Exploring the chemical space of post-translationally modified peptides in Streptomyces with machine learning

Kloosterman, A.M.

5

Characterization of a novel RiPP BGC identified in *Streptomyces* sp. MBT27

Alexander M. Kloosterman
Somayah S. Elsayed
Jasper van der Peet
Chao Du
Marnix H. Medema
Gilles P. van Wezel

Abstract

Actinobacteria are the most prolific producers of bioactive molecules. Their biosynthetic arsenal includes some two thirds of the clinical antibiotics and many other compounds of clinical and agricultural importance. The increased numbers of available sequenced genomes have revealed an enormous reservoir of biosynthetic gene clusters (BGCs). Mining of genomes for truly novel families of BGCs, however, requires a different approach. Here, we report the discovery of a candidate RiPP BGC, called *trc*, in *Streptomyces* sp. MBT27. The *trc* gene cluster was identified using our machine learning-based pipeline decRiPPter and encodes two candidate precursors containing a repeated TTGWQ-motif, as well as a radical SAM enzyme and a PGM1-like ATP-grasp ligase, which have been previously associated with RiPP biosynthesis. Constitutive expression of a *luxR*-like regulatory gene located within the BGC resulted in strongly increased expression of the *trc* gene cluster. Comparative LC-MS analysis of culture extracts revealed 113 mass features that were produced by strains expressing the *trc* gene cluster but were not detected in extracts of a *trc* null mutant. Grouping these mass features with GNPS networking revealed two major networks containing 73 of these mass features, suggesting they are derived from similar compounds. Taken together, our data support that the *trc* gene cluster specifies a range of small RiPPs, likely derived from a TTGWQ-motif present in all predicted precursor peptides. Further research is required to unveil how these compounds were modified, their biological role and possible application.

Introduction

Natural products are compounds synthesized by bacteria and fungi, which have widespread clinical applications [222]. Actinobacteria are filamentous bacteria that live in both soil and aquatic environments, and are the most prolific producers of bioactive molecules with clinical and biotechnological application. These include antibiotics and compounds with anticancer, antifungal, immunosuppressant or herbicidal activity [32, 158]. The majority of these natural products are produced by members of the genus *Streptomyces*. The enzymes specifying these natural products are encoded by clusters of genes, organized in one or more operons, which are referred to as biosynthetic gene clusters (BGCs). Due to chemical redundancy, the return of investment of high-throughput screening is decreasing rapidly [25, 223]. Still, it is expected that we have only scratched the surface of the chemical space of natural products [224]. Genome sequencing has uncovered that even the best-studied model actinomycetes possess many yet underexplored resources for natural products [27, 225, 226]. Most natural products discovered belong to previously characterized classes rather than new classes [227]. The key question that scientists need to answer now is, can we find truly novel classes of natural products that have hitherto been overlooked? These molecules are likely products of so-called cryptic BGCs that are poorly expressed under routine laboratory conditions, but require specific molecular signals or intensive genetic manipulation [30, 33].

The challenge of identifying novel BGCs poses an interesting conundrum: how can novel BGCs be detected without prior knowledge of specific genetic elements, while retaining a high detection accuracy? Many genome mining efforts aimed at BGC detection target one or more core enzymes required for the biosynthesis of these classes. These have proven highly effective in the detection of BGCs for natural products that have been well-characterized and contain highly conserved genes, such as those encoding nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) [36, 228]. When such conserved genetic markers are missing, a more innovative approach is required. Ribosomally synthesized and post-translationally modified peptides (RiPPs) form a class of natural products where this is often the case.

While RiPPs all share the same generic biosynthetic procedure, in which a precursor peptide is modified and cleaved to form a natural product, the modifications and responsible enzymes vary enormously. This makes it impossible to design a single genome mining strategy for the detection of all RiPP BGCs. Nevertheless, the large diversity covered by this class of natural products makes it an excellent candidate for the discovery of novel biosynthetic pathways.

