKSI pathways may be more chemically diverse than actinomycetal PKSIs. Common soil and street sediment PKIs are globally distributed.

Keywords Natural product discovery · Polyketides · Urban microbiology · Bioprospecting · Mycobactin

Introduction

Type I polyketide synthases (PKSIs) produce many and varied bioactive microbial secondary metabolites [11]. Many of these metabolites have been developed into pharmaceuticals including the macrolide antibiotics and the newly approved anti-cancer drug ixabepilone [22]. Most of these polyketide antibiotics were discovered through culturing actinomycetes, a subclass of the *Actinobacteria*,

Electronic supplementary material The online version of this article (doi:10.1007/s10295-013-1362-7) contains supplementary material, which is available to authorized users.

P. Hill (✉) · S. Aris-Brosou · C. N. Boddy
Department of Biology, University of Ottawa, Ottawa,
ON K1N 6N5, Canada
e-mail: phill@uottawa.ca

J. Piel
Institute of Microbiology, Eidgenössische Technische
Hochschule (ETH) Zurich, Wolfgang-Pauli-Strasse 10,
8093 Zurich, Switzerland

V. Krištůfek
Biology Centre AS CR, v. v. i.-Institute of Soil Biology,
Na Sádkách 7, 370 05 České Budějovice, Czech Republic

C. N. Boddy
Department of Chemistry, University of Ottawa, Ottawa,
ON K1N 6N5, Canada

L. Dijkhuizen
Department of Microbial Physiology, Groningen Biomolecular
Sciences and Biotechnology Institute (GBB), University
of Groningen, Nijenborgh7, 9747 AG Groningen,
The Netherlands

from soils. This approach worked well from the 1950s until the late 1980s, however, since then, antibiotic discovery has slowed, with serious implications for drug development [21, 38].

Several authors have suggested soil of the hyper-arid Atacama Desert as a new source of actinomycetal secondary metabolites. The actinomycetal fraction of 16S libraries from soils is normally between 0 and 19 % [18]; in soils of the Atacama Desert, actinomycetes can make up over 90 % of the bacterial community [6] and these actinomycetes have been the source of novel secondary metabolites such as the chaxamycins [26, 31]. Dust found on the street surfaces or that collects in between cobblestones may be another new source of actinomycetal PKSIs. Street sediments support a broad range of bacterial communities [19] and are enriched in actinomycetes compared to soils, with some 16S clone libraries being more than 90 % actinomycetal [16].

Even if the bacterial communities of street sediments are largely actinomycetal, this may not mean that street sediments are a good alternative to soil for PKSI bioprospecting. The PKSI genes in the actinomycetal-rich street sediment communities may also be in soil. Actinomycetal distribution may not be the best measure of PKSI distribution. Sequencing of amplicons from the conserved beta-ketosynthase (KS) domain of PKSI genes of environmental DNA from soil has found novel PKSI KS diversity ascribed to the *Proteobacteria*, Firmicutes, and Chloroflexi [7, 24, 27].

We compared PKSI distribution in soils and street sediments by lightly sampling (215 sequences) a broad range of samples ($n = 21$) to determine if there are PKIs genes that are specific to soil and street sediment. Previous work has characterized PKSIs in the environment using the KS domain, which clusters by phylogeny rather than function [12, 17]. We used a primer pair, which amplifies two PKSI domains, the last 20 % of the KS domain and the first 80 % of the acyltransferase (AT) domain, which clusters by substrate specificity [3]. These primers have previously been used to characterize soil environmental DNA [9].

Methods

Sampling and DNA extraction

Ten soil samples were taken from forest, grassland, and cultivated soil in Arctic tundra to Tropical savannah climates. Nine street samples were from six cities in Canada ($n = 1$), Europe ($n = 4$), and Pakistan ($n = 1$). Street and soil samples are named using the country or city sampled, land use, and site. Thus, Cz pasture-Palava was taken from a grassland soil in the Palava protected area of the Czech Republic; Brussels-cobblestones-Boucher is material from between cobblestones in Boucher Street, Brussels.

To eliminate the possibility of a sampling bias towards sequences present in all animal affected environments but absent from soils, we also sampled two earthworm-affected sites: the gut contents of *Martiodrilus heterostichon* from Rozo Colombia (Rozoworm gut-Martiodrilus) and vermicomposted cattle manure (Cz-Vermicompost). Sampling sites are described in Tables 1 and 2. Sampling method and DNA extraction followed Hill et al. [16].

PCR amplification

The degenerate PKSI-specific primers and reaction conditions of Ayuso–Sacido and Genilloud [3] were used (K1F 5′-TSAAGTCSAACATCGGBCA-3′ M6R 5′-CGCAGGTTSCSGTACCAGTA-3′) with an annealing temperature of 58 °C.

PCR gave two amplification products, a 1,100–1,390-bp band, which included sections of the AT and KS domains, and a 650–700-bp non-PKSI band (data not shown). PCR products were run on a 1 % agarose gel and the larger band was excised and extracted using the QIAquick Gel Extraction Kit (Hilden, Germany).

Cloning and sequencing

Gel-purified PCR products were cloned directly into the Promega pGEM-T easy cloning system. Ligation reactions were plated on X-gal and white colonies picked and used in colony PCR using the T7/SP-6 vector primers. Colony PCR products were sent to GATC Biotech (Konstanz, Germany) or Beckman Coulter Genomics (Danvers, MA, USA) for double-ended sequencing. The entire PCR product of 215 clones from 21 samples were sequenced (see Tables S1 and S2 for amplicon details). Nucleotide sequences were deposited in GenBank with the following accession numbers: Mbt sequence from street sediments KF764484-KF764498, other street sediment sequences KF781465-KF781488, KF826392-KF826435, soil sequences KF781438-KF781464, KF826317-KF826391, earth worm affected sequences KF826436-KF826458, AT domain only sequences KF826459-KF826465.

Sequence analysis

For six of the 215 sequences (Rozoworm gut-MartiodrillusP5, Cz cultivated-PlanaP7, České Budějovice cobblestones-Ceska P6, P7, P8, and Paris cobblestones-Rive GaucheP1) poor chromatogram quality at the start of the sequence meant that the KS domain could not be analyzed.

