COMMUNICATIONS
BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-019-0333-6 OPEN

Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters

Francesco Del Carratore[1], Konrad Zych[2], Matthew Cummings[1], Eriko Takano[1], Marnix H. Medema[3] & Rainer Breitling[1]

The biosynthetic machinery responsible for the production of bacterial specialised metabolites is encoded by physically clustered group of genes called biosynthetic gene clusters (BGCs). The experimental characterisation of numerous BGCs has led to the elucidation of subclusters of genes within BGCs, jointly responsible for the same biosynthetic function in different genetic contexts. We developed an unsupervised statistical method able to successfully detect a large number of modules (putative functional subclusters) within an extensive set of predicted BGCs in a systematic and automated manner. Multiple already known subclusters were confirmed by our method, proving its efficiency and sensitivity. In addition, the resulting large collection of newly defined modules provides new insights into the prevalence and putative biosynthetic role of these modular genetic entities. The automated and unbiased identification of hundreds of co-evolving group of genes is an essential breakthrough for the discovery and biosynthetic engineering of high-value compounds.

[1] Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, Faculty of Science and Engineering, University of Manchester, 131 Princess Street, Manchester, M1 7DN, United Kingdom. [2] Structural & Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany. [3] Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands. Correspondence and requests for materials should be addressed to R.B. (email: rainer.breitling@manchester.ac.uk)

Microbial specialised metabolism is a rich source of high-value and biochemically active compounds of immense biotechnological and biomedical potential[1,2]. The enzymatic pathways responsible for the biosynthesis of such compounds are encoded by physically clustered groups of genes called biosynthetic gene clusters (BGCs). These sometimes very large[3] genomic regions have a high modular structure at the genetic level[4,5]. For instance, it has been previously observed that certain classes of BGC share co-evolving multi-gene subclusters, which work together as a unit for the implementation of a specific biosynthetic function[4,5]. Numerous examples of such genetic entities have been described[6–19]. BGCs often harbour more than one subcluster, and can be composed almost exclusively of these genetic building blocks, as is the case for aminocoumarins: the groups of genes responsible for the biosynthesis of the deoxysugar ring moiety, the aminocoumarin core and the pyrrole ring moieties each form a discrete subcluster[20,21]. These subclusters are not constrained to the aminocoumarins, however: the pyrrole ring subcluster used in the biosynthesis of at least four different end compounds[20–25]. All these subclusters have been detected through the experimental characterisation of numerous BGCs and provide a very useful benchmark when developing a method able to automatically detect similar genetic entities. The deconstruction of BGCs into subclusters encoding discrete chemical moieties has been used to generate novel compounds by combinatorial biosynthesis[26]; therefore, their discovery and characterisation is potentially a great help to synthetic biology approaches to BGC reconstruction de novo, providing an extremely useful tool for the biotechnology research community aiming at exploiting the full commercial and clinical potential of microbial metabolism. The relevance and the (evolutionary) exchange of these modules across microbial species are still under study, however[4]. Recent advances in computational biology have allowed the identification of millions of putative BGCs[27] by the systematic analysis of DNA sequence[28] using, e.g., freely available BGC-mining tools, such as antiSMASH[29], BAGEL[30], PRISM[31], and ClusterFinder[4]. Here, we take this approach one step further: with the intent of elucidating the relevance and exchange of biosynthetic subclusters in the evolution of BGCs, we developed a statistical method for the detection of subclusters. This algorithm successfully detected 185,718 statistically significant putative subclusters (hereafter, called modules) in 12,842 predicted BGCs in microbial sequences from GenBank in a systematic and automated manner. Although accurately detecting already known subclusters, our method is able to rigorously define numerous novel modules and provide new insights into the prevalence of putative functional subclusters and their role in specialised metabolite biosynthesis. The resulting library of statistically defined modules is a rich resource for the specialised metabolite research community. In fact, these modules could be used for the design of BGCs that are likely to encode the biosynthesis of molecules with novel combinations of known chemical moieties. Moreover, these results yield an unprecedented insight into the degree of modular organisation of the specialised biosynthetic machinery across the entire microbial kingdom, and significantly reduce the role played by serendipity in the initial identification of individual modules by guiding the selection of statistically supported promising candidates through a comprehensive and unbiased automated approach. In addition, the method described here is able to identify strongly supported modules with currently unknown functions. Such modules are potentially responsible for the biosynthesis of novel chemical (sub)structures and represent valuable guides for the targeted mining of microbial genomes for new drug candidates. This work will also be a great asset for cluster prediction tools such as antiSMASH[29], as it can be used to annotate predicted BGCs by suggesting which genes in a BGC function together as discrete units in a complex biosynthetic pathway; in addition, this module-based annotation can also help in cluster boundary prediction.

Results

Module detection algorithm. After generating a collection of predicted BGCs through antiSMASH[29], the method relied on the orthoMCL package v 1.4[32] for the annotation of specialised metabolite Clusters of Orthologous Genes (smCOG). All the detected smCOGs were next organised into a network where two smCOG are connected if they share a statistically significant number of adjacency or colocalization interactions (the evaluation of the statistical significance of the number of interactions is described in the Methods). Finally, all the fully connected subgraphs (i.e., cliques) found in the network were considered as putative biosynthetic modules, this is a statistically very conservative approach, as it requires that all individual interactions between module members to be highly significant. This algorithm successfully detected 185,718 statistically significant putative subclusters (hereafter, called modules) in 12,842 predicted BGCs in microbial sequences from GenBank in a systematic and automated manner. Although this is a rather large number when compared with the number of BGCs present in our data set, it is important to consider that the numbers appear somewhat inflated by the common appearance of groups of nested modules, where smaller (less specific, but statistically highly significant) modules are contained within larger (more specific) modules in various combinations; examples of this biologically important (and expected) phenomenon are discussed below. More importantly, this statistically strongly supported subset is only a tiny fraction of the > 10^{34} possible modules. The method is briefly summarised in Fig. 1, whereas a more detailed description is available in the Methods. Given the large number of statistically significant putative modules detected by our algorithm, a method for ranking the 185,718 putative modules would strongly benefit the exploration of our database. This has been achieved through the Module Interest Benchmarking (MIB) score. After ranking the modules according to different metrics (number of BGCs containing the module, number of BGCs containing the module and present in the MIBiG database, module size, strictest *p*-value threshold, number of different compound classes and their Shannon entropy, and the percentage of smCOGs members of a specific category), the MIB score is simply computed as a weighted sum of all ranks obtained for each module. Subsequently, the obtained values are rescaled over the range from 1 (least interesting) to 100 (more interesting). The default values of the weights used are: 2 for the length of the modules (longer, and thus more specific, being better), 15 for the Shannon entropy, 10 for the number of BGCs containing the module, 5 for the maximum *p*-value threshold, 10 for the percentage of tailoring smCOGs, and 0 for all the remaining criteria. However, all the weights can be adjusted by the user to increase or decrease the contribution of each criterion in the prioritisation. A more detailed description of the prioritisation procedure can be found in the Methods. In order to evaluate the biological and evolutionary relevance of the detected modules, we first checked for the appearance of well-characterised and previously reported modules in our database, starting with the previously described aminocoumarin subclusters[20,21].