Recently, we reported on a novel pipeline for the detection of novel RiPP BGCs, called decRiPPter (Chapter 3). This pipeline combines Support Vector Machine (SVM) models to detect candidate genes encoding RiPP precursors, with a pan-genomic analysis to prioritize novel candidate RiPP BGCs. A thorough analysis of 1,295 *Streptomyces* genomes resulted in the identification of 42 novel candidate RiPP families. While these candidate BGCs were not detected by conventional RiPP genome mining methods, some of them contained genes found among many different RiPP subclasses, such as genes encoding radical S-adenosyl methionine (SAM) utilizing enzymes, YcaO enzymes or ATP-grasp enzymes. Some of these genes have been used previously as 'bait' queries to identify novel RiPPs, efforts which have led to the discovery of the spliceotides [103], the WGK RiPPs [104], and the thiovarsiolins [52]. The presence of such RiPP-associated genes inside a gene cluster is no guarantee for the discovery of novel RiPPs, however. For example, a search for homologs of *pgm1*, a gene encoding an ATP-grasp ligase involved in the biosynthesis of pheganomycin, led to the discovery of the ketomemicins, which are non-RiPP natural products [163]. This example nevertheless illustrates that characterizing BGCs that show some relation to known BGCs may be a fruitful approach leading to the discovery of novel types of natural products.

Here, we describe the characterization of a gene cluster from *Streptomyces* sp. MBT27, which was detected by decRiPPter as a candidate RiPP BGC. The gene cluster contains several interesting RiPP markers, including a *pgm1* homolog, as well as a gene encoding a radical SAM enzyme, of which we study their relation to homologs from known BGCs. In addition, two closely related predicted precursors are encoded, which contain highly conserved TTGWQ-repeats. Five other gene clusters with similar features are discovered,

which form the candidate RiPP family. We studied the expression of the gene cluster by quantitative proteomics, and showed that it is naturally expressed under laboratory conditions. Comparative LC-MS analysis reveals that when the gene cluster is expressed at a higher level, several masses within the mass range 400 – 600 Da are detected at significantly higher levels, providing interesting leads for further research and characterization of any peptides that might be produced.

Results and Discussion

Bioinformatic characterization of the *trc* gene cluster

Given decRiPPter's capabilities of detecting BGCs of new RiPP subclasses, we aimed to find additional RiPP BGCs encoded by the analysed Streptomyces species. The large number of candidate RiPP BGCs detected by decRiPPter allows for many possibilities to filter them for candidates of interest, using numbers of encoded enzymes, transporters and regulators. Many candidate RiPP BGCs were identified by decRiPPter using a "strict filter". This filter requires the presence of two enzyme-coding genes, one gene encoding a transporter, one encoding a peptidase, and one encoding a regulator. In addition, the "Clusters of Orthologous Genes" (COG) score (which signifies the fraction of genomes within a taxon that contain a homolog of a given gene) of all genes in the gene cluster should be no higher than 0.1 on average. In other words, no more than 10 percent of the genomes analysed should contain an orthologue of any of the genes in the (core operon of the) gene cluster (Chapter 3). While using this strict filter increased the saturation of known RiPP BGCs among the results, 93% of known RiPP BGCs identified by antiSMASH [39] were filtered out in the process, which suggests that many more unknown candidate RiPP BGCs were filtered out as well. To investigate this, we examined the BGCs mined from 1,295 *Streptomyces* genomes, also considering BGCs that passed the mild filter (two encoded enzymes, one encoded transporter, average COG score ≤ 0.25). From these, we selected a promising candidate RiPP BGC that encodes a unique combination of enzymes, and was discovered in *Streptomyces* sp. MBT27 (from now on referred to as 'MBT27') [229]. The gene cluster consists of two putative operons in an opposing strand orientation, each starting with a predicted precursor gene. The shared sequence of the putative precursors is completely identical, with one precursor being 14 amino acids longer than the other. Interestingly, the precursors contained 3 and 4 repeats of a TTGWQ sequence, respectively, lending the cluster its preliminary name *trc* (TTGWQ-Repeat Containing RiPP candidate).