Sequences were aligned using Muscle [8] in MEGA, version 5.00 [37] with default parameters. AT domain

Table 1 Soil samples

Area	Land use	Parent material	Identifier	pH	Organic matter (%)	Clay (%)	Season	No. of seq
Ottawa, Canada	Pine plantation	Glacial moraine	Cdnforest-Kemptville*	5.8	2.8	9.9	Summer	3
Chelton Beach Park, P.E.I. Canada	Scattered grassland	Beach sand	Cdnbeach-PEI	4.7	3.7	4.5	Summer	8
Resolute Bay, Nunavut, Canada	Tundra	Dolomite	Cdntundra-Resolute	8.3	4.3	7.3	Summer	18
Česke, Budéjovice, Czech Republic	Maize	Vltava floodplain	Czcultivated-Plana*	6.1	2.9	15.0	Summer	11
	Church garden	Unknown	CeskeBudejovice-Paradise*	7.5	2.8	3.0	Winter	8
	Native grassland	Limestone	Czpasture-Palava*	7.2	5.0	20.0	Summer	9
Cali, Colombia	Native grassland	Inactive alluvial fan	Colpasture-Pance*	4.3	1.1	18.3	Dry season	10
	Primary forest	Cauca River floodplain	Colforest-Hatico*	7.7	7.0	25.0	Dry season	17
	Bamboo forest		Colforest-CIAT*	6.3	5.7	33.5	Wet season	7
Budapest, Hungary	Scrub forest, Citadel Park	Limestone	Hungforest-Citadel*	7.3	18.6	5.0	Winter	12

* Samples previously used in Hill et al. [16]

Table 2 Street sediment and earthworm-affected samples

Area	Sediment	Identifier	pH	Organic matter (%)	Clay (%)	Season	No. of seq
Brussels, Belgium	Cobblestones Rue, Boucher	Brussels-cobblestones-Boucher*	7.0	2.9	4.0	Winter	9
Byward market/Rideau centre, Ottawa, Canada	Street bricks, William Street by Dubliner pub.	Ottawa-cobblestones-Dubliner	7.3	18.5	0.0	Winter	16
	Pedestrian crosswalk, Rideau Centre	Ottawa Street dust-Rideau	7.7	7.5	0.1	Winter	7
Central Česke Budéjovice, Czech Republic	Cobblestones, Česka Street	Česke Budéjovice cobblestones-Česka *	7.5	3.0	3.0	Spring	10
	Street dust near Koh-i-noor pencil factory	Česke Budéjovicestreetdust-Koh-i-Noor*	7.3	3.3	2.0	Spring	4
Budapest, Hungary	Pavement crack, Terez Korut, near West Railway station	Budapest pavement-Terez Korut	6.8	13.8	0.0	Summer	13
Left Bank, Paris, France	Cobblestones, Rue du Sabot, Rive Gauche nightclub	Paris cobblestones-Rive Gauche*	7.3	9.1	4.0	Summer	10
	Cobblestones, Café Procope	Paris cobblestones-café*	7.0	10.0	6.0	Summer	6
Faisalabad, Pakistan	Street dust, clock tower, central Faisalabad	Faisalabad Street dust-Clock*	6.8	12.0	2.0	Dry season	12
Cattle manure, vermicomposting system, Czech Republic	Vermicomposted (*Eisenia andrei*) manure	Czech-Vermicompost*	7.2	32.1	n/d	n/a	15
Gut contents of Colombian earthworm from microcosm	*Martiodrilus heterostichon* gut contents	Rozoworm gut-Martiodrilus*	Sample too small for analysis				10

n/d not determined as too much organic carbon in sample to be practicable

n/a not applicable as sample indoors

* Sample previously used in Hill et al. [16]

alignments were curated with GBLOCKS [4] with the minimum length of a block reduced from 10 to 6 and with gap positions allowed. Model selection was performed with ProtTest based on the Akaike Information Criterion [1], which identified WAG as the best substitution model. Phylogenetic analysis was performed with PhyML with the subtree pruning and regrafting (SPR) tree-searching algorithm [13]. Branch support was estimated using approximate likelihood ratio test (aLRT; [2]).

The AT and KS amino acid trees were analyzed with UniFrac with the significance and lineage-specific analyses (www.bmf2.colorado.edu/UniFrac, [23]). Soil, street sediment, and earthworm-affected environments were defined as three separate environments. UniFrac significance tests were performed within each environment with 100 permutations.

UniFrac lineage statistically significant soils and street sediment clades included sequences found between 700 and 10,000 km apart. Five other soil and street sediment clades, which were not UniFrac lineage significant, also included sequences from distant sites. To determine if these clades were narrow enough to represent a single PKSI pathway, their maximum pairwise and overall mean distance was determined with MEGA, version 5.00 [37] using the whole uncurated amplified region of the AT and KS domains. These values were compared with similar values for the methylmalonyl AT domains and their associated upstream KS domains picked from ten known PKSI pathways for the same regions of the KS and AT domains.

Individual AT and KS trees were used to define highly supported clades (A5, A6, A8-1, A8-3, and Mbt, Figs. 1 and 2). AT/KS cophylogenies were plotted with the APE library in R [28]. Tree mismatch was estimated by the Robinson and Foulds [34] distance (RFD) and its significance was computed from the distribution of all-against-all RFDs of bootstrapped trees.

Full-length AT and KS domain amino acid sequences were compared with GenBank using protein BLAST. AT domains were compared with all acyl transferase (Pfam family PF00698) domains downloaded from the Protein FAMily database on April 18, 2012 (pfam.sanger.ac.uk, [30]) containing 10,266 sequences from 2,781 species of Bacteria and 366 Eukaryotes.

Results

Sequence length and GC ratios

A total of 215 amplicons were sequenced from 21 different environmental DNA samples. GC content of environmental sequences ranged from 61 to 75 %. These values were in the range of proteobacterial (58–76 %) and actinomycetal (66–77 %) matches but exclude Firmicutes (35 %) and cyanobacteria (43–51 %) as likely sources of these sequences (Table S1).