Aminocoumarin, deoxysugar, and pyrrole ring. Canonical aminocoumarin-specialised metabolites are highly modular in structure and comprise an aminocoumarin core moiety, which is often decorated with deoxysugars and pyrrole rings (Fig. 2). Each of these distinct chemical moieties is encoded by a discrete suite

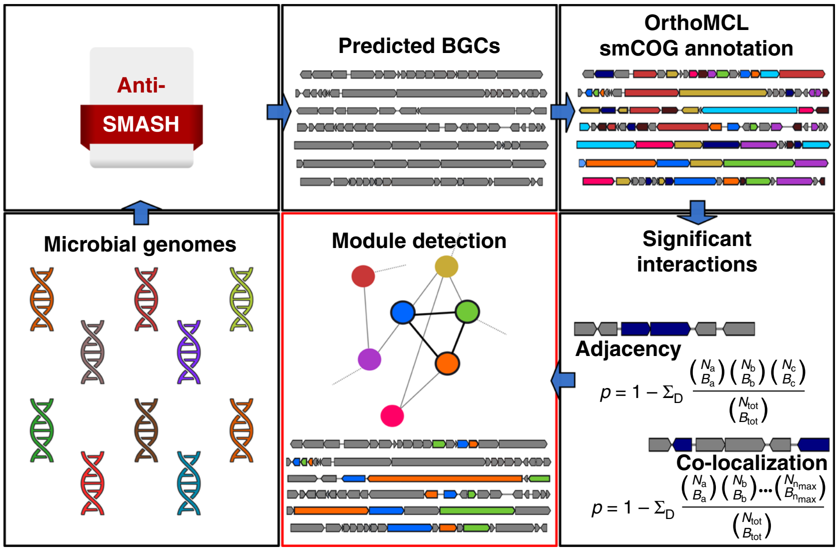
Anti-
SMASH
Predicted BGCs
OrthoMCL
smCOG annotation
Microbial genomes
Module detection
Significant
interactions
Adjacency
$p = 1 - \Sigma_D \frac{\binom{N_a}{B_a}\binom{N_b}{B_b}\binom{N_c}{B_c}}{\binom{N_{tot}}{B_{tot}}}$
Co-localization
$p = 1 - \Sigma_D \frac{\binom{N_a}{B_a}\binom{N_b}{B_b}\cdots\binom{N_{n_{max}}}{B_{n_{max}}}}{\binom{N_{tot}}{B_{tot}}}$

Fig. 1 Module detection algorithm. By processing the entire collection of microbial genomes considered in this study, antiSMASH predicted tens of thousands BGCs. OrthoMCL was then used for the smCOG annotation. Next, these smCOGs were organised into a network where two smCOGs are connected only if they share a significant number of adjacency or colocalization interactions. Fully connected sub-graphs (cliques) are considered as putative biosynthetic modules

of genes, a subcluster, which appear to mix-and-match in Nature to produce chemically diverse end compounds. Previously, experimentally validated subclusters encoding the biosynthesis of all three chemical moieties were detected by our bioinformatics analysis. Figure 2 shows 3 modules that perfectly cover three well-known subclusters that are all found both in the clorobiocin and the coumermycin BGCs. Specifically, module M142052 (MIB score = 60.22, number of BGCs covered = 4) covers the group of genes responsible of the biosynthesis the deoxysugar ring present in clorobiocin (*cloM, P, T, U, V, W*), novobiocin (*novM, P, T, U, V, W*), and coumermycin (*couM, P, T, U, V, W*)[20]. Module M2466 (MIB score = 68.32, BGCs = 6) targets the genes encoding the aminocoumarin group in the clorobiocin BGC (*cloI-L*) novobiocin (*novI-L*) and simocyclinone (*simI, J1, J2, K, L*)[20,21]. Module M113610 (MIB score = 95.36, BGCs = 33) covers the three-gene subcluster responsible for the biosynthesis of the pyrrole ring in a number of different BGCs: clorobiocin (*cloN3-N5*)[20], prodigiosin (*rphW, M, O*)[22], coumermycin (*couN3-N5*)[20], calcimycin (*schN1-N3*)[23], indanomycin (*idmI- K*)[24], and pyoluteorin (*pltE-L*)[25].

ACP-linked PKS extender modules (hierarchy of modularity). The previously described five-gene subcluster responsible for the formation of methoxymalonyl-ACP, another classical example of a functional pathway module, is covered by module M130554 (MIB score = 96.62; containing smCOG10382, smCOG10393, smCOG10031, smCOG10236, and smCOG10154). This module is found in a total of 28 BGCs, 12 of which are fully characterised BGCs found in the MiBiG database[33]: galbonolides[6], tautomycin[7,8], oxazolomycin[9,10], FK520[11,12], macbecin[13,14], ansatomycin[13,15], geldamycin[13,16], herbimycin[13,16], concanamycin[17], bafilomycin[18], apoptolidin[19], and nocathiacin[34]. Interestingly, the group of genes *noc-5–noc-9* from the nocathiacin BGC are unlikely to play a role in the biosynthesis of this ribosomally synthesised and post translationally modified peptide. Instead, it is more likely that, in this genetic context, module M130554 forms part of an additional, flanking BGC within the corresponding genome (which is yet to be completely sequenced). Module M130554 encompasses a smaller related module, M112949. This reduced module lacks the *O*-methyltransferase (O-MT) responsible for methylation of the *α*-carbon hydroxyl group (smCOG10382) and is found in a total of 73 BGCs, 15 of which have been fully characterised, and shows a higher MIB score than module M130554 (99.09). Interestingly, module M112949 appears to comprise a minimal complement of genes from which an array of unusual acyl-ACP extender units are derived, e.g., aminomalonyl-ACP (zwittermycin A)[35], alternative routes to methoxymalonyl-ACP (chondrochloren)[36] and an unusual glycolate containing-ACP substrate (pellasoren)[37]. In the case of the chondrochloren BGC, the module M112949 cooperates in the synthesis of methoxymalonyl-ACP despite the lack of a discrete gene encoding an O-MT, instead the O-MT function complemented by an enzymatic domain within the polyketide synthase subunit *cndE*[36] (smCOG10053). The *cndE* O-MT is not part of the module as defined here, as *cndE* is not annotated as a methyltransferase. This indicates how the statistically defined modules can be used for a targeted search of missing enzymes, redundancy within BGCs and functional gene homologues: the observed O-MT domain fusion (smCOG10053) is common, however, and forms part of its own module (M152817, MIB score = 92.83) comprising module M112949 with the addition of smCOG10053, highlighting two convergent routes to methoxymalonyl-ACP formation. The combinatorial complexities of acyl-ACPs do not stop here, however. Module M112949 covers 4 out of the 5 genes responsible for the biosynthesis of a glycolate type extender unit in the pellasoren BGC[37]. All five genes are covered by module M118907, which contains an additional acetyltransferase, ACP and O-MT multifunctional gene product (*PelG*, smCOG11537) predicted to be responsible for the loading, tethering, and methylation of 1,3 bisphosphoglycerate[37]. Furthermore, the zwittermycin A BGC contains a nine-gene subcluster encoding the biosynthesis of two different acyl-ACP PKS extender units: (2S)-aminomalonyl-ACP and (2R)-hydroxymalonyl-ACP[38]. This big subcluster is completely covered by M118911 (MIB score = 87.47), which is composed of module M112949 plus smCOG13061 (*zmaJ*). Deconstruction of this subcluster shows module M112949 to comprise two ACP genes (*zmaD* and *zmaH*, smCOG10236) and two acyl-CoA dehydrogenase genes (*zmaE* and *zmaI*, smCOG10031) which are