Enzymes encoded by the *trc* cluster include an ATP-grasp ligase (TrcC), a radical SAM protein (TrcD), oxidoreductases (TrcF, TrcG, TrcH3) and two aminotransferases (TrcJ, TrcK) (Figure 1). Directly adjacent to the operons lies a

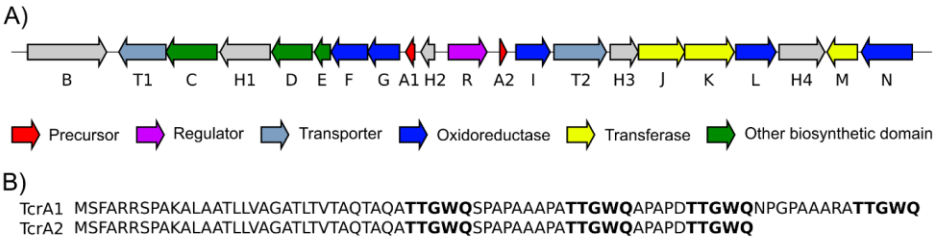

Figure 1. The *trc* gene cluster from *Streptomyces* sp. MBT27. A) The *trc* cluster consists of clusters of genes that likely form separate operons, each preceded by a putative RiPP precursor. B) Both predicted precursor proteins are highly similar to one another, and contain multiple repeats of a TTGWQ-motif. Further details on the annotation can be found in Table 1.

gene encoding a tryptophan halogenase (TrcN), although it is unclear whether this gene is part of the *trc* cluster. In addition, genes encoding a transporter (TrcT) and a regulator (TrcR) were found, which could have a cluster-specific role. We also searched for RiPP Recognition Elements (RREs), which facilitate precursor peptide binding in a wide variety of RiPPs [109], using RREFinder (Chapter 4). No RREs were found using either the conservative "precision mode" or the less restricted "exploratory mode". If this gene cluster indeed encodes proteins that produce a RiPP, precursor peptide recognition must be independent of an RRE.

To study the distribution of the *trc* cluster across bacteria, we employed MultiGeneBLAST [230] using all the genes in Figure 1 as a query. Five orthologous clusters of genes were found among the 1,295 streptomycetes studied (Figure S1). All of these contained both putative operons, showing little variation with respect to gene conservation and synteny. *trcN*, the gene encoding a tryptophan halogenase, was also well conserved, suggesting it was indeed part of the gene cluster. Outside of *Streptomyces*, only distantly related gene clusters were found in Actinobacteria, which share up to four different genes in a different genetic context (Figure S2). No gene clusters with significant similarity were identified outside of Actinobacteria, or among characterized BGCs from the MIBiG database [29].

Table 1. Annotation of the *trc* gene cluster.

Gene name	NCBI Accession	NCBI product annotation	Best Pfam hit	Pfam description
TrcB	WP_167162477.1	Threonine—tRNA ligase	PF00587	tRNA synthetase class II core domain (G, H, P, S and T)
TrcT1	WP_167162479.1	MFS transporter	PF07690	Major Facilitator Superfamily
TrcC	WP_167162481.1	Hypothetical protein	PF18604	Pre ATP-grasp domain
TrcH1	WP_167162483.1	Hypothetical protein		
TrcD	WP_167162485.1	Radical SAM protein	PF04055	Radical SAM superfamily
TrcE	WP_167162487.1	GFA family protein	PF04828	Glutathione-dependent formaldehyde-activating enzyme
TrcF	WP_167162489.1	Omega-3 fatty acid desaturase	PF00487	Fatty acid desaturase
TrcG	WP_167162880.1	2OG-Fe -dioxygenase family protein	PF10014	2OG-Fe dioxygenase
TrcA1	WP_167162491.1	Hypothetical protein		
TrcH2	WP_167162492.1	Hypothetical protein		
TrcR	WP_167162494.1	Helix-turn-helix transcriptional regulator		
TrcA2	WP_167162496.1	Hypothetical protein		
TrcI	WP_167162498.1	Phytanoyl-CoA dioxygenase family protein	PF05721	Phytanoyl-CoA dioxygenase (PhyH)
TrcT2	WP_167162500.1	Cation:proton antiporter	PF00999	Sodium/hydrogen exchanger family
TrcH3	WP_167162502.1	Hypothetical protein	PF00970	Oxidoreductase FAD-binding domain
TrcJ	WP_167162504.1	Aminotransferase class I/II-fold pyridoxal	PF00155	Aminotransferase class I and II
TrcK	WP_167162506.1	Aminotransferase class I/II-fold pyridoxal	PF01053	Cys/Met metabolism PLP-dependent enzyme
TrcL	WP_167162508.1	Hypothetical protein	PF02566	OsmC-like protein
TrcH4	WP_167162509.1	Hypothetical protein		
TrcM	WP_167162511.1	Class I SAM-dependent methyltransferase	PF13649	Methyltransferase domain
TrcN	WP_167162513.1	Tryptophan 7-halogenase	PF04820	Tryptophan halogenase

