DOI 10.1099/mic.0.000386

Editor's Choice

Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters

Michelle A. Schorn,[1] Mohammad M. Alanjary,[2] Kristen Aguinaldo,[3] Anton Korobeynikov,[4,5] Sheila Podell,[1] Nastassia Patin,[1] Tommie Lincecum,[3] Paul R. Jensen,[1,6] Nadine Ziemert[2] and Bradley S. Moore[1,6,7]

Correspondence
Bradley S. Moore
bsmoore@ucsd.edu

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, USA

[2]German Centre for Infection Research (DZIF), Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), University of Tuebingen, Tuebingen, Germany

[3]Thermo Fisher Scientific, Carlsbad, CA, USA

[4]Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia

[5]Department of Statistical Modeling, St. Petersburg State University, St. Petersburg, Russia

[6]Center for Microbiome Innovation, University of California, San Diego, USA

[7]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, USA

Received 16 August 2016
Accepted 25 October 2016

Traditional natural product discovery methods have nearly exhausted the accessible diversity of microbial chemicals, making new sources and techniques paramount in the search for new molecules. Marine actinomycete bacteria have recently come into the spotlight as fruitful producers of structurally diverse secondary metabolites, and remain relatively untapped. In this study, we sequenced 21 marine-derived actinomycete strains, rarely studied for their secondary metabolite potential and under-represented in current genomic databases. We found that genome size and phylogeny were good predictors of biosynthetic gene cluster diversity, with larger genomes rivalling the well-known marine producers in the *Streptomyces* and *Salinispora* genera. Genomes in the *Micrococcineae* suborder, however, had consistently the lowest number of biosynthetic gene clusters. By networking individual gene clusters into gene cluster families, we were able to computationally estimate the degree of novelty each genus contributed to the current sequence databases. Based on the similarity measures between all actinobacteria in the Joint Genome Institute's Atlas of Biosynthetic gene Clusters database, rare marine genera show a high degree of novelty and diversity, with *Corynebacterium*, *Gordonia*, *Nocardiopsis*, *Saccharomonospora* and *Pseudonocardia* genera representing the highest gene cluster diversity. This research validates that rare marine actinomycetes are important candidates for

Abbreviations: ABC, Atlas of Biosynthetic gene Clusters; BGC, biosynthetic gene cluster; gDNA, genomic DNA; IMG, Integrated Microbial Genomes; JGI, Joint Genome Institute; NRPS, nonribosomal peptide synthetase; PGM, Personal Genome Machine; PKS, polyketide synthase; RMA, rare marine actinomycete.

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers are as follows: CNU-125-MKJY01000000, CUA-896-MKKH01000000, CNJ-954-MKKI01000000, CNJ-863-MKKG01000000, CNJ-787-SRX873596*, CNJ-770-MKJW01000000, CUA-901-MKKB01000000, CNB-394-GCA_000374985.1, CNS-044-SRX873600*, CNY-236-GCA_000482385.1, CNR-923-MKKC01000000, CNS-639-GCA_000381685.1, CNT-312-GCA_000515115.1, CNJ-824-MKKA01000000, CNS-004-MKJV01000000, CNS-139-MKJX01000000, CUA-806-MKKD01000000, CUA-673-MKKE01000000, CNQ-490-GCA_000527075.1, CNJ-927-MKKF01000000 and CUA-874-MKJZ01000000. *These accession numbers are for the NCBI Sequence Read Archive (SRA) database.

Six supplementary figures and five supplementary tables are available with the online Supplementary Material.

exploration, as they are relatively unstudied, and their relatives are historically rich in secondary metabolites.

INTRODUCTION

The Actinobacteria represent a diverse phylum of bacteria capable of immense secondary metabolic capacity (Monciardini *et al.*, 2014). They are renowned for producing biomedically important molecules such as antibiotics and anticancer compounds and include human and plant pathogens. Many of the better-studied genera in the terrestrial and clinical spheres, such as *Streptomyces* and *Mycobacterium*, have hundreds of genomes sequenced (Doroghazi & Metcalf, 2013; Nett *et al.*, 2009). Though these terrestrial strains have been exploited for centuries, the rate of discovery of new chemical entities from terrestrial microbes has slowed in recent years (Bérdy, 2012). The conventional path to natural product discovery relies heavily on the ability to coax cultured strains of bacteria to produce metabolites at detectable levels by varying laboratory conditions. As time goes on, the likelihood of re-discovering a known compound inevitably increases, further reducing the productivity of traditional natural product discovery methods. However, a new era of fast and cheap sequencing has transformed the natural products discovery field by revealing the undetected majority of gene clusters harboured in bacterial genomes (Bentley *et al.*, 2002; Ikeda *et al.*, 2003; Udwary *et al.*, 2007). A bacterium's genomic sequence contains the blueprint of potential molecules the strain is capable of producing. Mining bacterial genomes has shown that their potential for producing secondary metabolites is much higher than what is observed in the laboratory (Bachmann *et al.*, 2014). As bioinformatic tools for assessing bacterial secondary metabolite biosynthesis advance, the power in genome mining amplifies, allowing for de-replication of known products, compound class identification, structural predictions and, in some cases, target identification (Jensen *et al.*, 2014; Tang *et al.*, 2015). Likewise, advances in heterologous expression and regulation manipulation allow increased access to silent natural product biosynthetic pathways (Tang *et al.*, 2015; Yamanaka *et al.*, 2014).

Marine actinobacteria have recently come into the spotlight as fruitful producers of structurally diverse secondary metabolites and remain relatively untapped (Fenical & Jensen, 2006; Moore *et al.*, 2005; Subramani & Aalbersberg, 2013; Zotchev, 2012). Over 70 bioactive compounds have been isolated from marine actinobacteria, most from the genus *Streptomyces* (Manivasagan *et al.*, 2014). However, in the marine ecosystem, genera previously underexplored for natural product research are being reported on a regular basis as a source of new metabolites (Tiwari & Gupta, 2012). These so-called rare actinomycetes are defined as strains from actinomycete genera other than *Streptomyces* (Bérdy, 2005) or strains from genera that are isolated less frequently than *Streptomyces* species, although they may not be rare in abundance (Baltz, 2006; Lazzarini *et al.*, 2001). Our understanding of the genetic potential of rare marine actinomycetes (RMAs) is incomplete. Although there are many reported compounds from rare actinomycetes (Bérdy, 2005; Choi *et al.*, 2015), most rare genera have few genomes published and little is known about their genomic capacity to produce natural products, especially those from the ocean. Insight into RMA genomes may reveal an important untapped resource for unique biosynthetic gene clusters (BGCs) and inform future collection efforts in the search for new bioactive natural products.

Aside from *Streptomyces*, *Salinispora* is arguably the most prolific marine actinomycete genus, in terms of secondary metabolite production, discovered to date, producing a suite of potent bioactive chemicals (Jensen *et al.*, 2015; Manivasagan *et al.*, 2014). Sequencing of 75 genomes from three *Salinispora* species has revealed that up to 10 % of the genomes are dedicated to secondary metabolite biosynthesis and that many BGCs are uniquely present in only one or two strains (Ziemert *et al.*, 2014). These rare BGCs show that sequencing just one strain of a species does not capture the entire repertoire of potential natural products, thus validating further sequencing efforts to discover new BGCs. *Salinispora* is a prime example of how new or poorly explored taxa can lead to the discovery of new molecules through genome mining (Eustáquio *et al.*, 2011; Udwary *et al.*, 2007). Therefore, further investigation of RMA genomes could give access to a pool of untapped natural product biosynthetic potential. Currently, the Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) database houses over 4600 actinobacteria genomes. The top four represented orders – *Bifidobacteriales*, *Corynebacteriales*, *Micrococcales* and *Streptomycetales* – account for over 80 % of the genomes in the database.