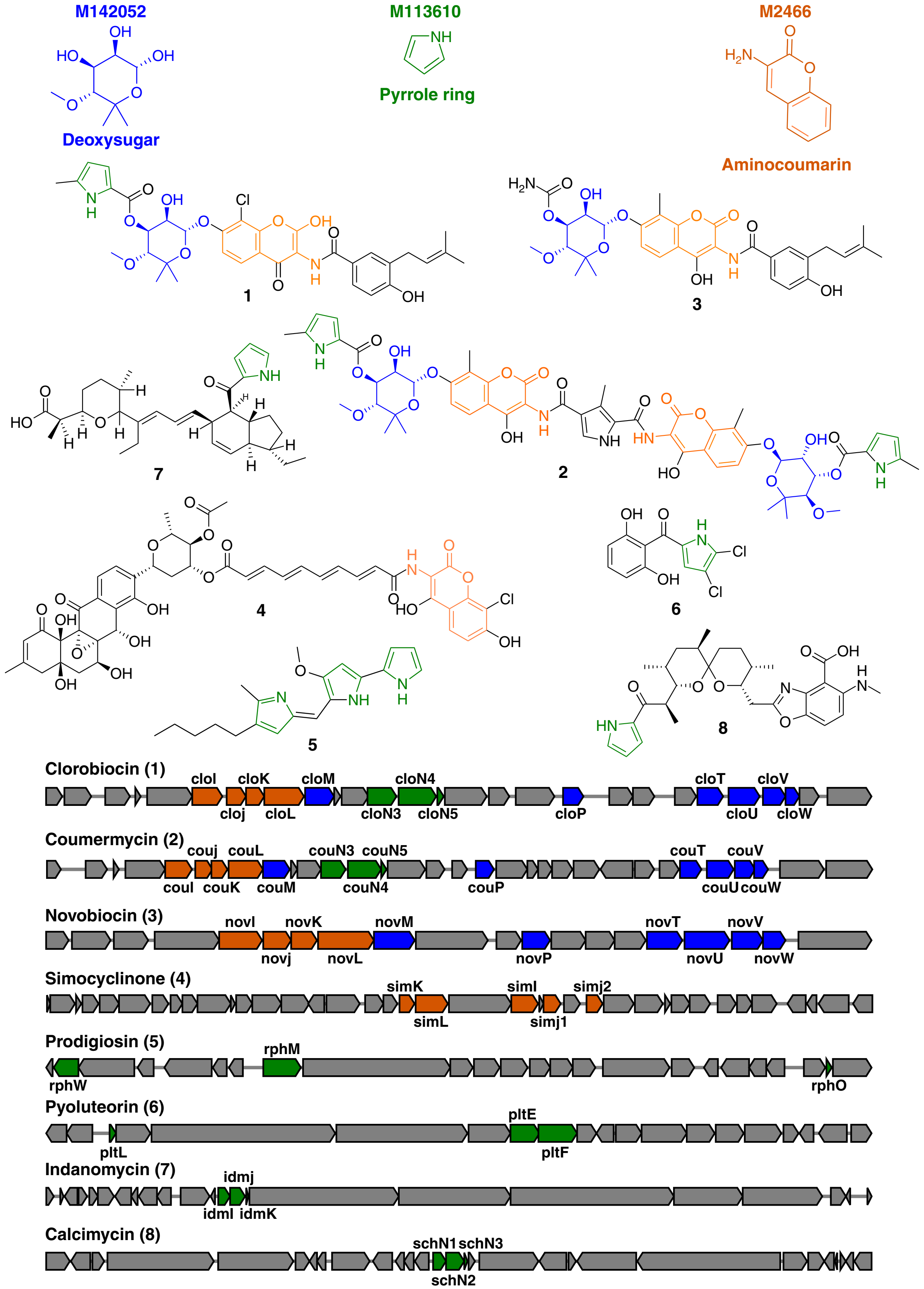

Fig. 2 Deoxysugar, pyrrole ring, and aminocoumarin modules. Overview of the clorobiocin, coumermycin, novobiocin, simocyclinone, prodigiosin, pyluteorin, indanomycin, and calcimycin BGCs. When present, the genes covered by Module M142052 (blue), Module M113610 (green), and Module M2466 (orange) are highlighted in the clusters. The chemical moieties related to the modules are highlighted in the chemical structures. Clusters are not drawn to scale

orthogonal for (2R)-hydroxymalonyl- and (2S)-aminomalonyl-ACP formation correspondingly (Fig. 3, yellow and green ORFs respectively), a single smCOG10154 acyl-ACP dehydrogenase predicted to be promiscous for both acyl-intermediates, and a glyceryl-S-*ZmaD* synthase (smCOG10393), *zmaN*, specific for (2R)-hydroxymalonyl-ACP formation. The biosynthesis of the (2S)-aminomalonyl-ACP necessitates a dedicated seryl-AMP synthetase, *zmaJ*, to load L-serine onto *ZmaH*. This enzymatic domain is annotated by an alternative smCOG, therefore falling outside of the module M112949, instead being completely covered by module M118912 (MIB score = 92.61), composed of smCOG10031, smCOG10154, smCOG10236, and smCOG13061, and eluding to a common small subcluster comprising smCOG10031, smCOG10154, and smCOG10236 (M176566, MIB score = 99.06, BGCs = 99). Figure 3 shows this representative example of the hierarchical organisation of subclusters. Several additional known subclusters have been identified by our method. For example, the biosynthesis of 4-methyl-3-hydroxyanthranilic acid has been associated to module M108999 (Supplementary Fig. 1), whereas the biosynthesis of 2,3-dihydroxy-benzoic acid (DBHA) has been associated to module M21869 (Supplementary Fig. 2). Moreover, the two overlapping modules M103444 and M131293 have been associated to the biosynthesis of 9- and 10-membered enediyne rings (Supplementary Fig. 3). Finally, module M107196 and module M11279 have been associated with the biosynthesis of β-carotene and ectoine, respectively (Supplementary Note 1).

Exploring the full collection of detected modules. As previously mentioned, the MIB score allows us to prioritise the detected modules by considering different criteria with different weights at the same time. Supplementary Data 1 show a selection of the available metrics for the modules previously discussed. It is noteworthy that these modules show a significantly high MIB score, demonstrating that experimentally verified modules are strongly favoured by this prioritisation method. All the previously mentioned modules are in the top 25% when considering their MIB scores. According to the Fisher's exact test, the probability of observing this situation simply by chance is equal to p-value = 5.809×10^{-10}. This increases our confidence that previously unreported modules with comparably high MIB scores are also biochemically interesting evolutionary units. When selecting the examples for the discussion below, we considered only modules with at least one hit in the MIBiG database[33], thus focusing on examples where at least some chemical information is available to validate our interpretation, and we iteratively filtered out all the modules containing at least one smCOG found in the modules previously discussed (so that at each step we only consider modules that are clearly different from modules already examined) we then selected the module showing the highest MIB score. With the first iteration we identified the four-smCOG module M63477 (MIB score = 100), composed of smCOG10029, smCOG10228, smCOG10252, and smCOG10310. This fairly common module (found in 77 different BGCs) is found in the streptomycin BGC of *Streptomyces griseus* covering four genes: *argD* SG7F10.54, *argB* SG7F10.53, *argC* SG7F10.51, and *argJ* SG7F10.52. These four genes are encoding the enzymes involved in the biosynthesis of L-ornithine from L-glutamate, and are therefore undoubtedly a functionally and evolutionarily coherent unit. In the case of the streptomycin gene cluster, however, Module M63477 is most a likely part of a neighbouring biosynthetic gene cluster, merged as a result of the greedy nature of the antiSMASH detection algorithm. The wide distribution of this module suggests that the ornithine precursor is more widely used than previously appreciated, possibly as a precursor to diaminopropionate biosynthesis (from ornithine and serine), as has been suggested for the formation of stenothricin[39].

The second iteration selected the module M100203 (MIB score = 99.55, BGCs = 47), which contains three smCOGs: smCOG10555, smCOG10642, and smCOG11025. Two of the BGCs targeted by this module are present in the MIBiG database, encoding the biosynthesis of A-503083 A and A-500359 A, respectively. In both of these clusters, the module covers three genes that are predicted to encode three subunits of a functional carbon monoxide dehydrogenase complex with an unclear role in biosynthesis[40,41]. The inclusion of this three-gene module across a wide variety of biosynthetic pathways (Shannon's entropy = 2.26) pinpoints these poorly characterised genes to either play a fundamental, and not yet understood, role in specialised metabolism, or alternatively to result from greedy BGC prediction in a similar fashion to module M63477. In either case this group of genes deserves closer experimental evaluation.