Predicted precursors encoded by the orthologous gene clusters are well conserved, particularly the N-terminal 31 aa (Figure S3). The C-terminal part of the peptides showed more variation, although all of them contain between two and four repeats of the TTGWQ sequence, a motif with unknown function. Possibly, the TTGWQ sequences form the core peptides, which are then processed to from the final product. This efficient usage of a precursor peptide, in which only a single leader peptide needs to be synthesized for multiple copies of the final product, has been reported for several other RiPPs, including cyanobactins [231, 232], orbitides [233], cyclotides [234], microviridins [235] and other omega-ester containing peptides (OEPs) [84], dikaritins [236-238], type II borosins [239], lyciumins [51] and pheganomycin [162].

Since decRiPPter also identifies putative precursors within non-RiPP BGCs, we looked for further evidence whether or not we could associate the *trc* cluster to the RiPP family of natural products. The gene *trcD* was predicted to encode a radical S-adenosyl methionine (radical SAM) enzyme. Radical SAMs typically share a conserved CxxxCxxC motif, containing a redox-active [4Fe-4S]-cluster binding an S-adenosyl methionine (SAM). These enzymes are highly divergent: a recent review grouped known examples of radical SAMs into 20 different families, which were further divided in almost 100 different subgroups [100]. Radical SAMs are encoded by the BGCs of many different RiPP subclasses [96]. Phylogenetic comparison of all radical SAM enzymes of characterized BGCs from MIBiG revealed several clades corresponding to enzymes involved in the biosynthesis of (multiple subclasses of) RiPPs (figure 2A). The protein sequences from these clades were mapped to the 20 different families of radical SAMs, by comparing their sequences with those of representatives of each family with BLAST. Nine of the 20 families were identified among all radical SAM enzymes, but only those containing a SPASM/Twitch-domain or a B12-binding domain were found among RiPP-related rSAMs. In this tree, the radical SAM enzymes with a B12-binding domain can be seen to clade together. These typically catalyse methylation, such as in bottromycins [97].