Advancements in four key areas will aid in overcoming the current lull in novel natural product discovery: studying and learning to cultivate strains from poorly explored habitats, advances in genome sequencing and assembly for unfragmented genome blueprints, improved bioinformatics tools for predicting elusive biosynthetic origins and reliable heterologous expression of cryptic pathways. Development in these areas will unveil unexploited resources, such as RMAs for novel BGC sequence information, and could therefore lead to new chemical entities for use in medicine and biotechnology. While individual metabolites have been reported from RMAs at an increasing rate over the past decade, there has not yet been an analysis of the biosynthetic potential of a large set of RMAs. In this study, we add to the growing pool of sequenced RMA genomes and utilize gene cluster similarity networks to compare RMA gene clusters with gene clusters from the JGI IMG database to assess the likelihood of discovering new natural products. These gene cluster similarity networks allow for the rapid comparison of tens of thousands of gene clusters to quickly determine RMA gene cluster novelty and the diversity of classes they are distributed in.

METHODS

Genomic DNA isolation and genome sequencing. Strains (listed in Table S1, available in the online Supplementary Material) were obtained using previously detailed methods from various locations, grown in pure culture and stored at −80 °C (Gontang *et al.*, 2007; Jensen *et al.*, 2005). Genomic DNA (gDNA) extraction, sequencing and assembly of strains by JGI are previously detailed (Ziemert *et al.*, 2014). Strains sequenced in-house were grown in 5 ml of A1 medium (28 g Instant Ocean distributed by United Pet Group, 10 g starch, 4.0 g yeast extract, 2.0 g peptone and 1 litre deionized water) for 7 days, shaking at 220 r.p.m. at 28 °C on an Innova 2350 platform shaker (New Brunswick Scientific). Five millilitres of culture was then used for gDNA extraction using the Qiagen Genomic-tip 20/G kit (Qiagen). Strains that yielded little or no DNA using this kit and were extracted using a modified protocol developed for *Salinispora* spp. (Gontang *et al.*, 2007). gDNA was checked for quality by running on a 1 % agarose gel at 70–80 V for 1.5–2 h and stained using 1× GelRed (Biotium) in the gel. gDNA was quantified using the Qubit dsDNA HS Assay Kit with the Qubit Fluorometer (Thermo Fisher Scientific).

Ion Torrent 400 bp sequencing libraries were made using the Ion Xpress Plus gDNA Fragment Library kit according to the user guide (Thermo Fisher Scientific). The Covaris S2 (Covaris) was used to shear 1 µg of gDNA to about 500 bp. The samples were then processed according to the user guide for end repair and adapter ligation. The Pippen Prep (Sage Science) instrument was used, according to the provided protocol, to size select using a 2 % agarose gel cassette with DNA marker B, which allows for size selection between 100 and 600 bp, using a narrow size range targeting 475 bp. Libraries were not amplified and were analysed for quality and quantitated using the BioAnalyzer High Sensitivity DNA kit on the BioAnalyzer 2100 System (Agilent). The Ion Personal Genome Machine (PGM) 400 Template OT2 400 bp Kit (Thermo Fisher Scientific) was used for sample preparation with the Ion OneTouch 2 System according to the protocol with a modified 400 bp thermoprofile. The melting temperature was increased to 97 °C with elongated extension times. Sequencing was performed using an Ion Torrent PGM (Thermo Fisher Scientific) with an Ion PGM Hi-Q Sequencing Kit (Thermo Fisher Scientific), according to the standard protocol, on a 318 v2 sequencing chip (Thermo Fisher Scientific).

Genome assembly and annotation. SPAdes version 3.1.1 with Ion Torrent and single cell options was run with each fastq sequencing file (Bankevich *et al.*, 2012; Nurk *et al.*, 2013). K-mer sizes of 21, 33, 55 and 77 and single cell mode are recommended for high G+C genomes and were run for each genome assembly. Scaffolds smaller than 1 kb in length were discarded, unless 16S rRNA gene information was present. Each genome assembly was submitted to the JGI IMG Expert Review pipeline for genome annotation and is publicly available through http://genomeportal.jgi.doe.gov/. Additionally, this Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers can be found in Table S1.

Contigs containing potential contaminants were identified based on a combination of assembly coverage depth, nucleotide composition (percent G+C) and number of predicted amino acid sequences having closest matches in GenBank nr from actinobacterial versus non-actinobacterial sources, as determined using DarkHorse software, version 1.5 (Podell & Gaasterland, 2007). One genome, CUA-806 *Rhodococcus* sp., was contaminated by a low G+C *Staphylococcus* sp., sequenced at a much higher coverage. In this case, we were able to extract the low coverage, high G+C scaffolds such that the final assembly only contains scaffolds attributed to *Rhodococcus*.

Genome quality was assessed as described in *Standards in Genomic Science* in 2014 (Land *et al.*, 2014). Scaffold number; length of continuous, un-gapped nucleotides; length of 5S, 16S and 23S rRNAs; number and amino acid specificity of tRNAs and the presence or absence of genes encoding 102 housekeeping proteins found in nearly all bacteria were combined into a single normalized, composite score, enabling direct comparison to previously published quality data for 32 000 microbial genomes already in public databases.

All genome assemblies were subsequently analysed using antiSMASH v3.0 with and without ClusterFinder enabled (Weber *et al.*, 2015). The gene clusters identified without ClusterFinder were further curated using NaPDoS (Natural Product Domain Seeker) (Ziemert *et al.*, 2012) to determine which polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) partial clusters most likely belonged in the same gene cluster (Fig. S1). The final numbers of gene clusters were corrected to incorporate this additional information on inferred connections, consolidating gene clusters likely to have been split across multiple contigs during sequence assembly. This correction step was necessary to avoid potential overestimation of the number of PKS and NRPS clusters in each genome. The resulting table (Table S4) of genome versus gene cluster type and number was visualized using Circos (Krzywinski *et al.*, 2009).

16S phylogenetic tree. The 16S rRNA sequences were extracted from the whole-genome assemblies for all strains. Type strains within the same families and, if possible, from marine sources were selected from the list of prokaryotic names with standing in nomenclature (LPSN; bacterio.net) and 16S gene sequences for these strains were collected from NCBI. All sequences were aligned with MAFFT (Katoh *et al.*, 2005) using the accurate L-INS-I mode. Alignments were manually inspected and trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009) using the automated1 setting resulting in approximately 1500 bp of aligned sequence. The GTR+I+G evolutionary model was selected using MrModeltest (Nylander, 2004) and a maximum-likelihood tree was built using RAxML (Stamatakis, 2006) with 1000 bootstrap replicates. Visual aids were then added using the Interactive Tree of Life (iTOL) (Letunic & Bork, 2016).

Gene cluster family networking. BGC similarity was determined using a distance measure based on Pfam composition as detailed in Cimermancic *et al.* (2014). The method uses an optimized weighted distance that incorporates the Jaccard index and domain duplication similarity measures as employed in a previous method to determine protein similarity (Lin *et al.*, 2006). Pairwise distances between BGCs in RMA strains and Actinobacteria from JGI were generated using a custom python script and were then visualized in a network using Gephi (Bastian *et al.*, 2009). A similarity threshold was set at 0.6 to limit network complexity while retaining meaningful connections and to minimize the number of connected nodes with different cluster type annotations (Fig. S2). Clustering using the OpenORD and the YifanHu force-directed algorithm was performed for the final layout (Hu, 2006). Before final networking, a de-replication step was necessary to account for the number of re-sequencing projects of the same isolate present in the JGI dataset. Replicate gene clusters, ones with 0.99 similarity score and matching organism identification, were condensed into a single node before visualization. This was accomplished using an initial network of nodes over the 0.99 threshold. Attributes were inherited by the new nodes with the addition of a node size attribute, used to visualize the number of de-replicated gene clusters in the final network.

Annotations for secondary metabolite products from the JGI dataset were augmented by including homology results to the MIBiG database

(Medema *et al.*, 2015). A random sampling of clusters in the set were screened using MultiGeneBlast (Medema *et al.*, 2013) against MIBiG and paired with the highest scoring hit. All hits that did not have 80 % of the genes in the query cluster were filtered out. Nodes used as example compounds were also manually screened for accuracy and MIBiG BGC numbers are provided in Fig. 1 caption.