The third iteration highlighted the module M80260 (MIB score = 99.47) as an interesting candidate. This module contains smCOG10066, smCOG10118, and smCOG10495 and is found in 58 BGCs. This module targets three genes in three fully characterised modules: R1128 (*zhuF, D, E*), polyketomycin (*pokAC1, AC2, AC3*), and xantholipin (*xanB3, B1, B2*). This three-gene module appears to be encoding the biosynthesis of malonyl-CoA[14,42,43]. Surprisingly, this module has a rather high Shannon entropy (1.96), and not all of the BGCs are predicted to be involved in polyketide biosynthesis. For example, the module is found in clusters putatively responsible for the biosynthesis of non-ribosomal peptides, terpenes, ectoines, and bacteriocins.

Module M105700 (MIB score = 99.01, BGCs = 47) is the next module selected with this procedure. This module targets only one BGC present in the MIBiG database, which encodes the biosynthesis of the polyketide tetronasin[44]. The module contains three smCOGs: smCOG10139 (covering the tsn3 gene), smCOG10264 (*tsn1*/*tsn2*), and smCOG10631 (*tsn4*), which encode enzymes similar to the components of pyruvate dehydrogenase and related multi-enzyme complexes; whether these have a role in boosting the supply of acetyl-CoA precursors for specialised metabolite biosynthesis has not been demonstrated, but seems plausible.

The next module selected by our procedure is module M156303 (MIB score = 98.67, BGCs = 31). Interestingly, this novel three-smCOG module (smCOG10062, smCOG10123, and smCOG10687) is found in 15 clusters present in the MIBiG database: landomycin (*lanT, Q, S*)[45], granaticin (*gra-orf23, orf26, orf27*)[46], sch47554/sch47555 (*schS1, S3, S2*)[47], rubradirin (*rubN3, K, L, N4*)[48,49], spinosad (*spnN, R, Q, O*)[50], lactonamycin (*lct44, 45, 46*)[51], kijanimicin (*kijD10, D2, D7, D1*)[52], tetrocarcin A (ACB37729.1, ACB37732.1, ACB37737.1 and ACB37754.1)[53], polyketomycin (*pokS4, S5, S3*)[42,54], streptolydigin (*slgS4, S6, S3*) [55], pristinamycin (*cpp28, hpaA, cpp32*)[56], nocathiacin (*nocS4, S6, S5*)[34], BE-7585A (*bexQ, T, V*)[57], and amicetin (*amiD, N, C*)[58]. In all cases, this module appears to be involved in sugar modification/biosynthesis. The biochemical details of their function have not been fully elucidated in any of these cases, but the modularity analysis will facilitate a comparative approach to understanding their action. Module M156303 is just one example of a large number of sugar-related modules of high statistical support, highlighting the pervasive modularity of sugar decorations in specialised metabolite biosynthesis, which could be the target of a more detailed evolutionary and biochemical evaluation in the future.

Finally, our procedure selected module M134122 (MIB score = 97.33, BGCs = 20) for further discussion. This module, which is composed of smCOG10818, smCOG11294, smCOG12468, and smCOG12628, targets two BGCs present in the MIBiG

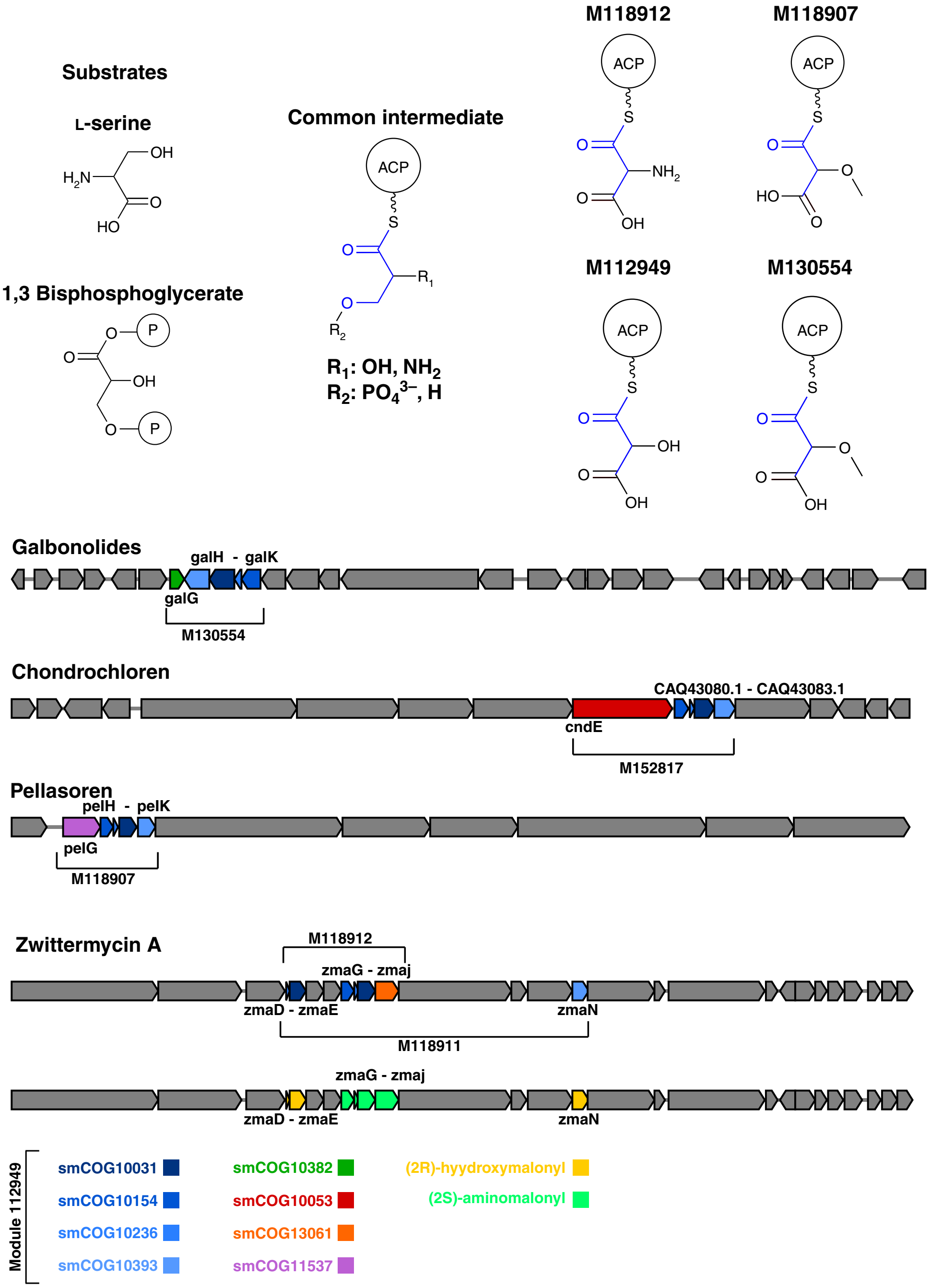

Fig. 3 Schematic representation of the chemical moieties produced and the hierarchical organisation of the ACP-linked PKS extender modules. L-Serine and 1,3-biphosphoglycerate are the two possible substrates accepted by the modules. The galbonolides cluster has been chosen as a representative example of the known clusters containing the methoxymalonyl-ACP module. It is also noteworthy that all modules share a common intermediate

database: kanamycin (BAE95427.1, BAE95426.1, BAE95429.1, and BAE95428.1)[59,60] and tallysomycin (*orf35–37*)[61]. The genes covered by module M134122 code for the subunits of nitrate reductase (*α*, *β*, *γ*, and *δ* chain) and have no known role in the biosynthesis of specialised metabolites. They represent a true (functional) module, but play their role in primary, rather than specialised metabolism. Nevertheless, the fact that this module is found with such high statistical support in the close neighbourhood of many BGCs might provide relevant information about the genetic context in which these BGCs are situated, and potentially the physiological context in which they are activated[62]. Importantly, this example also illustrates that our module detection method could be of more general value beyond specialised metabolism.