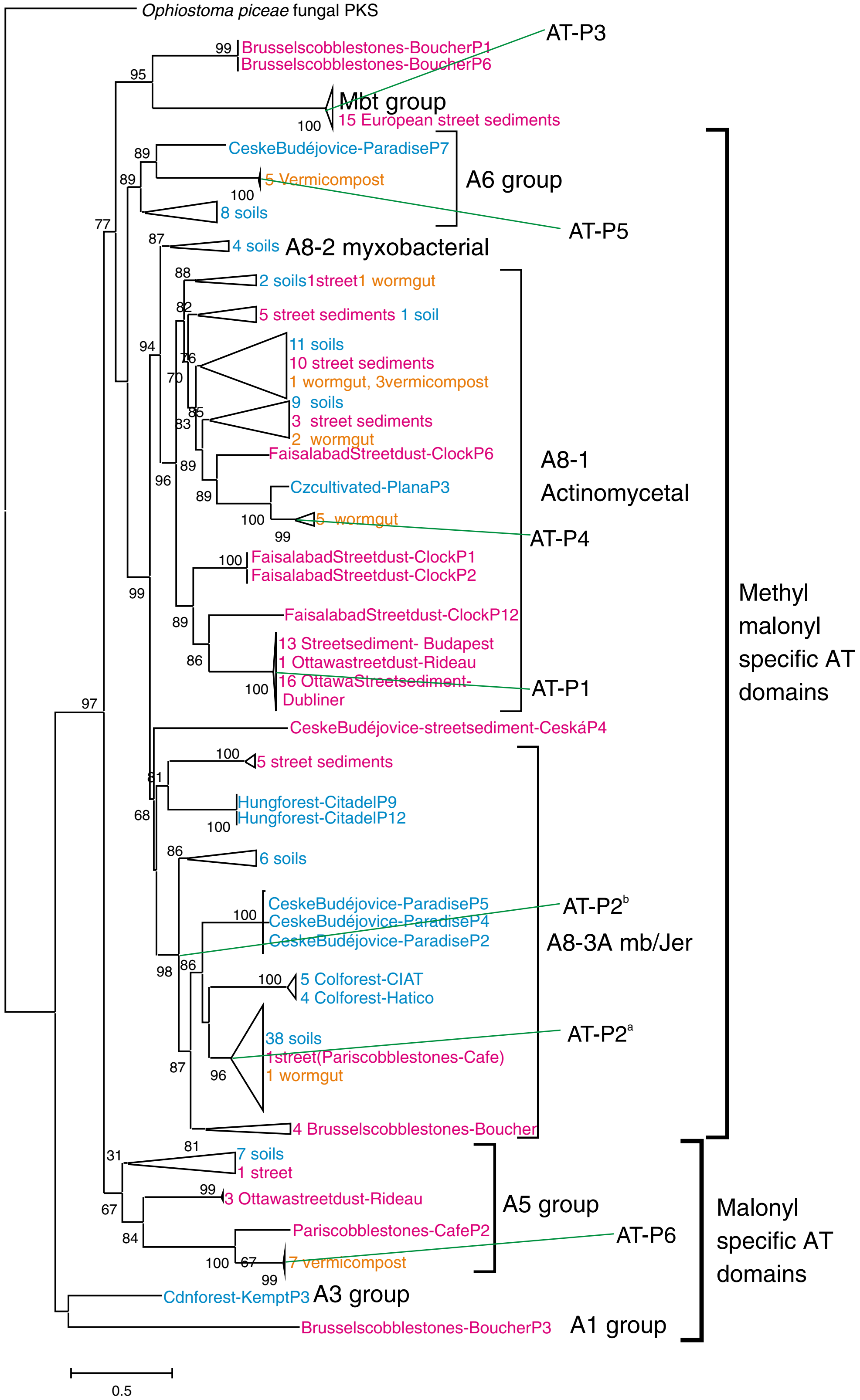
Fig. 1 Rooted maximum likelihood phylogenetic amino acid tree of AT domains. WAG substitution model used. Branch support values at nodes are approximate likelihood-ratio test for branches (aLRT) values. The *scale bar* represents 0.5 amino acid substitutions per position. Mbt refers to the mycobactin siderophore. Nodes identified as significant by UniFrac lineage-specific analysis are P1, P2, P2, P5, P6, and P7. *p* values and actual/expected values are: AT-P1 $p = 6.39 \times 10^{-11}$ Soils-0/14.37 Street sediments-30/12.14 Earthworm sediments-0/3.49, AT-P2^{A} $p = 8.83 \times 10^{-9}$ Soils-38/19.16 Street sediments-1/13.19 Earthworm sediments-1/4.65, AT-P2^{B} $p = 2.22 \times 10^{-3}$ Soils-18/8.17 Street sediments-4/10.81 Earthworm sediments-0/3.02 (note this analysis was run with descendants from P2^{A} removed), AT-P3 $p = 2.96 \times 10^{-4}$ Soils-0/7.19 Street sediments-15/6.07 Earthworm sediments-0/1.74, AT-P4 $p = 4.92 \times 10^{-3}$ Soils-0/2.40 Street sediments-0/2.02 Earthworm sediments-5/0.58, AT-P5 $p = 6.23 \times 10^{-3}$ Soils-0/2.40 Street sediments-0/2.02 Earthworm sediments-5/0.58, AT-P6 $p = 4.36 \times 10^{-5}$ Soils-0/3.35 Street sediments-0/2.83 Earthworm sediments-7/0.81

Amplicon sequence lengths ranged from 1,048 to 1,392 bp. PKSI primers [3] amplified the C-terminal 80 amino acids of the 400–460 amino acid long KS domain, an interspacer region, and the first 240 amino acids of the 280–310 amino acid-long AT domain. The three shortest sequences Brussels-cobblestones-Boucher P2, 4, and 9 had truncated AT domains of 473 to 574 bp (versus >700 bp of most sequences). KS (Fig. 2) and AT (Fig. 1) domains were analyzed separately as they have been shown to cluster differently. Interspacer regions were too variable in length and composition for phylogenetic analysis (data not shown).