RESULTS AND DISCUSSION

Rare actinomycete sequencing and genome assembly

We chose 21 strains for genome sequencing and secondary metabolite analysis (Table S1). In their paper, Gontang *et al.* (2010) probed a variety of marine sediment actinomycetes for pathways indicative of secondary metabolite production. The majority of the microbes assayed contained either or both PKS and NRPS pathways, suggesting the presence of biosynthetic machinery for making yet to be identified molecules. We selected nine strains from the Gontang study, as well as five newly isolated RMAs, for whole-genome sequencing in-house, and another seven strains for sequencing at the JGI. We sequenced the 14 strains using the Ion Torrent PGM, with 400 bp sequencing libraries and 400 bp sample preparation and sequencing chemistry (Rothberg *et al.*, 2011; Yamanaka *et al.*, 2014). The limitations of using next-generation sequencing, specifically for the discovery of secondary metabolites, are many and varied (Gomez-Escribano *et al.*, 2016). Read length and high G+C content of the DNA are the two biggest hurdles to overcome. At the time of sequencing, 400 bp contiguous sequences were considered quite long. To address the high G+C nature of these actinomycete genomes, we created a modified thermoprofile which incorporated high denaturation temperatures and longer extension times to improve amplification of these long, high G+C sequencing libraries. Each genome was run on one 318 chip, giving 1–1.8 Gb of information per genome.

We used SPAdes (Bankevich *et al.*, 2012; Nurk *et al.*, 2013) for genome assembly for many reasons. First, it is one of the only non-commercial assemblers that can be used with Ion Torrent reads, and has an error correction module, Ion-Hammer, specific to Ion Torrent errors. Additionally, the single cell option, which was developed for improved assembly of multiple displacement amplified genomes from a single cell, aided with assembly of these high G+C genomes because of the non-uniform coverage profiles associated with these genomes. Preliminary assembly using commercial CLC Genomics Workbench (Qiagen) software resulted in very few complete secondary metabolite gene clusters; most clusters were truncated at the beginning or at the end of a contig. However, after using SPAdes, over 50 complete gene clusters, including notoriously difficult to assemble PKS and NRPS clusters, emerged. The single cell option in SPAdes helped us to solve the complications associated with high G+C sequences.

All genome assemblies were confirmed to be of sufficiently high quality for comparative analysis (Table S2) based on recommended guidelines set forth in 2014 in *Standards in Genomic Science* (Land *et al.*, 2014). This definitive monograph states that genome assemblies with quality scores above 0.8 can be safely used for comparative genomic analysis, but those with scores below 0.6 should not be used. Of the 21 RMA genomes, 19 in the current study had quality scores above 0.8. The two assemblies below 0.8 were *Actinomadura* sp. CNU-125 (0.76) and *Kytococcus* sp. CUA-901 (0.79), but both of these were still well above the 0.6 minimum threshold value.

Variety of secondary metabolite biosynthetic gene clusters

Many of the under-exploited genera from this study contain pathways of various classes that warrant further exploration (Table S3). The variety and number of pathways, as found by antiSMASH 3.0 without ClusterFinder, for a representative strain of each genus sequenced are displayed in the Circos diagram (Fig. 2) (Krzywinski *et al.*, 2009). Non-NRPS-PKS hybrid clusters were separated into their component parts so as to more easily visualize the variety of categories present, while the ‘Hybrid’ category retains only NRPS-PKS hybrid clusters (Table S4). As is expected, the number and variety of pathways generally increases as the size of the genome increases, with few exceptions. Also not surprising are the ubiquitous terpene pathways, present in every genome. Recent genome mining efforts revealed wide distribution of terpene synthases in bacterial genomes and led to the creation of a new hidden Markov model to identify bacterial terpene synthases (Cane & Ikeda, 2012; Yamada *et al.*, 2015).

The second most pervasive and most abundant class of BGCs are PKS pathways, present in all large genomes, and some small genomes as well. The next most represented class of BGCs is NRPS clusters. Identification of siderophores based on bioinformatics criteria alone can be difficult, and many siderophores are made by NRPS pathways that antiSMASH identifies as NRPS and not Siderophore. In fact, all genomes that did not have an explicitly identified siderophore cluster contained at least one NRPS pathway with >50 % gene homology, according to antiSMASH, to known siderophore pathways, such as coelichelin and fuscachelin. Interestingly, one of the larger genomes, *Pseudonocardia* strain CNS-004, contains a relatively low number of secondary metabolite pathways, with no PKS or Hybrid clusters and only two NRPS pathways (Table S2). In contrast, *Pseudonocardia* strain CNS-139 contains a large number of clusters, including Hybrid, PKS and many NRPS clusters. This difference in total pathway abundance between two members of the same genus is also seen in another *Pseudonocardineae* genus, *Saccharomonospora* (Fig. 3).

The actinomycete 16S phylogeny (Fig. 3) reveals some patterns in the number of pathways present in the RMA genomes. Some of the strains sequenced have numbers nd MIBiG BGC numbers are provand varieties of pathway