The first six modules prioritised using our iterative procedure are summarised in Supplementary Data 2. It should be emphasised that the biological roles suggested for these modules are purely based on plausibility arguments, but experimental validation will be required in each case to establish their precise functional relevance. Exploring the whole collection of putative functional biosynthetic modules detected by the method presented here will take a community effort and is beyond the aims of this work. The modules described above illustrate the power of the statistical detection and definition of putative biosynthetic modules, and the database provided will be a helpful resource for the whole community to explore further.

Discussion

In this work, we present a statistical method to automatically detect putative functional (and evolutionary) modules in BGCs. The fact that the method is unsupervised makes it powerful in associating genes that may have not been associated before by looking at individual gene clusters. For the future, a key next step will entail correcting for phylogenetic bias in the input data (i.e., having many similar gene clusters from closely related genomes), which currently can lead to the detection of artificial modules (although the Shannon entropy score (see Methods) can be used to down-rank such artefacts). This could be done either by performing additional redundancy filtering on the input data, or by correcting for phylogenetic structure in the statistical tests (which would of course increase the computational demands even more). Nonetheless, the obtained library of automatically predicted modules allows the efficient definition of BGCs that are likely to encode the biosynthesis of molecules with novel combinations of known chemical moieties. Moreover, strongly supported modules with currently unknown functions can be identified in our data, which potentially are responsible for the biosynthesis of discrete and novel chemical (sub)structures. In addition, the whole collection of automatically detected modules will help understand the degree of modularity in the organisation of microbial BGCs, and it will provide useful tools to be used in conjunction with other screening modalities for drug discovery by genome mining. For example, this work will also be a great asset for cluster prediction tools such as antiSMASH[29], as it can be used to annotate BGCs by predicting, which genes in a BGC function together as discrete units in a complex biosynthetic pathway, i.e., the enediynes (see Supplementary Note 1), and can also help in cluster boundary prediction. In the future, we intend to integrate the module library with public web services such as antiSMASH[29], antiSMASH database[63,64], and MIBiG[33]. In this context, it is noteworthy that the MIBiG database already allows users to upload chemically characterised subclusters.

Methods

Data acquisition. All the available bacterial and fungal genomic sequences were obtained from GenBank (access date 08 October 2012). We used antiSMASH version 1.0[65] to detect all BGCs in this set of genomes, resulting in a collection of 482,040 genes in 14,869 BGCs.

Cluster trimming. The boundaries of BGCs reported by antiSMASH[65] are expanded to include genes neighbouring the actual biosynthetic cluster in order to assure that the complete genomic entity is extracted. This greedy approach is a rational choice for molecular biology purposes, but would result in extra computational burden in our downstream analysis. To minimise this problem, we analysed the data set with ClusterFinder[4] to obtain the most probable cluster borders, based on each cluster's constituent PFAM domains. We shortlisted PFAM domains that are important for specialised metabolites (Supplementary Data 3) and trimmed genes from the extremes of each of the clusters if their probability of having a PFAM domain from the list was lower than 0.1. The threshold value was selected based on the observation that increasing the threshold up to 0.1 resulted in more genes being trimmed out, whereas any further increase (up to 0.5) had hardly any effect. Trimming removed 135,298 genes from the set. The trimmed data set consisted of 346,742 genes in 14,809 BGCs.

Clusters of orthologous genes. A set of smCOGs was constructed from all genes in the set of BGCs using the orthoMCL v 1.4 package[32] with standard settings. OrthoMCL analysis resulted in 19,292 smCOGs. From this set, we removed smCOGs having fewer than three genes, resulting in 12,756 smCOGs. We also removed 45,906 genes that did not belong to any of the remaining smCOGs. This left us with 211 BGCs that were empty (i.e., did not include any gene belonging to any smCOG), further narrowing the data set down to 300,612 genes in 14,598 BGCs. In order to reduce redundancy, if two or more clusters showed the same smCOG composition (regardless the order), we only kept the shortest one, narrowing the data set down to 12,842 non-redundant BGCs. Subsequently, we annotated the smCOGs based on the annotation of the genes they contain. For this purpose, we divided the annotations of individual genes into five major categories: core biosynthesis, regulator, tailoring, transport, and other. This taxonomy is described in more detail in Supplementary Data 4. Each smCOG was annotated: (1) if > 60% of genes from a smCOG share the same annotation category, this category was used a main annotation of the smCOG with exception of (2) if the most common category is other but the second most frequent one occurs in > 40% of genes, the second category was used; (3) if the most common category is present in < 60% of genes but first and second most common categories are together present in > 75% of the genes the smCOG was annotated with a double category (e.g., tailoring/core); (4) otherwise, the smCOG was annotated as mixed. Exact descriptions of all smCOG annotations are available in the Supplementary Data 5.

Interactions between smCOGs. A putative module is defined as a set of smCOGs of any size found in more BGCs than expected by chance. Considering all the 12,842 different smCOGs present in our data set, the number of possible modules is enormous, even if the size of a module is constrained to be between three and ten genes:

$$\sum_{n=3}^{10} \binom{12842}{n} \cong 3.35\times 10^{34} \tag{1}$$

Handling such an enormous number of potential modules represents a very complex computational challenge. To address this issue, we focused on the pairwise interactions between different smCOGs, i.e., the detection of smCOGs that co-occur surprisingly often within the same BGCs. In practice, we distinguished between two types of interactions: adjacency, when two smCOGs appear side-by-side in a BGC, and colocalization, when they are found together within the same cluster independently of their relative position. Adjacency and colocalization interactions were counted for each possible pair of smCOGs. Genes in the BGCs that did not belong to any of the smCOGs did not contribute to the number of interactions, but their positions were not skipped; i.e., a smCOG neighbouring two genes that did not belong to any smCOG would be counted as having no adjacency interactions in this cluster.

Assessment of statistical significance of interactions. Considering two smCOGs sharing N adjacency or colocalization interactions, one can assess the statistical significance of such interactions by computing the probability of observing at least N interactions between the two COGs, if they were randomly distributed among the clusters present in our data set. Although easy to state, this calculation is far from trivial computationally, and tackling it with a permutation-based approach would be too computationally demanding. However, for both kinds of interactions, it is relatively easy to compute the probability of observing at least N interactions, keeping fixed the positions of one of the two smCOGs in our data set. Using this approach, it is possible to obtain two (different) p-values per each pair of smCOGs, depending on which of them is considered as fixed. Aiming for the most conservative approach, we considered only the larger of the two p-values when determining the statistical significance of an smCOG interaction in the subsequent analysis.