AT and KS tree topology

Clusters in the AT and KS trees are classified using Jenke–Kodama et al.'s [17] system. Ginholac et al. [12] found four AT domain clusters: a malonyl-CoA, a methylmalonyl-CoA selective cluster, and two unresolved clusters (set A and B). Jenke–Kodama et al. [17] further refined this classification into eight AT groups: A1-A8. Groups A1-A5 are malonyl-CoA-specific, groups A6 and A8 are methylmalonyl-CoA specific, and the A7 group contains sequences for mycobacterial cell wall lipids. Sets A and B are groups A5 and A6.

Most (173) of the 215 sequenced amplicons had methylmalonyl-CoA-specific AT domains from the A8 (159) and A6 (14) groups (Fig. 1). A8 AT domains consisted of three subgroups: the A8-1 (86), A8-2 (6) clusters clustered with actinobacterial and myxobacterial GenBank sequences, respectively (Figs. S1 and S3) and the A8-3 (69) subgroup, which clustered with matches from the ambruticin (amb) and jerangolid (jer) pathways from *Sorangium cellulosum* [20] (Fig. S2). Twenty-one malonyl-CoA-specific AT domains were from the A1 (1), A3 (1), and A5 (19, Fig. S4)



groups. Fifteen sequences clustered with PKSI in the mixed NRPS/PKS pathway of the mycobacterial siderophore mycobactin (Mbt, Fig. S5) [5].

Jenke–Kodama [17] classified KS domains into the K1 to K7 groups. K1 is a mixed non-ribosomal peptide synthase (NRPS)/polyketide synthase (PKS) group and K2

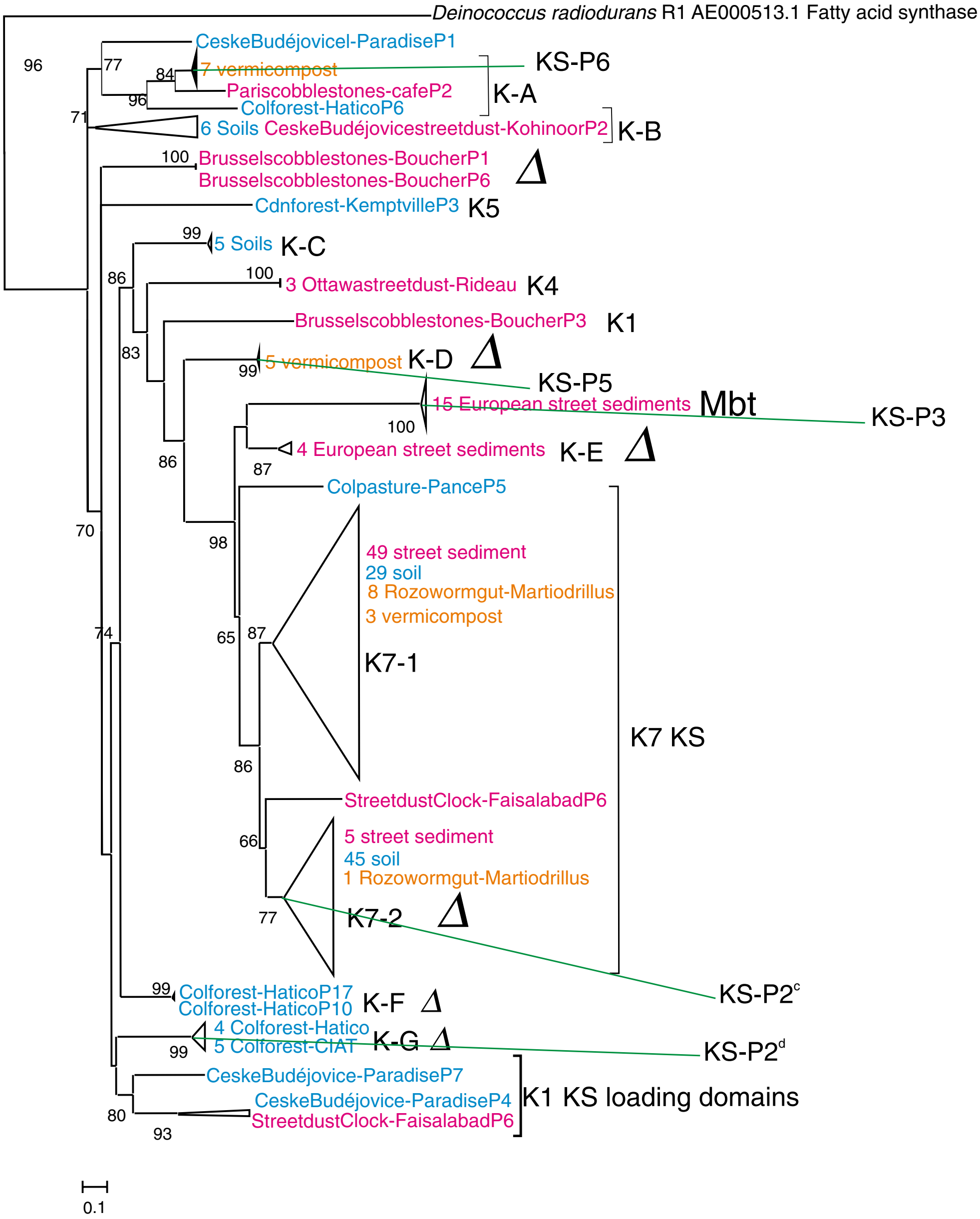

Deinococcus radiodurans R1 AE000513.1 Fatty acid synthase
96
77
84
96
71
CeskeBudéjovicel-ParadiseP1
7 vermicompost
Pariscobblestones-cafeP2
Colforest-HaticoP6
6 Soils CeskeBudéjovicestreetdust-KohinoorP2
KS-P6
K-A
K-B
100
Brusselscobblestones-BoucherP1
Brusselscobblestones-BoucherP6
Δ
Cdnforest-KemptvilleP3
K5
99
5 Soils
K-C
86
100
3 Ottawastreetdust-Rideau
K4
83
Brusselscobblestones-BoucherP3
K1
99
5 vermicompost
K-D
Δ
KS-P5
15 European street sediments
Mbt
100
KS-P3
4 European street sediments
K-E
Δ
86
87
Colpasture-PanceP5
70
98
49 street sediment
29 soil
8 Rozowormgut-Martiodrillus
3 vermicompost
K7-1
74
65
87
K7 KS
86
StreetdustClock-FaisalabadP6
66
5 street sediment
45 soil
1 Rozowormgut-Martiodrillus
77
K7-2
Δ
KS-P2c
99
Colforest-HaticoP17
Colforest-HaticoP10
K-F
Δ
4 Colforest-Hatico
5 Colforest-CIAT
K-G
Δ
99
KS-P2d
CeskeBudéjovice-ParadiseP7
CeskeBudéjovice-ParadiseP4
StreetdustClock-FaisalabadP6
K1 KS loading domains
80
93
0.1