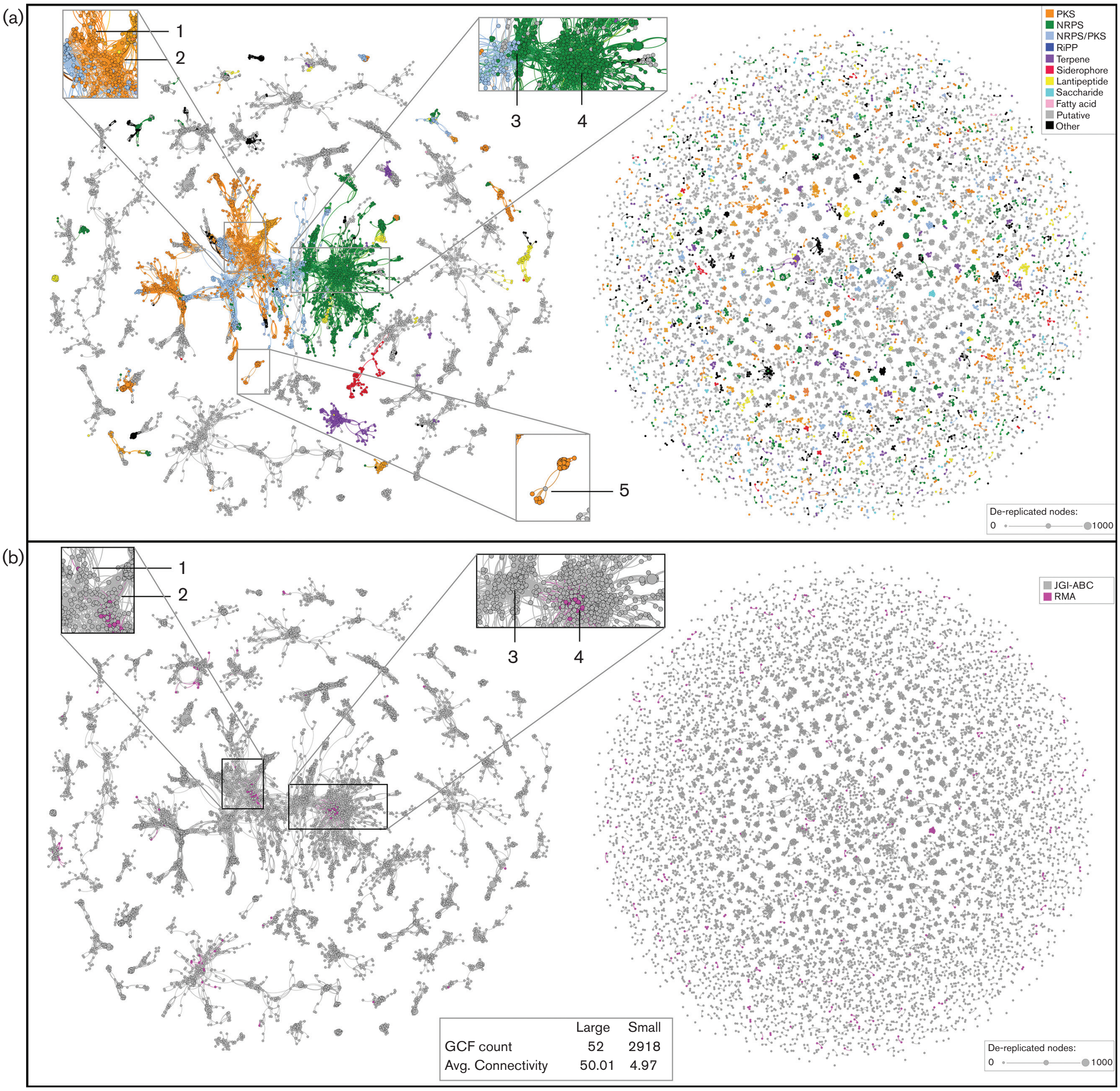

Fig. 1. JGI ABC gene cluster similarity network. Gene cluster similarity networks generated using >3000 Actinobacteria genomes from JGI ABC and the RMA genomes introduced in this study. Gene clusters were identified using antiSMASH3.0 with ClusterFinder (probability >0.8). Each node represents one sequenced gene cluster; any identical gene clusters from multiple sequencings of the same genome were de-replicated and that information is stored in the size of the node. If cluster category was assigned in JGI ABC, those colours were included and correspond to the antiSMASH colouring scheme in (a). RMA BGCs are highlighted in pink in (b). For better visualization, the network was split into large communities (left) and small communities (right). Selected zoom panels in (a) show examples of BGCs with known products. Type I PKS macrolides such as oligomycin (1) (BGC0000117) and erythromycin (2) (BGC0000055) are contained within the larger PKS GCF. Siderophores, such as mycobactin (3) (BGC0001021), lie within the hybrid NRPS-PKS section, while cyclic depsipeptides, including homologues to pristinamycin (4) (BGC0000952), reside in the NRPS GCF. Rifamycin (5) (BGC0000136) and analogues form their own separate GCF. The two zoom panels in (b) show RMA nodes found in proximity to the identified BGCs for known compounds. General network statistics are shown at the bottom of (b).

classes that rival the well-known *Streptomyces* and *Salinispora* producers, which can contain as many as 30 BGCs. These include strains in the genera *Nocardia*, *Rhodococcus*, *Actinomadura*, *Micromonospora*, *Nocardiopsis* and *Gordonia*.

Perhaps more surprising are the genera that lack a large number of secondary metabolite BGCs. These are small genome (<4 Mb) actinomycetes, which also lack NRPS and non-fatty-acid PKS pathways, that belong to the suborder

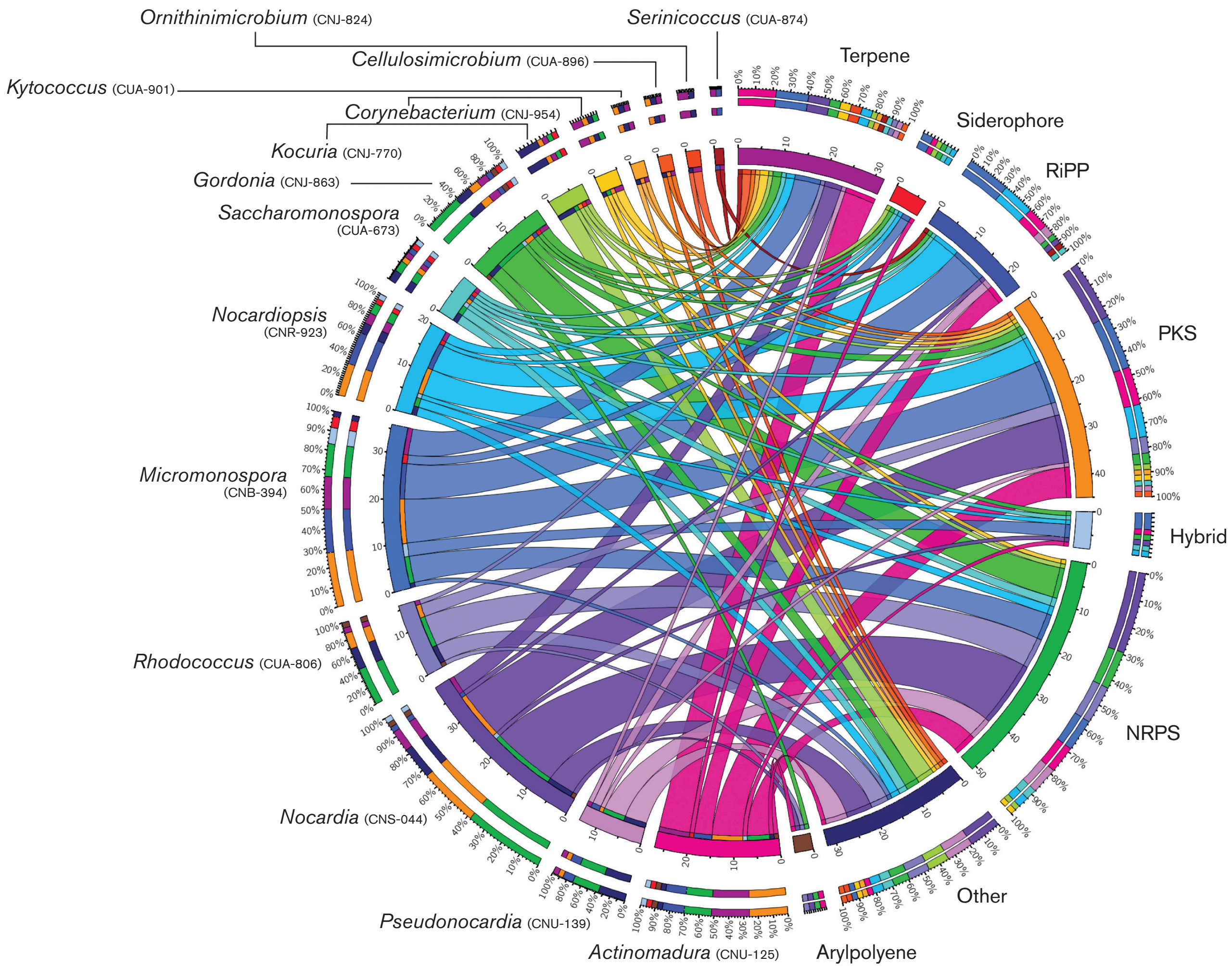

Fig. 2. Circos diagram of RMA pathway diversity. The genomes in this Circos figure include one representative from each genus sequenced and are arranged in ascending genome size, from the top of the circle, counterclockwise to the bottom. Each genome is represented by a different coloured band (left half of the circle) that can be traced from the organism to the types of gene clusters found in that genome (right half of the circle). The width of these bands indicates the number of pathways of that type. The cluster types are also assigned colours based on the colours antiSMASH uses to represent pathway types. The cluster types are designated by their respective colours that make up the outer two rings next to each genome to easily see what portion of the pathways belong to each category. Conversely, the outer two rings next to the gene cluster categories show the proportion of that pathway attributed to each genome represented by the genome colour. The 'Other' category includes clusters that antiSMASH calls Other and Ectoine, as well as uncommon cluster types found in only one or two genomes, such as the following: Butyrolactone, Phenazine, Homoserine lactone, Aminoglycosides, Oligosaccharide and Nucleoside. The Hybrid category includes only Hybrid NRPS-PKS gene clusters. All other hybrid clusters were split into their component parts to get a better overview of gene cluster category diversity (see Tables S3 and S4 for full gene cluster counts). The RiPP category includes clusters identified as Bacteriocin, Lantipeptide or Thiopeptide.