Adjacency interactions. Consider a pair of smCOGs (called COG A and COG B) that occur adjacently at least once. If we keep all genes annotated as belonging to COG A in their original positions, one could divide all the remaining positions in the BGCs into three classes: (a) positions that are not adjacent to any COG A gene; (b) positions adjacent to one COG A gene; and (c) positions adjacent to two COG A genes. N_a, N_b, and N_c represent the number of available positions in each of these three classes. The total number of adjacency interactions between the two COGs is:

$$i_{orig} = B_b + 2 \cdot B_c \tag{2}$$

where B_b represent the number of COG B genes occupying a position adjacent with one COG A genes and B_c represent the number of COG B gene occupying a position adjacent with two COG A genes. The number of COG B genes occupying a position not adjacent to any COG A gene is indicated as B_a. It should be noticed that the same number of interactions can be observed with more than one distribution of the COG B. The probability of observing one specific distribution d (i.e., a specific set of values for B_a, B_b, and B_c) when the COG A genes are fixed and the COG B genes are randomly distributed among all the available positions is expressed by the following hypergeometric equation:

$$P_D = \frac{\binom{N_a}{B_b}\binom{N_b}{B_b}\binom{N_c}{B_c}}{\binom{N_{tot}}{B_{tot}}} \tag{3}$$

where B_{tot} and N_{tot} represent the total number of COG B genes present in the data set and the total number of available positions. The probability of observing at least

i_{orig} adjacency interactions (i.e., the p-value) can be easily computed as:

$$p = P_{i \geq i_{orig}} = 1 - P_{i < i_{orig}} = 1 - \sum_{\mathbf{D}} P_d \tag{4}$$

where $\mathbf{D}$ represent the set of all the possible combinations of B_a, B_b, and B_c, leading to a number of interactions lower than the observed one.

Colocalization interactions. The calculations for the colocalization interactions are analogous to those described for the adjacency interactions. However, the number of gene classes is not limited to three, but to the maximum number of occurrences of COG A genes in a single cluster (n_{max}):

$$P_D = \frac{\binom{N_a}{B_a}\binom{N_b}{B_b}\cdots\binom{N_{n_{max}}}{B_{n_{max}}}}{\binom{N_{tot}}{B_{tot}}} \tag{5}$$

This creates a computational problem whenever n_{max} is large. In order to avoid this problem, we removed such redundant smCOGs for the colocalization calculations: when genes from the same smCOG are found more than once in a cluster, they are substituted by an empty position and the genes from the affected smCOG are attached at the end of the cluster separated by an empty position from the rest of the cluster. Althoughthis approach deletes some adjacency interactions and slightly changes the topology of some clusters, it leads to conservative p-value estimates and, most importantly, makes the p-value computations easier, as equation 5 simplifies to:

$$P_D = \frac{\binom{N_a}{B_a}\binom{N_b}{B_b}}{\binom{N_{tot}}{N_{tot}}} \tag{6}$$

All the p-values computed for both kinds of interactions were corrected for multiple testing using the Benjamini–Yekutieli method[66] for controlling the false-discovery rate under dependency.

Module detection. The obtained multiple-testing corrected p-values were used for detecting putative modules. By selecting an initial arbitrary p-value threshold for significant interactions, it is possible to compute a binary matrix M of dimension ($C \times C$), where C is the total number of smCOGs present in our data set, and the $m_{i,j}$ element of the matrix is equal to 1 if either the adjacency or the colocalization p-value is lower or equal to the chosen threshold. This matrix represents an undirected graph, where two smCOGs (nodes) are connected by an edge if they share a statistically significant number of adjacency or colocalization interactions. All the maximal cliques found in this graph and containing at least three elements are detected and added to our list of putative modules. A maximal clique is a fully connected sub-graph, where connections are based on either significant adjacency or significant colocalization, which is not a subset of any other fully connected sub-graph. All p-values occurring in the data set smaller or equal to 0.1 are iteratively considered as the arbitrary p-value threshold. The graph analysis and the maximal clique detection was performed using the *igraph* package[67]. Using this approach, we ended up with a total of 197,564 putative modules. It is important to remember that each three-member module is supported by three individually significant interactions, and that depending on the intended use case, stricter p-value thresholds can easily be applied to reduce the number of modules for further analysis. Although a rigorous estimation of the false discovery rate associated with our module detection method is not provided, we considered all the modules found together in less than two BGCs as false positives. Such modules represent ~ 6% of all the detected modules and they have been removed from our database.

Modules prioritisation and trimming. A number of different metrics were computed for each detected module in order to prioritise and filter them according to user-defined criteria:

- Number of BGCs containing the module. As the modules are defined on the basis of individual pairwise interactions between smCOGs, it is possible that all pairwise interactions between members of a module are significant, even when the complete module never occurs together in the same BGC. To remove such spurious modules, we removed ~ 6% of the detected modules that were found together in less than two of the BGCs, resulting on a total of 185,718 modules.
- Number of BGCs containing the module and present in the MIBiG data set. The MIBiG data set of experimentally characterised BGCs[33] was queried in order to identify which of the BGCs present in our data set have been associated with experimental data. For these clusters, the enzymatic pathway is at least partially defined and the chemical structures of the end compound known.
- Module size. The size of the detected modules (i.e., the number of smCOGs) ranges from 3 (minimum value allowed by the module detection algorithm) to a maximum of 42 (the largest maximal clique using the most lenient interaction threshold). In total, 80% of modules are smaller than 16 smCOGs, the median size is 9, and the most likely value is 3. Very large modules are in fact typically entire BGCs, rather than biosynthetic modules (subclusters), and are therefore usually not of interest for the subsequent analysis (although they obviously are modules in the sense of being coherent evolutionary entities, the detection of which confirms the validity of the module detection algorithm).
- Strictest p-value threshold. While detecting the modules, a different p-value threshold is chosen at each iteration. For each module, it is possible to identify the strictest p-value threshold that can be used to detect it. This value can be considered as a measure of the overall statistical significance of the module.
- Number of different compound classes and their Shannon entropy. While looking for putative BGCs, antiSMASH is also predicting the chemical class of the end compound. Using this information, we can focus specifically on modules that occur in clusters responsible for the most diverse set of predicted compound classes—these are the most likely to be responsible for carrying out well-defined chemical functions and to act as independent evolutionary units. To accurately estimate the diversity of compound classes covered by the BGCs containing a specific module, we used Shannon's informational Entropy (SE), which is computed as follows:

$$\sum_{i=1}^{I} f_i \cdot \log(f_i) \tag{7}$$

where f_i is the ratio between the number of times the module is involved in the biosynthesis of the i^{th} compound class and the total number of BGCs containing the module, and I is the total number of different putative end compound classes produced by these BGCs. The higher the SE, the larger is the number of targeted compound classes, and the lower the bias toward one or more compound classes.

- Percentage of smCOGs members of a specific category. As mentioned above, all smCOGs were annotated according to functional categories (e.g., transport, tailoring, core biosynthesis etc.). For each module, we computed the percentage of each functional category. Using this information, we can, for example, focus on modules composed only or mostly of tailoring smCOGs (again, these are most likely to represent evolutionary units of interest).

For the overall prioritisation, we used a weighted combination of all of these metrics, the MIB score. This score is simply computed as a weighted sum of all the ranks obtained considering each metric individually. Subsequently, the MIB scores obtained are rescaled over the range from 1 (least interesting) to 100 (most interesting). Depending on the user particular interest, the weights can be adjusted to increase the contribution of individual criteria to the overall prioritisation. The default values of the weights used in the subsequent discussion are: 2 for the length of the module (longer, and thus more specific, being better), 15 for the Shannon entropy, 10 for the number of BGCs containing the module, 5 for the maximum p-value threshold, 10 for the percentage of tailoring smCOGs, and 0 for all remaining criteria.

Code availability. The code used for the p-value evaluation, module detection, and metrics computation was written in R[68] with the use of the *igraph*[67] and *Rmpfr*[69] packages, and it is available at https://github.com/francescodc87/Modules_Detection together with a detailed documentation.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The complete data set is available at https://github.com/francescodc87/Modules-explorer together with a *Shiny*-based[70] web application, which provides users with a simple graphical interface to explore the data set containing all the detected modules. A detailed documentation is present the github pages. In addition, all the supplementary material mentioned in the manuscript can be also found at https://github.com/francescodc87/Modules_Detection/tree/master/Supplemetary_Files.