Fig. 2 Rooted maximum likelihood phylogenetic amino acid tree of KS domains. WAG substitution model used. Branch support values at nodes are aLRT values. The *scale bar* represents 0.1 amino acid substitutions per position. Clusters are labeled with the Jenke-Kodama classification of the KS domain except in the case of Mbt which refers to the mycobactin siderophore. When KS domains do not match the Jenke-Kodama classification and are not Mbt they are labeled with *letters*. Clusters with no GenBank members annotated with *unfilled triangle*. Nodes identified as significant by UniFrac lineage-specific analysis are P1, P2c, P2d, P, P6, and P7. *p* values and actual/expected values are: KS-P1 Found within A8-1/A8-2 AT not shown. $p = 1.47 \times 10^{-11}$ Soils-0/14.64 Street sediments-30/11.91 Earthworm sediments-0/3.44, KS-P2^{c} $p = 4.70 \times 10^{-8}$ Soils-45/24.89 Street sediments-5/20.25 Earthworm sediments-1/5.86, KS-P2^{d} $p = 8.41 \times 10^{-2}$ Soils-9/4.39 Street sediments-0/3.57 Earthworm sediments-0/1.03, KS-P3 $p = 1.00 \times 10^{-4}$ Soils-0/7.32 Street sediments-15/5.96 Earthworm sediments-0/1.72, KS-P5 $p = 1.25 \times 10^{-3}$ Soils-0/2.44 Street sediments-0/1.98 Earthworm sediments-5/0.57, KS-P6 $p = 2.76 \times 10^{-5}$ Soils-0/3.42 Street sediments-0/2.78 Earthworm sediments-7/0.80

contains loading KS domains. The remaining groups did not cluster by KS function but by phylogeny, the K3 and K5 groups being cyanobacterial, the K4 and K6 groups mixed myxobacterial and actinomycetal, and the K7 group actinomycetal.

The actinomycetal Mbt and A8-1 subgroups both corresponded to a single KS domain clade, an Mbt cluster and the K7-1 KS cluster. The A8-3, A5, and A6 AT domains each corresponded to several different KS domain clusters (Supporting data S1–S5). These clusters were often not found in Jenke–Kodama et al.'s classification and in several cases did not include any GenBank sequences. These new KS clusters (K-A to K-G) are shown in Fig. 2. Most (48) of the environmental sequences from the A8-3 group clustered in a K7-2 subgroup, which did not contain any sequences from GenBank, nine were also found in the K-G subgroup. Lastly, neither the AT nor KS domains from Brusselscobblestones-Boucher P1 and P6 clustered with GenBank sequences.

Soil and street sediments are distinct PKSI habitats

UniFrac analyses showed that the soil AT and KS trees were significantly different from non-soil trees (AT and KS; $p < 0.01$) and that street sediment trees were significantly different from non-street sediment trees (AT and KS; $p < 0.01$). The earthworm-affected sediment trees, however, could not be differentiated from non-earthworm affect sediment trees (AT and KS; p values = 0.20 and 0.24, respectively).

Soil and street sediment-specific clades

UniFrac lineage-specific analysis identified two actinomycetal nodes that were specific to street sediments. The first, within the A8-1 subgroup, contained 30 sequences from three street sediment samples in Ottawa and Budapest (Fig. 1. AT-P1; $p = 6.39 \times 10^{-11}$). The KS sequences corresponding to these AT sequences were found in the K7-1 subgroup and were also statistically significant (Fig. 2, KS-P1; $p = 1.47 \times 10^{-11}$). The closest matches to these sequences were all actinomycetal (Figs. S1 and S2).

A second clade contained 15 sequences from European cities and was significant for both AT (Fig. 1, AT-P3; $p = 2.96 \times 10^{-4}$) and KS domains (Fig. 2, KS-P3; $p = 1.00 \times 10^{-4}$). These sequences clustered with proteins from the mixed NRPS/PKS pathway that produces a siderophore, Mbt, needed for virulence in a variety of mycobacterial and *Nocardia* strains [5]. Unlike other PKSI KS and AT sequences, the Mbt KS and AT domains were separated by a stop codon.

AT domains from the A8-3 group were soil-specific at several levels. One of the largest clusters of A8-3 sequences (40 of 69 sequences) is significant for soil (AT-P2^{a}; $p = 1.02 \times 10^{-8}$). Even with this cluster removed from the UniFrac environmental file, a node remains statistically significant for soil (AT-P2^{b}; $p = 1.88 \times 10^{-3}$).

KS domains associated with A8-3 AT domains were also strongly soil-specific. Fifty-one sequences with A8-3 sequences had KS domains in the K7-2 group. The K7-2 group was significantly selected for in soils (Fig. 2. KS-P2^{c}; $p = 4.70 \times 10^{-8}$). A second cluster of nine KS domains (KS-G) associated with A8-3 AT domains from two Colombian soils was selected for in soils with a less significant p value (Fig. 2. KS-P2^{d}; $p = 8.41 \times 10^{-2}$). Neither of these sequences clustered with any sequences from GenBank (Fig. S2).

The phylogeny of the A8-3 group is unclear. BLASTP matches for the 69 A8-3 sequences show that the closest matches to their AT domains are myxobacterial, either from the first AT domains in the amb and jer biosynthetic pathways from two strains of *Sorangium cellulosum* [20] or from a single *Myxococcus xanthus* sequence. However, a single unpublished sequence from an actinomycete isolated from mangrove soil (AEE69401) also clusters in the A8-3 group (supplemental data Fig. S2). KS domains from all of these matches clustered in the K7-1 group, while most KS domains from environmental A8-3 sequences were from the K7-2 and K-G clusters (Fig. 2).