Micrococcineae. Actinomycetes are generally lauded for the production of multiple complex metabolites made by NRPS and PKS pathways; thus, to observe a subset of genomes without this capacity is quite unusual. This is not to say that they do not produce interesting natural products but that these small genome actinomycetes may have different mechanisms of producing secondary metabolites undetectable by current bioinformatic analysis. A taxonomic search in MarinLit for all families in the Fig. S5 suborder returned three recently published groups of compounds: the dermacozines A–J, phenazine compounds isolated from the deep sea *Dermacoccus abyssi* (Abdel-Mageed *et al.*, 2010; Wagner *et al.*, 2014); microluside A, a glycosylated xanthone from a sponge associated *Micrococcus* sp. EG45 (Eltamany *et al.*, 2014); and indole alkaloids from the deep sea *Serinicoccus profundi* (Yang *et al.*, 2013). Additionally, seriniquinone was isolated from *Serinicoccus* strain CNJ-927, a strain sequenced as part of this study (Trzoss *et al.*, 2014). Kocurin, a thiozolyl peptide, has been reported from marine sponge-derived strains *Kocuria marina* F-276310, *Kocuria palustris* F-276345 and *Micrococcus yunnanensis* F-256446 and is hypothesized to be a product of a RiPP (ribosomally synthesized and post-translationally modified peptide) pathway (Palomo *et al.*, 2013). Aside from kocurin, the reported structures from Fig. S5 strains have bioinformatically elusive biosynthetic origins that would likely not be identified by current automated genome mining programs.

Gene cluster similarity network

To address the novelty of the BGCs within the RMA genomes, we classified gene cluster families (GCFs) via a similarity network (Fig. 1). GCF similarity networks have been increasingly utilized to compare gene clusters from large sequencing datasets (Cimermancic *et al.*, 2014; Doroghazi *et al.*, 2014; Ziemert *et al.*, 2014). Grouping gene clusters into larger families allows for quick prioritization of potentially new classes of gene clusters and dereplication of known gene clusters and their associated products (Medema & Fischbach, 2015). In this case, the degree to which the gene clusters we found in the RMAs network with other sequenced gene clusters can give insight into how rare the pathways are in this subset of marine bacteria and whether they are worth pursuing for novel gene cluster discovery.

In order to determine which RMA gene clusters are similar to already sequenced, but not necessarily experimentally confirmed, gene clusters, we used a gene cluster similarity networking approach with the JGI Atlas of Biosynthetic gene Clusters (ABC) database as our comparison set (Hadjithomas *et al.*, 2015). This dataset is composed of all genomes deposited in JGI and run with ClusterFinder and antiSMASH to locate and annotate secondary metabolite gene clusters. We only used actinomycete genomes from the database for comparison to our RMA genomes and dereplicated identical gene clusters from replicate sequencings of the same genome. The resulting network is split into large and small GCFs coloured by BGC type (Fig. 1a), and to show where RMA clusters are incorporated in the network, Fig. 1(b) shows RMA BGCs highlighted in pink. Each BGC is represented by a node, and BGCs that do not have a similarity score over the 0.6 threshold do not appear in the network. Related BGC nodes are connected by edges, and an inclusive set of connected nodes is called a GCF, as was used in Cimermancic *et al.* (2014). In the JGI ABC gene cluster similarity network, 311 of the 1382 (22 %) RMA BGCs appear in the network as nodes, and of those, only 179 nodes (13 % of the total number of predicted gene clusters) are directly linked with JGI ABC nodes. This suggests that 87 % of the RMA gene clusters have similarity scores lower than the threshold used for the network in comparison with any known actinomycete sequence in the JGI ABC database. The dispersal of RMA nodes can be seen in Fig. 1 (b); note that the RMA nodes in the smaller half of the network tend to group into GCFs comprised only of RMA nodes. The distribution of similarity scores (Fig. S3) also shows an average drop for RMA sequences relative to the JGI dataset. Additionally, because a significant fraction of the JGI dataset is represented in the network, we expect that the exclusion of the many RMA gene clusters is not significantly explained by shortcomings with the distance method.

For comparison, we looked at the uniqueness of pathways in the well-characterized genomes of *Streptomyces coelicolor* (non-rare) and *Saccharopolyspora erythraea* (rare). For all *Streptomyces coelicolor* strains included in the JGI ABC set, 54 of the 73 (74 %) gene clusters network with other JGI ABC clusters. For all *Saccharopolyspora erythraea* strains in the JGI ABC set, 25 of the 183 (14 %) gene clusters network with other JGI ABC clusters. This analysis sets a benchmark that, for non-rare strains, the similarity of clusters is quite high while, for rare strains, the connectivity of their clusters is lower, as is seen with the RMA gene clusters.

To assess if the marine environment plays a role in pathway uniqueness, we looked at *Streptomyces xinghaiensis*, the first marine-derived streptomycete to be sequenced (Zhao & Yang, 2011). As a streptomycete, we would expect a high number of pathways shared with other genomes in the database. However, we see that only 9 clusters out of 60 total (15 %) are included in the network, which is on par with the RMA genomes. To further explore the uniqueness of marine streptomycetes, we analysed 24 additional marine isolates (strains listed in the online Supplementary Material). Of the 1925 gene clusters in this group, 412 clusters are in the network (21 %) (Table 1). This low number suggests that marine-derived streptomycetes also harbour unique gene clusters, and we propose that the under-sampling of marine genomes is likely the reason for such novelty.

While the number of unique pathways is important, the diversity of networked gene clusters can also direct which strains to pursue for greater novelty and variety. Diversity indices are used to measure species diversity (Tuomisto, 2010); however, here, we have applied the measure for True diversity (as a function of the Shannon Index) (Jost, 2006) to BGC diversity. To compare the diversity between BGCs in RMA and marine streptomycete genomes, we calculated diversity and normalized it by number of BGCs (Table 1). This measure gives insight into the degree of re-occurring GCFs and, by extension, re-occurring classes of compounds, where a higher value represents a wider range of GCFs and therefore increased likelihood of product diversity. While both RMAs and marine streptomycetes have a low amount of clusters that network, those that do are more diverse in RMAs (0.28) than in marine streptomycetes (0.18). Furthermore, the overlap between RMA and marine streptomycetes is very low, with only four GCFs in common between the two groups (3 %) (Fig. S4). Thus, it is worthwhile to continue sequencing both marine streptomycetes