Received: 13 November 2018 Accepted: 22 January 2019
Published online: 28 February 2019

References

1. Smanski, M. J. et al. Synthetic biology to access and expand nature's chemical diversity. *Nat. Rev. Microbiol.* **14**, 135–149 (2016).

2. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
3. Laureti, L. et al. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proc. Natl. Acad. Sci.* **108**, 6258–6263 (2011).
4. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
5. Fischbach, M. A., Walsh, C. T. & Clardy, J. The evolution of gene collectives: how natural selection drives chemical innovation. *Proc. Natl. Acad. Sci.* **105**, 4601–4608 (2008).
6. Karki, S. et al. The methoxymalonyl-acyl carrier protein biosynthesis locus and the nearby gene with the β-ketoacyl synthase domain are involved in the biosynthesis of galbonolides in *Streptomyces galbus*, but these loci are separate from the modular polyketide synthase gene cluster. *FEMS Microbiol. Lett.* **310**, 69–75 (2010).
7. Li, W., Ju, J., Osada, H. & Shen, B. Utilization of the methoxymalonyl-acyl carrier protein biosynthesis locus for cloning of the tautomycin biosynthetic gene cluster from *Streptomyces spiroverticillatus*. *J. Bacteriol.* **188**, 4148–4152 (2006).
8. Li, W., Ju, J., Rajski, S. R., Osada, H. & Shen, B. Characterization of the tautomycin biosynthetic gene cluster from *Streptomyces spiroverticillatus* unveiling new insights into dialkylmaleic anhydride and polyketide biosynthesis. *J. Biol. Chem.* **283**, 28607–28617 (2008).
9. Zhao, C. et al. Oxazolomycin biosynthesis in *Streptomyces albus* JA3453 featuring an "acyltransferase-less" type I polyketide synthase that incorporates two distinct extender units. *J. Biol. Chem.* **285**, 20097–20108 (2010).
10. Zhao, C. et al. Utilization of the methoxymalonyl-acyl carrier protein biosynthesis locus for cloning the oxazolomycin biosynthetic gene cluster from *Streptomyces albus* JA3453. *J. Bacteriol.* **188**, 4142–4147 (2006).
11. Wu, K., Chung, L., Revill, W. P., Katz, L. & Reeves, C. D. The FK520 gene cluster of *Streptomyces hygroscopicus* var. *ascomyceticus* (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. *Gene* **251**, 81–90 (2000).
12. Mo, S. et al. Biosynthesis of the allylmalonyl-CoA extender unit for the FK506 polyketide synthase proceeds through a dedicated polyketide synthase and facilitates the mutasynthesis of analogues. *J. Am. Chem. Soc.* **133**, 976–985 (2010).
13. Kang, Q., Shen, Y. & Bai, L. Biosynthesis of 3, 5-AHBA-derived natural products. *Nat. Prod. Rep.* **29**, 243–263 (2012).
14. Zhang, M.-Q. et al. Optimizing natural products by biosynthetic engineering: discovery of nonquinone Hsp90 inhibitors. *J. Med. Chem.* **51**, 5494–5497 (2008).
15. Yu, T.-W. et al. The biosynthetic gene cluster of the maytansinoid antitumor agent ansamitocin from *Actinosynnema pretiosum*. *Proc. Natl. Acad. Sci.* **99**, 7968–7973 (2002).
16. Rascher, A., Hu, Z., Buchanan, G. O., Reid, R. & Hutchinson, C. R. Insights into the biosynthesis of the benzoquinone ansamycins geldanamycin and herbimycin, obtained by gene sequencing and disruption. *Appl. Environ. Microbiol.* **71**, 4862–4871 (2005).
17. Haydock, S. F. et al. Organization of the biosynthetic gene cluster for the macrolide concanamycin A in *Streptomyces neyagawaensis* ATCC 27449. *Microbiology* **151**, 3161–3169 (2005).
18. Li, Z. et al. Complete elucidation of the late steps of bafilomycin biosynthesis in *Streptomyces lohii*. *J. Biol. Chem.* **292**, 7095–7104 (2017).
19. Du, Y. et al. Biosynthesis of the apoptolidins in *Nocardiopsis* sp. FU 40. *Tetrahedron* **67**, 6568–6575 (2011).
20. Pojer, F., Li, S.-M. & Heide, L. Molecular cloning and sequence analysis of the clorobiocin biosynthetic gene cluster: new insights into the biosynthesis of aminocoumarin antibiotics. *Microbiology* **148**, 3901–3911 (2002).
21. Trefzer, A. et al. Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. *Antimicrob. Agents Chemother.* **46**, 1174–1182 (2002).
22. Williamson, N. R. et al. Biosynthesis of the red antibiotic, prodigiosin, in *Serratia*: identification of a novel 2-methyl-3-n-amyl-pyrrole (MAP) assembly pathway, definition of the terminal condensing enzyme, and implications for undecylprodigiosin biosynthesis in *Streptomyces*. *Mol. Microbiol.* **56**, 971–989 (2005).
23. Wu, Q. et al. Characterization of the biosynthesis gene cluster for the pyrrole polyether antibiotic calcimycin (A23187) in *Streptomyces chartreusis* NRRL 3882. *Antimicrob. Agents Chemother.* **55**, 974–982 (2011).
24. Li, C., Roege, K. E. & Kelly, W. L. Analysis of the indanomycin biosynthetic gene cluster from *Streptomyces antibioticus* NRRL 8167. *Chembiochem* **10**, 1064–1072 (2009).
25. Nowak-Thompson, B., Chaney, N., Wing, J. S., Gould, S. J. & Loper, J. E. Characterization of the pyoluteorin biosynthetic gene cluster of *Pseudomonas fluorescens* Pf-5. *J. Bacteriol.* **181**, 2166–2174 (1999).
26. Medema, M. H., Breitling, R., Bovenberg, R. & Takano, E. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat. Rev. Microbiol.* **9**, 131–137 (2011).
27. Hadjithomas, M. et al. IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. *Nucleic Acids Res.* **45**, D560–D565 (2017).
28. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
29. Blin, K. et al. antiSMASH 4.0 – improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
30. van Heel, A. J., de Jong, A., Montalban-Lopez, M., Kok, J. & Kuipers, O. P. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. *Nucleic Acids Res.* **41**, W448–W453 (2013).
31. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
32. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
33. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
34. Ding, Y. et al. Moving posttranslational modifications forward to biosynthesize the glycosylated thiopeptide nocathiacin I in *Nocardia* sp. ATCC202099. *Mol. Biosyst.* **6**, 1180–1185 (2010).
35. Kevany, B. M., Rasko, D. A. & Thomas, M. G. Characterization of the complete zwittermicin A biosynthesis gene cluster from *Bacillus cereus*. *Appl. Environ. Microbiol.* **75**, 1144–1155 (2009).
36. Rachid, S., Scharfe, M., Blöcker, H., Weissman, K. J. & Müller, R. Unusual chemistry in the biosynthesis of the antibiotic chondrochlorens. *Chem. Biol.* **16**, 70–81 (2009).
37. Jahns, C. et al. Pellasoren: structure elucidation, biosynthesis, and total synthesis of a cytotoxic secondary metabolite from *Sorangium cellulosum*. *Angew. Chem. Int. Ed.* **51**, 5239–5243 (2012).
38. Chan, Y. A. & Thomas, M. G. Recognition of (2S)-aminomalonyl-acyl carrier protein (ACP) and (2R)-hydroxymalonyl-ACP by acyltransferases in zwittermicin A biosynthesis. *Biochemistry* **49**, 3667–3677 (2010).
39. Liu, W.-T. et al. MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in *Streptomyces roseosporus*. *J. Antibiot. (Tokyo)* **67**, 99–104 (2014).
40. Cai, W. et al. The biosynthesis of capuramycin-type antibiotics identification of the A-102395 biosynthetic gene cluster, mechanism of self-resistence, and formation of uridine-5′-carboxamide. *J. Biol. Chem.* **290**, 13710–13724 (2015).
41. Funabashi, M. et al. Identification of the biosynthetic gene cluster of A-500359s in *Streptomyces griseus* SANK60196. *J. Antibiot. (Tokyo)* **62**, 325–332 (2009).
42. Daum, M. et al. Organisation of the biosynthetic gene cluster and tailoring enzymes in the biosynthesis of the tetracyclic quinone glycoside antibiotic polyketomycin. *Chembiochem* **10**, 1073–1083 (2009).
43. Marti, T., Hu, Z., Pohl, N. L., Shah, A. N. & Khosla, C. Cloning, nucleotide sequence, and heterologous expression of the biosynthetic gene cluster for R1128, a non-steroidal estrogen receptor antagonist insights into an unusual priming mechanism. *J. Biol. Chem.* **275**, 33443–33448 (2000).
44. Cooper, H. N., Cortes, J., Bevitt, D. J., Leadlay, P. F. & Staunton, J. Analysis of a gene cluster from *S. longisporoflavus* potentially involved in tetronasin biosynthesis. *Biochem. Soc. Trans.* **21**, 31S (1993).
45. Westrich, L. et al. Cloning and characterization of a gene cluster from *Streptomyces cyanogenus* S136 probably involved in landomycin biosynthesis. *FEMS Microbiol. Lett.* **170**, 381–387 (1999).
46. Ichinose, K. et al. The granaticin biosynthetic gene cluster of *Streptomyces violaceoruber* Tü22: sequence analysis and expression in a heterologous host. *Chem. Biol.* **5**, 647–659 (1998).
47. Basnet, D. B. et al. Angucyclines Sch 47554 and Sch 47555 from *Streptomyces* sp. SCC-2136: cloning, sequencing, and characterization. *Mol. Cells* **22**, 154–162 (2006).
48. Kim, C.-G. et al. Biosynthesis of rubradirin as an ansamycin antibiotic from *Streptomyces achromogenes* var. rubradiris NRRL3061. *Arch. Microbiol.* **189**, 463–473 (2008).
49. Sohng, J., Oh, T., Lee, J. & Kim, C. Identification of a gene cluster of biosynthetic genes of rubradirin substructures in *S. achromogenes* var. rubradiris NRRL3061. *Mol. Cells* **7**, 674–681 (1997).
50. Waldron, C. et al. Cloning and analysis of the spinosad biosynthetic gene cluster of *Saccharopolyspora spinosa*. *Chem. Biol.* **8**, 487–499 (2001).
51. Zhang, X., Alemany, L. B., Fiedler, H.-P., Goodfellow, M. & Parry, R. J. Biosynthetic investigations of lactonamycin and lactonamycin Z: cloning of the biosynthetic gene clusters and discovery of an unusual starter unit. *Antimicrob. Agents Chemother.* **52**, 574–585 (2008).
52. Zhang, H. et al. Elucidation of the kijanimicin gene cluster: insights into the biosynthesis of spirotetronate antibiotics and nitrosugars. *J. Am. Chem. Soc.* **129**, 14670–14683 (2007).
53. Fang, J. et al. Cloning and characterization of the tetrocarcin A gene cluster from *Micromonospora chalcea* NRRL 11289 reveals a highly conserved