Both soil and street sediment-specific clades were found in samples from distant sites. The P1 and AT-P2^{a}/KS-P2^{c} clades that UniFrac lineage analysis identified as statistically significant for soils or street sediments included sequences from sites that were from different continents. The P3 clade included sequences from cities at least 700 km apart in Europe. Five other small clades of either soil or street sediment sequences (Clusters C1–C5, Table S4; Figs. S1, S2, S3), while not significant for UniFrac lineage-specific analysis, included sequences from either soil or street sediments sites that were from different continents (C3, C4, C5) or from European cities that were at least 700 km apart (C1, C2).

The sequences within each of these widely geographically distributed (cosmopolitan) clades are similar to known PKSI proteins. The overall pairwise divergence within these cosmopolitan clades is between 0.012 and 0.117 for AT sequences and 0.00 and 0.167 for KS sequences (Table S4). In contrast, for ten known PKSI pathways, the overall pairwise divergence is between 0.00 and 0.614 for AT sequences and 0.064 and 0.528 for KS sequences (Table S4). Thus, these clades could represent either dissimilar pathways that share similar modules or similar pathways that produce identical or related polyketides products. In all of these cases, at least a single step in the PKSI assembly process is widely distributed over the landscape.

Earthworm-specific clades

UniFrac lineage-specific analysis identified three clades as earthworm-specific. As each clade was only found in a single sample, it is uncertain whether they were specific to earthworm species (*Martiodrilus heterostichon*, *Eisenia andrei*, environment (Worm Gut, Vermicompost) or some other factor.

Five sequences from Rozoworm gut-Martiodrilus formed a cluster within the actinomycetal A8-1 subgroup that is

Table 3 Reeve's signature model for AT domain motifs

E. coli FabD residue position	63[a]	92	201
Malonyl specific	ZTX$(T/A)(Q/E0)	GHS(L/V/I)GE	H(A/G)FH or (H/T/V/Y)AFH
Methylmalonyl specific	R(V/A/I)(D/E)VVQ or (R/Q/S/E/D)V(D/E)VVQ	GHS(Q/M)GE	(Y/V/W)ASH

X any amino acid, *Z* any hydrophilic amino acid, *$* any aromatic amino acid

[a] Consensus patterns given with all possible variations in *parentheses*. After Reeves et al. [33], Ginolhac et al. [12], Smith and Tsai [36]

UniFrac lineage significant (Fig. 1, AT-P4; $p = 4.92 \times 10^{-3}$). As the KS domain from one of these sequences was not included in analysis (Rozoworm gut-MartiodrillusP5), this node was not significant for the KS tree.

Two clusters were specific to Cz-vermicompost. The first was in the A6 group (Fig. 1, AT-P5; $p = 6.23 \times 10^{-3}$). The corresponding KS sequences were also selected for in vermicompost (Fig. 2, KS-P5 $p = 1.25 \times 10^{-3}$). A second cluster in the A5 group was selected for in the AT tree (Fig. 1, AT-P6*;* $p = 4.36 \times 10^{-5}$) and KS tree (Fig. 2, KS-P6; $p = 2.76 \times 10^{-5}$).

Sequence comparisons

Amino acid sequences were compared to the GenBank database by BlastP. Two amino acid sequences, Faisalabad street dust-Clock P11 and P12, had 89 and 98 % identity to GenBank matches and the percentage identity of the 15 Mbt group sequences to GenBank was between 75 and 80 %. Most (180/209, Table S2) full-length amino acid sequences had 65 % or less identity to GenBank. Sequences with A5 AT domains were the most dissimilar (42–53 %).

AT domains catalyze a highly selective self-acylation by malonyl-CoA, methylmalonyl-CoA, or other malonyl-CoA derivatives followed by transfer to the acyl carrier protein (ACP). The specificity of the AT domain is controlled by a set of motifs identified by Haydock et al. [15], modified by Reeves et al. [33] and Yadav et al. [39] and most recently summarized by Smith and Tsai [36]. These motifs are at the 63, 92 and 201 amino acid positions of the AT domain (number based on the AT domain of FabD in *Escherichia coli* Table 3).

Individual motifs of all 215 AT sequences and 99 comparable GenBank sequences were compared to the motif code in Table 3 (see Table S1 for full results). Few of the GenBank hits (30/82) and environmental clones (62/215) matched known, characterized motifs at all three positions (63, 92, and 201). Most AT domains with motifs that corresponded to the Reeves model at all three positions were A8-AT domains (21 GenBank matches and 57 environmental matches). Sequences that did not match known motifs, most often varied at position 63. Sequences from the Mbt, A1, A3, A5, and A6 groups generally did not match the known model for at least one of three motifs; this was also the case for their GenBank matches.

Motifs from environmental clones were compared with the Pfam database. Forty-three environmental sequences contained Reeve's motifs not found in the database; 34 of these were from street sediments or earthworm-affected samples.

Discussion

Our study compared 215 amplicons encoding part of the KS and AT domains of type I PKSI from 21 samples. Previous studies of PKSIs in the environment have amplified and sequenced the KS domain. Having both the AT and KS domains allowed cophylogenetic analysis, which showed discrepancies of clustering between domains. The length of the amplicons meant that next-generation sequencing could not be used, limiting the sequence coverage of any one sample. However, UniFrac lineage-specific analysis can identify habitat-specific sequences at a level of sequencing that does not fully characterize any one sample. Shallow sequencing across many samples can identify domains common to environments. Strategies that give the most information for the least sequencing will become important if sequencing costs stabilize [14].

Distinct PKSI synthases are specific to soil and street sediments

UniFrac analyses showed that soil and street sediment are different PKSI habitats. This is probably as street sediments are enriched in actinomycetes [16]. UniFrac lineage analysis identified two clades that clustered with actinobacterial sequences from street sediments. Thirty sequences with A8-1 AT and K7-1 KS domains were found in Ottawa and Budapest (AT-P1, KS-P1). A second cluster of 15 sequences related to the Mbt group of siderophores for mycobacterial virulence was found in European streets (AT-P3, KS-P3).