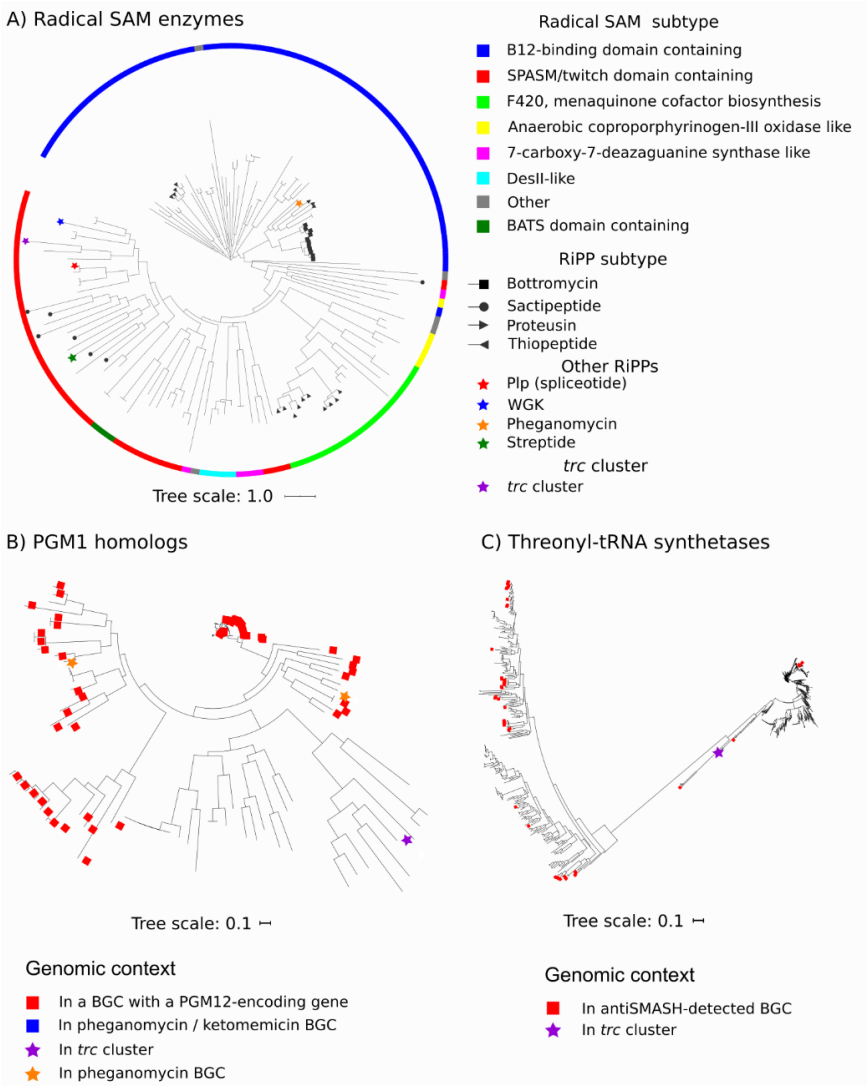

Figure 2. Homologs of TrcB, TrcC and TrcD overlap different types of BGCs predicted by antiSMASH. The relevant protein from the *trc* gene cluster is marked by a purple star. A) phylogenetic tree of rSAM enzymes detected in the MIBiG database. TrcD did not clearly overlap with any RiPP-associated clades. B) Phylogenetic tree of PGM1 homologs annotated in the antiSMASH database. Homologs are closely related when a PGM12-encoding gene was found in the same BGC (red). Other PGM1 homologs, including TrcC, are identified in a wider variety of BGCs and form a separate clade. C) Phylogenetic tree of 1,594 homologs of ThrRS found among 1,295 *Streptomyces* genomes. Among these homologs a closely related group was found, containing the majority of the homologs (1,255; righthand clade). The other homologs, including TrcB, showed a larger diversity, and were frequently found encoded in antiSMASH-detected BGCs (red).

The radical SAM of pheganomycin also belongs to this clade, marking it as distinct from TrcD, which fell within the clade of SPASM/Twitch-domain containing proteins. These radical SAM enzymes often form crosslinks, such as in sactipeptides and ranthipeptides [55, 99], or perform structural rearrangements, such as in spliceotides [103]. WGK and several recently identified RiPPs are all modified by radical SAM enzymes that belong to this family [104, 105, 107, 108]. These RiPPs are small (~500 Da) and contain a single crosslink applied by the radical SAM enzyme. The WGK radical SAM enzyme is closely related to TrcD, suggesting a similar modification is applied here. Still, as the majority of radical SAM enzymes (118 out of 142) were not encoded by a RiPP BGCs, and several of these enzymes were also closely related to TrcD, the presence of *trcD* alone did not provide conclusive evidence on whether the *trc* cluster encoded a RiPP.

trcC encodes a putative ATP-grasp ligase. ATP-grasp ligases catalyse nucleophilic additions onto carboxylic acids, which are activated as acylphosphate intermediates with ATP [181]. Members of this family have been shown to be involved in the biosynthesis of two RiPPs, pheganomycin [162] and omega-ester containing peptides (OEPs or graspetides) [85]. The ATP-grasp ligase predicted here does not show any significant similarity to any ATP-grasp ligase related to the OEP-family RiPPs found in a recent genome mining effort [85], but it is similar to PGM1, the ATP-grasp ligase encoded by the pheganomycin BGC. During biosynthesis of pheganomycin, the ATP-grasp ligase PGM1 catalyses the addition of an unnatural amino acid to a precursor peptide, via ATP-dependent formation of aminoacyl-phosphate derivative of this amino acid, followed by nucleophilic attack of the N-terminus of the peptide precursor. Several other proteins encoded by the pheganomycin BGC are thought to be involved in the biosynthesis of the unnatural amino acid, including PGM12, which shows high similarity to amidinotransferases. AntiSMASH detects BGCs related to the pheganomycin BGC using pHMMs built for the detection of PGM1 and PGM12. TrcC was detected by the pHMM built for PGM1, but a protein matching the pHMM for PGM12 was not encoded by the *trc* cluster. A phylogenetic tree containing all PGM1 homologs from the antiSMASH database [137] provided further insights in the relationship between the TrcC and PGM1 (figure 2B). Mapped onto this tree were those BGCs that contained genes for