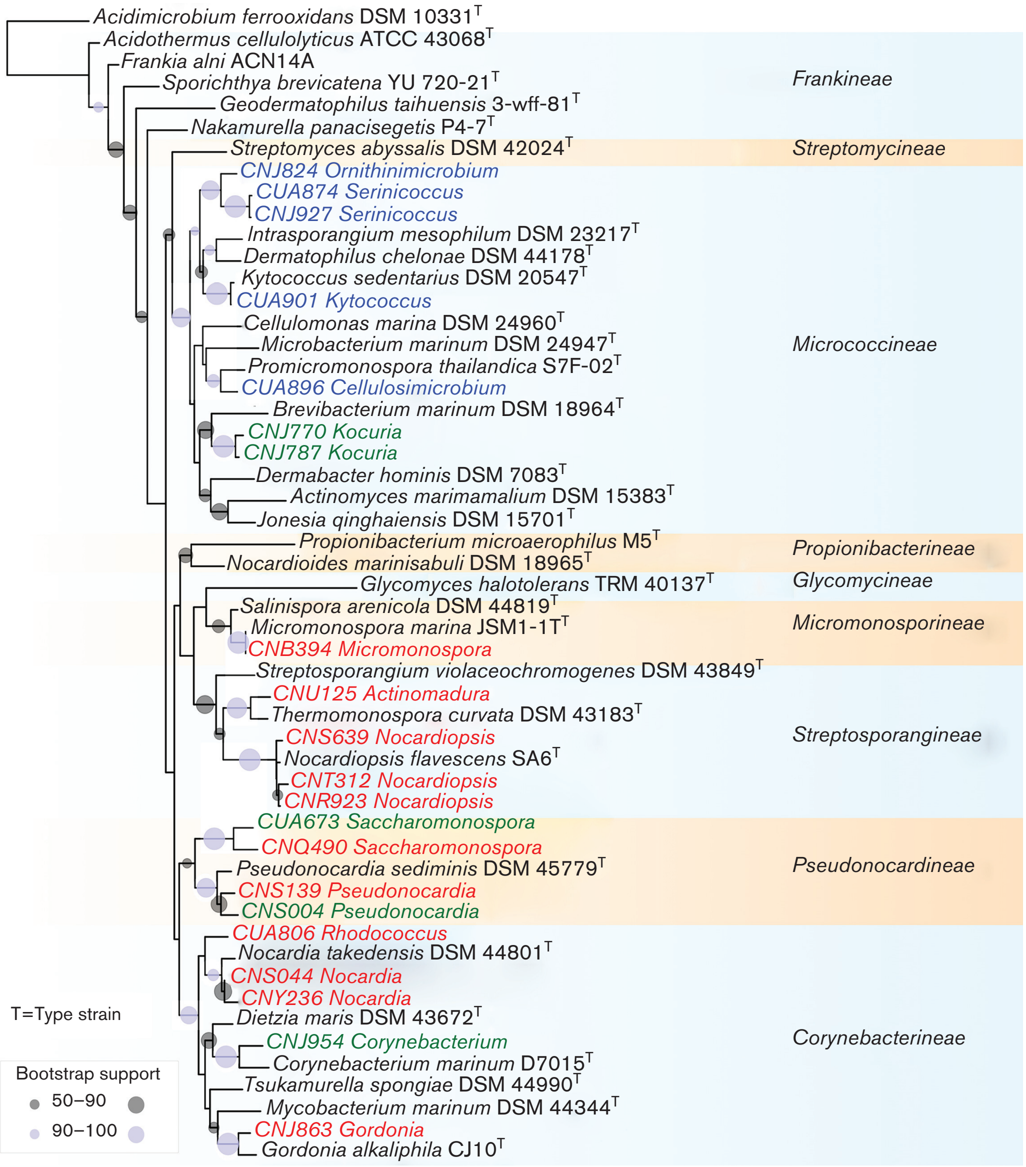
Acidimicrobium ferrooxidans DSM 10331T
Acidothermus cellulolyticus ATCC 43068T
Frankia alni ACN14A
Sporichthya brevicatena YU 720-21T
Geodermatophilus taihuensis 3-wff-81T
Nakamurella panacisegetis P4-7T
Frankineae
Streptomyces abyssalis DSM 42024T
Streptomycineae
CNJ824 Ornithinimicrobium
CUA874 Serinicoccus
CNJ927 Serinicoccus
Intrasporangium mesophilum DSM 23217T
Dermatophilus chelonae DSM 44178T
Kytococcus sedentarius DSM 20547T
CUA901 Kytococcus
Cellulomonas marina DSM 24960T
Microbacterium marinum DSM 24947T
Promicromonospora thailandica S7F-02T
CUA896 Cellulosimicrobium
Brevibacterium marinum DSM 18964T
CNJ770 Kocuria
CNJ787 Kocuria
Dermabacter hominis DSM 7083T
Actinomyces marimamalium DSM 15383T
Jonesia qinghaiensis DSM 15701T
Micrococcineae
Propionibacterium microaerophilus M5T
Nocardioides marinisabuli DSM 18965T
Propionibacterineae
Glycomyces halotolerans TRM 40137T
Glycomycineae
Salinispora arenicola DSM 44819T
Micromonospora marina JSM1-1TT
CNB394 Micromonospora
Micromonosporineae
Streptosporangium violaceochromogenes DSM 43849T
CNU125 Actinomadura
Thermomonospora curvata DSM 43183T
CNS639 Nocardiopsis
Nocardiopsis flavescens SA6T
CNT312 Nocardiopsis
CNR923 Nocardiopsis
Streptosporangineae
CUA673 Saccharomonospora
CNQ490 Saccharomonospora
Pseudonocardia sediminis DSM 45779T
CNS139 Pseudonocardia
CNS004 Pseudonocardia
Pseudonocardineae
CUA806 Rhodococcus
Nocardia takedensis DSM 44801T
CNS044 Nocardia
CNY236 Nocardia
Dietzia maris DSM 43672T
CNJ954 Corynebacterium
Corynebacterium marinum D7015T
Tsukamurella spongiae DSM 44990T
Mycobacterium marinum DSM 44344T
CNJ863 Gordonia
Gordonia alkaliphila CJ10T
Corynebacterineae
T=Type strain
Bootstrap support
50–90
90–100

Fig. 3. 16S rRNA phylogenetic tree of RMA and select type strains. The 21 RMAs used in this study are compared with representative type strains. When possible, marine type strains were chosen. Strains coloured red contain a high number of pathways (>13), those coloured green have a medium number of pathways (5–8) and at least one PKS and/or NRPS pathway (with the exception of CNJ-787) and those coloured blue have a low number of pathways (1–3) and do not contain NRPS or non-fatty-acid PKS pathways. The number of gene clusters in each genome was determined using antiSMASH 3.0 without ClusterFinder. Bootstrap values are indicated by grey circles (50–90) and blue circles (90–100) with increasing size representing increasing confidence.

but, perhaps more importantly, RMAs that have novelty as well as diversity.

Marine-derived genera warranting further study

While we can see that RMAs in general maintain unique gene clusters, it would be beneficial to know which genera contain the least replicated and more diverse gene clusters. We therefore analysed each genus individually and compared the number of new GCFs present from RMAs (Table S5); these represent GCFs not present in the currently sequenced genomes within the same genus. To isolate the effect that the marine environment has, we combined any genome in JGI ABC that had any metadata indicating isolation from a marine environment with those strains sequenced as part of this study. For example, 89 GCFs make

Table 1. RMA versus marine *Streptomyces* true diversity

This table compares 21 RMA genomes from this study and 24 marine streptomycete genomes. While the two groups have similar percentages for in network BGCs, the diversity of RMA BGCs is greater.

	Total BGCs	No. of strains	BGCs in network	No. of GCFs	% BGCs in network	RMA GCFs shared	True diversity	True diversity/ BGC
RMA	1386	21	311	153	22.44	153	86.1595	0.2770
Marine *Streptomyces*	1925	24	412	143	21.40	4	73.6128	0.1787

up the *Nocardiopsis* network (Fig. 4), which is composed of 18 non-marine JGI ABC strains and four RMA strains (three from this study and one from JGI). Of the 38 GCFs that contain RMA gene clusters, 26 (68 %) are exclusively composed of RMA gene clusters. This indicates that the marine *Nocardiopsis* strains include substantial biosynthetic gene diversity not observed in the 18 sequenced *Nocardiopsis* strains that came from sources other than the marine environment. Furthermore, the True diversity, normalized by number of BGCs, is higher for RMA BGCs (0.54) than for non-marine *Nocardiopsis* BGCs (0.17). Because the *Nocardiopsis* strains in the JGI ABC database are from terrestrial, host-associated, aquatic non-marine and other non-marine sources (Fig. S5), the added variety of clusters from the RMA genomes may be due to their marine nature. A recent review details the current natural products from *Nocardiopsis* species and the unique potential of marine *Nocardiopsis* strains (Bennur *et al.*, 2016). This genomic analysis corroborates their assertion that marine *Nocardiopsis* strains are promising in the pursuit of novel bioactive small molecules. In fact, for all genera examined here, all but two (*Actinomadura* and *Kocuria*) have higher normalized True diversity ratios for marine genomes when compared with their non-marine counterparts from JGI ABC (Table S5).

In order to account for how phylogenetic distance between strains affects BGC diversity, we plotted 16S rRNA percent identity against GCF overlap for each pair in the following groups: RMA genomes sequenced as part of this study, JGI marine streptomycetes and non-marine rare actinomycete genomes within the genera examined in this study (Fig. S6). The general trend shows that, with increasing phylogenetic similarity, more GCFs are shared for all three groups.