Actinomycetal distribution may not be a good guide to PKSI distribution. One hundred and eight sequences (50.2 %) with A5, A6, A8-2, and A8-3 AT domains do not cluster with actinomycetal GenBank matches. These sequences are probably not from actinomycetes but we cannot say which taxon they are from.

The largest of these clusters of uncertain origin is the A8-3 AT domains that is soil-specific (AT-P2) and make up 56 % (58/103) of the amplicons from soil. The A8-3 AT domains cluster with a *Myxococcus xanthus* sequence, domains from the start of the amb and jer pathways from two strains of *Sorangium cellulosum* (myxobacteria) and a single actinobacterial isolate from a Chinese mangrove soil. However, most KS domains from this group do not cluster with any GenBank matches.

Most GenBank sequences that cluster with the A6 and A8-2 sequences are myxobacterial. A5 sequences cluster with GenBank sequences from the Firmicutes, cyanobacteria, *Gamma proteobacteria*, myxobacteria, Planctomycetes, and soil metagenomic clones. Comparison of GC ratios shows that the A5 sequences are unlikely to be gammaproteobacterial, cyanobacterial, or firmicute (Tables S1, S3; Fig. S4). Even when PKSI domains consistently cluster with a single bacterial phylum with a similar GC content it is not certain that they are from that phylum. For example, Parsley et al. [29] found no consistent homology to any one bacterial phyla over the insert length of several soil metagenomic PKSI clones.

These putatively non-actinomycetal sequences may produce more chemically diverse compounds than better characterized actinomycetal PKSIs. Sequences with the actinomycetal A8-1 AT domain/K7-1 KS domain sequences were more likely to have known Reeve's motifs. Sequences from the A8-3, A5, and A6 AT domains contained wider ranges of KS domains, some of which did not cluster with any matches from GenBank. The A5 and A6 groups also contained many novel and non-standard Reeve's motifs (Table S1), and especially for the A5 group, lower percentage identities to GenBank (Table S2).

Previous studies using PKSI KS domain-specific primers on soil DNA ascribed most of the PKSI diversity to non-actinomycetes in a rhizosphere [40], a marine sediment [41] and a broad range of soils [7, 24, 27]. Two studies of soil metagenomic libraries have also found non-actinomycetal PKSI synthases in soil [12, 29]; in one case, these clones appeared to be acidobacterial.

In contrast, Reddy et al. [32], using pyrosequencing of amplicons from PKSI KS primers found that most (89 %) domains amplified from desert soils from Arizona, California, and Utah were actinomycetal and few (0.02 %) were proteobacterial. Desert soils may have bacterial communities that are more actinomycetal than other soils [6, 10, 25, 35], like street sediments and unlike most soils they contain few Acidobacteria [16, 25].

Many common environmental PKSIs are habitat-specific but widely distributed

Several soil or street sediment-specific clades included sequences from sites hundreds to thousands of kilometers from each other. The best example of this was a sub-clade of the A8-3 group (AT-P2a; Fig. 1) with sequences from the Canadian Arctic, the Czech Republic, and Colombia. In total, 96 of the 190 soil and street sediments sequences were in clades with sequences from two or more sites that were over 700 km from each other. Overall, the mean distance of these clades is low enough for each of them to be from a single PKSI pathway (Table S4).

This suggests that many common polyketides are specific to particular environments such as soil or street sediments but cosmopolitan (i.e., found in a particular environment wherever that environment may be). If this is true, efficient polyketide discovery will depend on identifying environments that have their own specific PKSIs rather than extensively sampling a single environment such as soil. Identifying these environments may not be straightforward; the two earthworm-affected samples contained different sequences while a single PKSI sequence predominated in street samples from Ottawa and Budapest but was not found in other street samples. Finding new polyketides will mean surveying anthropogenic habitats such as street sediment that are rarely studied as they neither correspond to biodiverse eukaryotic habitats nor provide ecosystem services.

Again, our results differ from those of Reddy et al. [32] who found that even at 85 % similarity, 90 % of PKSI sequences were only found in a single sample. Reddy et al. compared three soils far closer and more similar to each other than our sites were. This difference could be due to methods and/or the nature of desert soils.

Acknowledgments Dr. Meng Wang was kind enough to translate the paper "Screening identifying and functional analysis of polyketide synthase I domains from soil and seawater of Yangshan Harbour" into English for us. We thank Dr. A. Poulain for providing the sample from Resolute Bay, Canada, Dr. Rashid Nazir for providing the Faisalabad Clock tower sample, Dr. Declan Hill for providing the street sample from Budapest, Hungary, and Keavin M. Stanford-Finnerty for the beach sand sample from Prince Edward Island, Canada. Our laboratory work was helped by Jean Bjornson of the Geography Dept of the University of Ottawa who assisted with soil analysis and Daniël van Middelkoop and Sarina de Roos who carried out cloning and sequencing as part of a work experience program for the School for Biomedical Laboratory Techniques in Groningen, The Netherlands. Funding for this work was provided by the NATO Programme Security Through Science ESP.EAP.CLG 981785, the Ministry of Education, Youth and Sports of the Czech Republic (LC06066, 2B06154), the Research Plan of the Institute of Soil Biology (AV 0Z 6066 0521), the Natural Sciences and Engineering Council of Canada, the Ministry of Research and Innovation, and the University of Ottawa.