PGM1 homologues. All PGM1 homologues in close proximity to a PGM12 homolog grouped together in two large clades, while the remaining PGM1 hits formed a third clade. The PGM1 homologue encoded by the *trc* cluster claded with the latter group of PGM1 homologues. In other words, the relationship of the *trc* cluster to pheganomycin was no stronger than that of other BGCs containing genes encoding PGM1 homologs, but no PGM12 homologs. Altogether, while *trcC* is not specific to BGCs that specify RiPPs, its presence, in combination with the presence of the radical SAM-encoding gene *trcD* and small ORFs encoding putative precursors with RiPP precursor-like conservation patterns, is at least strongly suggestive.

Lastly, a threonine-tRNA synthase (ThrRS) is likely encoded by *trcB*. Aminoacyl–tRNA synthases (AARS) are an essential part of primary metabolism, as they provide amino-acyl tRNA precursors used in ribosomal biosynthesis. Aminoacyl-tRNAs are used as a precursor in the biosynthesis of various secondary metabolites and antibiotics, such as type I lanthipeptides [47], 3-thiaglutamate [113, 147], cyclodipeptides [240] and valanimycin [241]. The BGCs for these antibiotics sometimes contain copies of the AARS to synthesize additional amino-acyl tRNA precursors, as is suspected for valanimycin [242, 243]. Another function of the gene could be to provide a variant that is resistant to the product of the BGC. This is the case for borrelidin, which targets AARSs, but has a resistant variant encoded in its BGC [244]. Bioinformatic analysis showed that genomes containing secondary copies of genes for ThrRS are not uncommon; among the 1,295 *Streptomyces* genomes, 1,594 ThrRS homologs were detected, averaging 1.23 per genome (Figure 2C; cutoff: 300 bitscore). 1,258 of these formed a closely related and highly conserved group. Only 11 of these overlapped with an antiSMASH-detected BGC, suggesting these were the ThrRS homologs involved in primary metabolism. In contrast, 57 of the remaining 337 homologs overlapped with an antiSMASH-detected BGC, showing that the presence of secondary household genes in BGCs is not an uncommon occurrence. TrcB was well removed from the highly conserved clade, further supporting the idea that the *trc* cluster from *Streptomyces* sp. MBT27 specifies a yet uncharacterized natural product.

Enhanced expression of the *trc* gene cluster and identification of the biosynthetic proteins

To experimentally characterize the *trc* cluster and its products, we aimed to enhance its expression *in vivo* by constitutive and strong expression of the transcriptional regulator. BGCs typically encompass a gene for a pathway-specific activator, which determines largely the timing and level of gene expression of the cluster [186, 245]. This property can be harnessed to efficiently over-express a BGC, and thus allow identification of the natural products that are overrepresented in the culture fluid of the recombinant strains [246]. The *trc* cluster contains a regulatory gene, *trcR* that encodes a putative LuxR-family regulator. These regulators often function as activators of BGCs in Actinobacteria [247]. We therefore placed a copy of *trcR* behind the strong and constitutive *gapdh* promoter from *Streptomyces coelicolor* [212, 213]. For this, we amplified the entire gene plus 30 nucleotides downstream from the *Streptomyces* sp. MBT27 genome and inserted it as an NdeI/XbaI fragment behind the *gapdh* promoter in the integrative vector pSET152. This construct was then introduced into MBT27, whereby the empty vector was used as the control. In this way, we created recombinant strains MBT27-*gaptrcR* and MBT27-pSET152. In parallel, we replaced the core region spanning the genes encoding both predicted precursors, the regulator and *trcH3* with the apramycin resistance cassette *aac(3)IV*. As this region contained both the regulator and both predicted precursors, we expected any secondary metabolite production of the gene cluster to be abolished in this strain. To this end, we used a method based on the unstable multi-copy vector pWHM3 [248]. We cloned the flanks upstream (-1507/-39) and downstream (+136/1641) of this region in pWHM3-oriT, inserted the *aac(3)IV* apramycin resistance cassette in-between and introduced this knock-out construct into MBT27 via conjugation. After several rounds of growth on non-selective media, followed by selection for the appropriate phenotype (apramycin^{R}, thiostrepton^{S}), we confirmed the replacement of the genes with PCR. The mutant strain was named MBT27-CR (Centre Replaced).