strategy for tetronate biosynthesis in spirotetronate antibiotics. *J. Bacteriol.* **190**, 6014–6025 (2008).
54. Paululat, T., Zeeck, A., Gutterer, J. M. & Fielder, H.-P. Biosynthesis of polyketomycin produced by *Streptomyces diastatochromogenes* Tü 6028. *J. Antibiot. (Tokyo)* **52**, 96–101 (1999).
55. Gómez, C., Horna, D. H., Olano, C., Méndez, C. & Salas, J. A. Participation of putative glycoside hydrolases SlgC1 and SlgC2 in the biosynthesis of streptolydigin in *Streptomyces lydicus*. *Microbial. Biotechnology* **5**, 663–667 (2012).
56. Mast, Y. et al. Characterization of the "pristinamycin supercluster" of *Streptomyces pristinaespiralis*. *Microb. Biotechnol.* **4**, 192–206 (2011).
57. Sasaki, E., Ogasawara, Y. & Liu, H.-w A biosynthetic pathway for BE-7585A, a 2-thiosugar-containing angucycline-type natural product. *J. Am. Chem. Soc.* **132**, 7405–7417 (2010).
58. Zhang, G. et al. Characterization of the amicetin biosynthesis gene cluster from *Streptomyces vinaceusdrappus* NRRL 2363 implicates two alternative strategies for amide bond formation. *Appl. Environ. Microbiol.* **78**, 2393–2401 (2012).
59. Yanai, K. & Murakami, T. The kanamycin biosynthetic gene cluster from *Streptomyces kanamyceticus*. *J. Antibiot. (Tokyo)* **57**, 351–354 (2004).
60. Yanai, K., Murakami, T. & Bibb, M. Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of *Streptomyces kanamyceticus*. *Proc. Natl. Acad. Sci.* **103**, 9661–9666 (2006).
61. Tao, M. et al. The tallysomycin biosynthetic gene cluster from *Streptoalloteichus hindustanus* E465-94 ATCC 31158 unveiling new insights into the biosynthesis of the bleomycin family of antitumor antibiotics. *Mol. Biosyst.* **3**, 60–74 (2007).
62. Sawers, R., Falke, D. & Fischer, M. Chapter one–oxygen and nitrate respiration in *Streptomyces coelicolor* A3 (2). *Adv. Microb. Physiol.* **68**, 1–40 (2016).
63. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45**, D555–D559 (2016).
64. Blin, K. et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **47**, D625–D630 (2018).
65. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
66. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**,1165–1188 (2001).
67. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter. Complex Syst.* **1695**, 1–9 (2006).
68. R. Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (2015).
69. Maechler, M. Rmpfr: R mpfr-multiple precision floating-point reliable. *R package version 0.6-0*, http://rmpfr.r-forge.r-project.org (2015).
70. Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. shiny: Web application framework for R [Computer software]. http://CRAN.R-project.org/package=shiny (R package version 1.0. 0) (2017).

Acknowledgements

M.C. was funded by a BBSRC DTP studentship (Grant No. BB/J014478/1). This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM) and acknowledges the Biotechnology and Biological Sciences Research Council (BBSRC) and Engineering and Physical Sciences Research Council (EPSRC) for financial support (Grant No. BB/M017702/1). This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720793 TOPCAPI—thoroughly Optimised Production Chassis for Advanced Pharmaceutical Ingredients.

Author contributions

F.D.C. developed and designed the module detection algorithm, implemented the module detection method, and the module explorer, analysed, and manually curated the detected modules and wrote the manuscript. K.Z. performed the clusters prediction and the annotation of the cluster of orthologous genes annotation, and contributed to the writing of the manuscript. M.C. provided useful feedbacks and contributed in the analysis and manual curation of the modules. E.T. provides useful feedbacks on the biosynthetic role of the detected modules. M.M. provided useful feedbacks, and supervised the work of K.Z. R.B. designed the methods, supervised the project, contributed in the manual curation of the modules and wrote the manuscript. All authors contributed to the revision of the manuscript.

Additional information

Supplementary information accompanies this paper at https://doi.org/10.1038/s42003-019-0333-6.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.