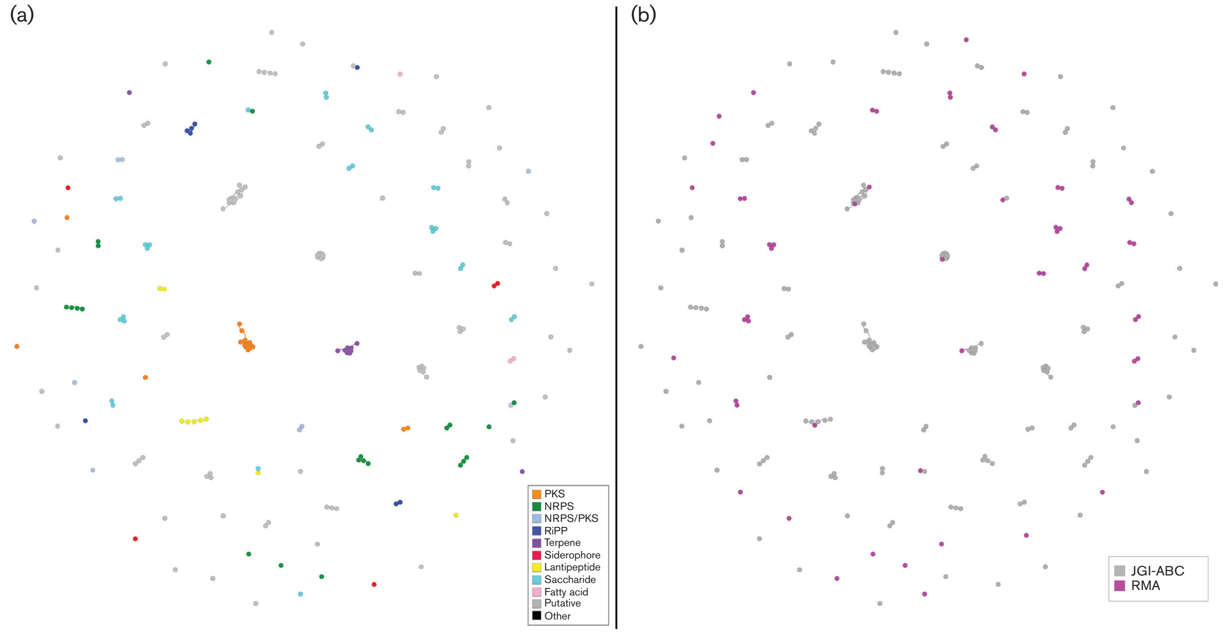

Fig. 4. Gene cluster similarity network of *Nocardiopsis* strains. Gene clusters from all non-marine *Nocardiopsis* genomes in JGI ABC and the three RMA *Nocardiopsis* genomes from this study and one marine *Nocardiopsis* genome from JGI that network with any gene cluster in the large network are retained in this *Nocardiopsis* subset network. Gene clusters coloured by type are shown in (a), while (b) highlights in pink those clusters from the four marine *Nocardiopsis* genomes. RMA clusters tend to form their own GCFs, with only seven RMA nodes connecting with other *Nocardiopsis* nodes.

However, only the non-marine group has pairs that share greater than 20 % GCFs at low phylogenetic distances (below 94 %). This could be due to the lower number of marine genomes in current sequence databases, but it could also indicate more diversity in marine genomes at lower phylogenetic distances. This observed increase in diversity for marine BGCs indicates a hidden biosynthetic potential that warrants more sequencing of RMAs. Although the number of RMA strains included in these analyses is low, with more sequencing, a more complete picture of their full potential will emerge.

The prospect of leaping from genomic code to molecule is fast approaching, with technological advances in sequencing, bioinformatics and heterologous gene expression paving the way to discover and manipulate unknown molecules from proposed biosynthetic pathways. The conventional path to discovering new compounds from microbes is lengthy, work intensive and does not always capture the full potential of high secondary metabolite producing bacteria. Additionally, a large percentage of the microbes from environmental samples have yet to be obtained in culture, representing an impressive biodiversity that remains largely inaccessible to natural product discovery. This new era of fast, easy sequencing can open doors to exploring microbial communities harbouring unprecedented natural chemistry. Though they have been examined on an individual basis for natural product discovery, RMAs as a group represent a widely untapped resource for new BGCs. Certain suborders of RMAs have very high potential to possess unique gene clusters, which may encode unprecedented chemical scaffolds. Further efforts should focus on culturing RMAs and sequencing their genomes to survey their full biosynthetic potential.

Gene cluster similarity networks provide a powerful tool to quickly assess the uniqueness of a given genome's biosynthetic pathways. However, there are limitations to such an automated method. We encountered gene cluster trimming to be a setback in some cases where antiSMASH would overcall the number of genes in a gene cluster. Automated and accurate gene cluster boundary delineation will only improve the precision of gene cluster similarity networks.

It has been estimated that all Actinobacteria biosynthetic diversity can be reached by sequencing only 15 000 actinomycete genomes (Doroghazi *et al.*, 2014). However, as the authors state, this estimation is based on what is currently in our sequenced databases, which is largely terrestrial streptomycetes and clinical isolates. With the amount of novelty seen in RMA and even marine streptomycete genomes, this estimation can be re-visited. We may be saturating the pool with terrestrial strains, but thus far, the potential from marine genomes has yet to be fully realized. Aside from the marine environment, other unexplored habitats will likely expand the number of genomes we need to sequence to see full secondary metabolite pathway potential. Bacteria in symbioses (such as endophytes in plants and endosymbionts in sponges), microbes living in extreme environments and un-cultured bacteria all represent large potential reservoirs of unknown biosynthetic capacity (Brader *et al.*, 2014; Chávez *et al.*, 2015). Using the RMA genomes as a glimpse into the potential of under-sampled genomes showcases the importance of expanding our sequencing efforts from the mainly terrestrial and clinical isolates that exist today. Improved sampling and culturing practices, along with enhanced molecular biology, sequencing and metagenome assembly techniques, will pave the way for accessing previously inaccessible genomes in the search for new biosynthetic potential.

Availability of data and material

The genome assemblies for each strain in this article are available at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers can be found in Table S1. Additionally, these strains can be found through the JGI IMG Expert Review portal at https://img.jgi.doe.gov. Genome assemblies can be accessed using the NCBI accession numbers or JGI OIDs provided in Table S1. Custom python scripts written to create the gene cluster similarity network are deposited and annotated at: https://bitbucket.org/malanjary_ut/clustsimscore.

ACKNOWLEDGEMENTS

M. A. S. acknowledges support from the Edna Bailey Sussman Foundation, the Louise Torraco Memorial Doctorate in Science Scholarship from UNICO and the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). B. S. M. and P. R. J. acknowledge support from the National Institutes of Health (NIH) grant R01-GM085770. The genome sequencing conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. A. K. was supported by the Russian Science Foundation (grant 14-50-00069). N. Z. and M. M. A. acknowledge support from the German Center for Infection Research (DZIF) (TTU 09.704).

REFERENCES

Abdel-Mageed, W. M., Milne, B. F., Wagner, M., Schumacher, M., Sandor, P., Pathom-aree, W., Goodfellow, M., Bull, A. T., Horikoshi, K. & other authors (2010). Dermacozines, a new phenazine family from deep-sea dermacocci isolated from a Mariana Trench sediment. *Org Biomol Chem* **8**, 2352–2362.

Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. (2014). Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol* **41**, 175–184.

Baltz, R. H. (2006). Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J Ind Microbiol Biotechnol* **33**, 507–513.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. & other authors (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477.

Bastian, M., Heymann, S. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third International*

AAAI Conference on Weblogs and Social Media. San Jose McEnery Convention Center.

Bennur, T., Ravi Kumar, A., Zinjarde, S. S. & Javdekar, V. (2016). *Nocardiopsis* species: a potential source of bioactive compounds. *J Appl Microbiol* **120**, 1–16.

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H. & other authors (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147.

Bérdy, J. (2005). Bioactive microbial metabolites. *J Antibiot* **58**, 1–26.

Bérdy, J. (2012). Thoughts and facts about antibiotics: where we are now and where we are heading. *J Antibiot* **65**, 441.

Brader, G., Compant, S., Mitter, B., Trognitz, F. & Sessitsch, A. (2014). Metabolic potential of endophytic bacteria. *Curr Opin Biotechnol* **27**, 30–37.

Cane, D. E. & Ikeda, H. (2012). Exploration and mining of the bacterial terpenome. *Acc Chem Res* **45**, 463–472.

Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Choi, S.-S., Kim, H.-J., Lee, H.-S., Kim, P. & Kim, E.-S. (2015). Genome mining of rare actinomycetes and cryptic pathway awakening. *Process Biochem* **50**, 1184–1193.

Chávez, R., Fierro, F., García-Rico, R. O. & Vaca, I. (2015). Filamentous fungi from extreme environments as a promising source of novel bioactive secondary metabolites. *Front Microbiol* **6**, 903.

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M. & other authors (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421.

Doroghazi, J. R. & Metcalf, W. W. (2013). Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **14**, 611.

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L. & Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* **10**, 963–968.

Eltamany, E. E., Abdelmohsen, U. R., Ibrahim, A. K., Hassanean, H. A., Hentschel, U. & Ahmed, S. A. (2014). New antibacterial xanthone from the marine sponge-derived *Micrococcus* sp. EG45. *Bioorg Med Chem Lett* **24**, 4939–4942.

Eustáquio, A. S., Nam, S.-J., Penn, K., Lechner, A., Wilson, M. C., Fenical, W., Jensen, P. R. & Moore, B. S. (2011). The discovery of Salinosporamide K from the marine bacterium *Salinispora pacifica* by genome mining gives insight into pathway evolution. *ChemBioChem* **12**, 61–64.

Fenical, W. & Jensen, P. R. (2006). Developing a new resource for drug discovery: marine actinomycete bacteria. *Nat Chem Biol* **2**, 666–673.

Gomez-Escribano, J., Alt, S. & Bibb, M. (2016). Next generation sequencing of Actinobacteria for the discovery of novel natural products. *Mar Drugs* **14**, 78.

Gontang, E. A., Fenical, W. & Jensen, P. R. (2007). Phylogenetic diversity of Gram-positive bacteria cultured from marine sediments. *Appl Environ Microbiol* **73**, 3272–3282.

Gontang, E. A., Gaudêncio, S. P., Fenical, W. & Jensen, P. R. (2010). Sequence-based analysis of secondary-metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487–2499.

Hadjithomas, M., Chen, I. M., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., Huang, J., Reddy, T. B., Cimermančič, P. & other authors (2015). IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932-15.

Hu, Y. (2006). Efficient, high-quality force-directed graph drawing. *Mathematica J* **10**, 37–71.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* **21**, 526–531.

Jensen, P. R., Gontang, E., Mafnas, C., Mincer, T. J. & Fenical, W. (2005). Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environ Microbiol* **7**, 1039–1048.

Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. (2014). Challenges and triumphs to genomics-based natural product discovery. *J Ind Microbiol Biotechnol* **41**, 203–209.

Jensen, P. R., Moore, B. S. & Fenical, W. (2015). The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738–751.

Jost, L. (2006). Entropy and diversity. *Oikos* **113**, 363–375.

Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511–518.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645.

Land, M. L., Hyatt, D., Jun, S. R., Kora, G. H., Hauser, L. J., Lukjancenko, O. & Ussery, D. W. (2014). Quality scores for 32,000 genomes. *Stand Genomic Sci* **9**, 20.

Lazzarini, A., Cavaletti, L., Toppo, G. & Marinelli, F. (2001). Rare genera of actinomycetes as potential producers of new antibiotics. *Antonie Van Leeuwenhoek* **79**, 399–405.

Letunic, I. & Bork, P. (2016). Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242–W245.

Lin, K., Zhu, L. & Zhang, D. Y. (2006). An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* **22**, 2081–2086.

Manivasagan, P., Venkatesan, J., Sivakumar, K. & Kim, S. K. (2014). Pharmaceutically active secondary metabolites of marine actinobacteria. *Microbiol Res* **169**, 262–278.

Medema, M. H. & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nat Chem Biol* **11**, 639–648.

Medema, M. H., Takano, E. & Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* **30**, 1218–1223.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J. & other authors (2015). Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* **11**, 625–631.

Monciardini, P., Iorio, M., Maffioli, S., Sosio, M. & Donadio, S. (2014). Discovering new bioactive molecules from microbial sources. *Microb Biotechnol* **7**, 209–220.

Moore, B. S., Kalaitzis, J. A. & Xiang, L. (2005). Exploiting marine actinomycete biosynthetic pathways for drug discovery. *Antonie van Leeuwenhoek* **87**, 49–57.

Nett, M., Ikeda, H. & Moore, B. S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* **26**, 1362–1384.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A. & other authors (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**, 714–737.

Nylander, J. A. A. (2004). MrModeltest v2. Evolutionary Biology Centre, Uppsala University: Program distributed by the author.

Palomo, S., González, I., de la Cruz, M., Martín, J., Tormo, J. R., Anderson, M., Hill, R. T., Vicente, F., Reyes, F. & other authors (2013). Sponge-derived *Kocuria* and *Micrococcus* spp. as sources of the new thiazolyl peptide antibiotic Kocurin. *Mar Drugs* **11**, 1071–1086.

Podell, S. & Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**, R16.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J. & other authors (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Subramani, R. & Aalbersberg, W. (2013). Culturable rare Actinomycetes: diversity, isolation and marine natural product discovery. *Appl Microbiol Biotechnol* **97**, 9291–9321.

Tang, X., Li, J., Millán-Aguiñaga, N., Zhang, J. J., O'Neill, E. C., Ugalde, J. A., Jensen, P. R., Mantovani, S. M. & Moore, B. S. (2015). Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841–2849.

Tiwari, K. & Gupta, R. K. (2012). Rare actinomycetes: a potential storehouse for novel antibiotics. *Crit Rev Biotechnol* **32**, 108–132.

Trzoss, L., Fukuda, T., Costa-Lotufo, L. V., Jimenez, P., La Clair, J. J. & Fenical, W. (2014). Seriniquinone, a selective anticancer agent, induces cell death by autophagocytosis, targeting the cancer-protective protein dermcidin. *Proc Natl Acad Sci U S A* **111**, 14687–14692.

Tuomisto, H. (2010). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **164**, 853–860.

Udwary, D. W., Zeigler, L., Asolkar, R. N., Singan, V., Lapidus, A., Fenical, W., Jensen, P. R. & Moore, B. S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc Natl Acad Sci U S A* **104**, 10376–10381.

Wagner, M., Abdel-Mageed, W. M., Ebel, R., Bull, A. T., Goodfellow, M., Fiedler, H. P. & Jaspars, M. (2014). Dermacozines H-J isolated from a deep-sea strain of *Dermacoccus abyssi* from Mariana Trench sediments. *J Nat Prod* **77**, 416–420.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R. & other authors (2015). antiSMASH 3.0 — a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237–W243.

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D. E. & Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. *Proc Natl Acad Sci U S A* **112**, 857–862.

Yamanaka, K., Reynolds, K. A., Kersten, R. D., Ryan, K. S., Gonzalez, D. J., Nizet, V., Dorrestein, P. C. & Moore, B. S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* **111**, 1957–1962.

Yang, X.-W., Zhang, G.-Y., Ying, J.-X., Yang, B., Zhou, X.-F., Steinmetz, A., Liu, Y.-H. & Wang, N. (2013). Isolation, characterization, and bioactivity evaluation of 3-((6-methylpyrazin-2-yl)methyl)-1*H*-indole, a new alkaloid from a deep-sea-derived actinomycete *Serinicoccus profundi* sp. nov. *Mar Drugs* **11**, 33–39.

Zhao, X. & Yang, T. (2011). Draft genome sequence of the marine sediment-derived actinomycete *Streptomyces xinghaiensis* NRRL B24674T. *J Bacteriol* **193**, 5543.

Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E. & Jensen, P. R. (2012). The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **7**, e34064.

Ziemert, N., Lechner, A., Wietz, M., Millan-Aguinaga, N., Chavarria, K. L. & Jensen, P. R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci U S A* **111**, E1130–E1139.

Zotchev, S. B. (2012). Marine actinomycetes as an emerging resource for the drug development pipelines. *J Biotechnol* **158**, 168–175.

Edited by: P. W. O'Toole and Y. Ohnishi