References

1. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinforma 21: 2104–2105
2. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552

3. Ayuso–Sacido A, Genilloud O (2005) New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microb Ecol 49:24–49
4. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552
5. Chavadi SS, Stirrett KL, Edupuganti UR, Vergnolle O, Sadhanandan G, Marchiano E, Martin C, Qiu WG, Soll CE, Quadri LE (2011) Mutational and phylogenetic analyses of the mycobacterial *mbt* gene cluster. J Bacteriol 193:5905–5913
6. Connon S A, Lester ED, Shafaat HS, Obenhuber DC, Ponce A (2007) Bacterial diversity in hyperarid Atacama Desert soils. J Geophys Res 112: G04S17, doi: 10.1029/2006JG000311
7. Dong XY, Wang LH, Sun MJ, Zong Y, Jiao YL, Jiao BH (2008) Screening, identifying and function analysis of polyketide synthase I domains from soil and seawater of Yangshan Harbor. Microbiology 35:1359–1366
8. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
9. Faizal I, Lestari R, Kurnia F, Latif A, Hadianto D, Kusumawati N, Rachmawati I, Marwoto B, Purbowasito W (2008) Polymorphism analysis of polyketide synthase gene from *Actinomycetes* genome DNA of Taman Nasional Gunung Halimun soil by using metagenome method. J Biotechnology Res Tropical Reg 1: biotechindonesia.org/journal/jbr/jbr-2008-00-01/jrb-1-08-2.pdf
10. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DA, Caporaso JG (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. PNAS 109:21390–21395
11. Fischbach MA, Walsh CT (2006) Assembly-line enzymology for polyketide and non-ribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106:3468–3496
12. Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K, Bertrand H, Vogel TM, Perrière G, Simonet P, Nalin R (2004) Phylogenetic analysis of polyketide I domains from soil metagenomic libraries allows selection of promising clones. Appl Environ Microbiol 70:5522–5527
13. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321
14. Hall N (2013) After the gold rush. Genome Biol 14:115–117
15. Haydock SF, Aparicio JF, Molnár I, Schwecke T, Khaw LE, König A, Marsden AFA, Galloway IS, Staunton J, Leadlay PF (1995) Divergent sequence motifs correlated with the substrate specificity of (methlyl) malonyl-CoA: acyl carrier protein during transcyclase domains in modular polyketide synthases. FEBS Lett 374:246–248
16. Hill P, Krištůfek V, Dijkhuizen L, Boddy C, Kroetsch D, van Elsas D (2011) Land use intensity controls actinobacterial community structure. Microb Ecol 61:286–302
17. Jenke-Kodama H, Sandmann A, Mű̈ller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22:2027–2039
18. Jensen PH (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microbiol 72:1719–1728
19. Johnson MJ, Lee KY, Scow KM (2003) DNA fingerprinting reveals links among agricultural crops, soil properties, and the composition of soil microbial communities. Geoderma 114:279–303
20. Julien B, Tian ZQ, Reid R, Reeves CD (2006) Analysis of the ambruticin and jerangolid gene clusters of sorangium cellulosum reveals unusual mechanisms of polyketide biosynthesis. Chem Biol 13:1277–1286
21. Lam KS (2007) New aspects of natural products in drug discovery. Trends Microbiol 15:279–289
22. Lee FYF, Borzilleri R, Fairchild CR, Kamath A, Smykla R, Kramer R, Vite G (2008) Preclinical discovery of ixabepilone, a highly active antineoplastic agent. Cancer Chemother Pharmacol 63:157–166
23. Lozupone C, Hamady M, Knight R (2006) UniFrac: an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7:371
24. Luo K, Du G-P, Zhao Z-X, Xie BY, Li D-J (2010) Phylogenetic analysis of type I polyketide synthase and non-ribosomal peptide synthase genes from Mila Mountain in Tibet plateau. J Hunan Agric Uni (Nat Sci) 36:506–511
25. Neilson JW, Quade J, Ortiz M, Nelson WM, Legatzki A, Tian F, LaComb M, Betancourt JL, Wing RA, Soderlund CA, Maier RM (2012) Life at the hyperarid margin: novel bacterial diversity in arid soils of the Atacama Desert, Chile. Extremophiles 16:553–566
26. Okoro CK, Brown R, Jones AL, Andrews BA, Asenjo JA, Goodfellow M, Bull AT (2009) Diversity of culturable actinomycetes in hyper-arid soils of the Atacama Desert, Chile. Antonie Leeuwenhoek 95:121–133
27. Pang MF, Tan G-YA, Abdullah N, Lee C-W, Ng C-C (2008) Phylogenetic analysis of type I and type II polyketide synthase from tropical forest soil. Biotechnology 7:660–668
28. Paradis A (2006) Analysis of phylogenetics and evolution. Springer, Berlin Heidelberg New York
29. Parsley LC, Linneman J, Goode AM, Becklund K, George I, Goodman RM, Lopanik NB, Liles MR (2011) Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. FEMS Microbiol Ecol 78:176–187
30. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301
31. Rateb ME, Houssen WE, Arnold M, Abdelrahman MH, Deng H, Harrison WTA, Okoro CK, Asenjo JA, Andrews BA, Ferguson G, Bull A, Goodfellow M, Ebel R, Jaspars M (2011) Chaxamycins A-D, bioactive ansamycins from a hyper-arid desert *Streptomyces* sp. J Nat Prod 74:1491–1499
32. Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z, Brady SF (2012) Natural product biosynthetic gene diversity in geographically distinct soil micro biomes. Appl Environ Microb 78:3744–3752
33. Reeves CD, Murli S, Ashley GW, Piagentini M, Hutchinson CR, McDaniel R (2001) Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. Biochemistry 40:15464–15470
34. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147
35. Saul-Tcherkas V, Steinberger Y (2011) Soil microbial diversity in the vicinity of a Negev Desert shrub–*Reaumuria negevensis*. Microb Ecol 61:64–81
36. Smith S, Tsai S-C (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. Nat Prod Rep 24:1041–1072
37. Tamura K, Peterson D, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
38. Watve MG, Tickoo R, Jog MM, Bhole BD (2001) How many antibiotics are produced by the genus *Streptomyces*? Arch Microbiol 176:386–390

39. Yadav G, Gokhale RS, Mohanty D (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J Mol Biol 328:335–363

40. Zhao B, Gao Z, Shao Y, Yan J, Hu Y, Yu J, Liu Q, Chen F (2012) Diversity analysis of type I ketosynthase in rhizosphere soil of cucumber. J Basic Microbiol 52:224–231

41. Zhao J, Yang N, Zeng R (2008) Phylogenetic analysis of type I polyketide synthase and non-ribosomal peptide synthetase genes in Antarctic sediment. Extremophiles 12:97–105