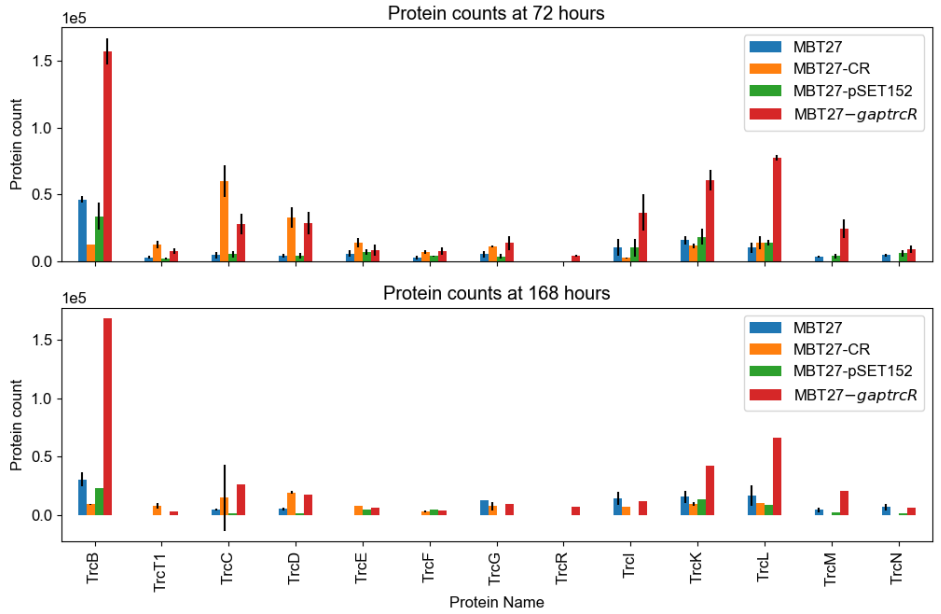

Figure 3. Expression of the *trc* cluster is affected by deletion of the core region and additional expression of *trcR*. All of the detected proteins were expressed at a higher level in a strain over-expressing TrcR. Surprisingly, in the proteome of the mutant MBT227-CR, which lacked the genes *trcA1-trcH3-trcR-trcA2*, several proteins encoded by the *trc* cluster were still detected, sometimes at a higher level than in the parental strain.

To establish the expression level of the *trc* gene cluster, and to see how gene expression would depend on the expression of *trcR*, we performed quantitative proteomics. As published previously, the expression level of BGCs corresponds very well to that of the metabolite produced from it, and hence the expression of the Trc proteins is a good measure of the expression of its cognate metabolite [249, 250]. For this, all strains were cultured in liquid minimal medium containing 0.5% (w/v) mannitol and 1% (w/v) glycerol as the carbon sources. Experiments were performed in triplicate. Mycelia were harvested after 72h and 168h, from which all proteins were isolated and analysed (Materials and Methods). Protein fragments were detected for 13 out of 19 proteins encoded by the *trc* cluster (Figure 3). Ten of these proteins were detected in all strains, showing that the gene cluster is expressed without genetic modifications. Strain MBT27-*gaptrcR* showed the highest overall expression of the *trc* cluster, in agreement with the function of TrcR as a transcriptional activator for the pathway.

Interestingly, in the knockout strain MBT27-CR the products of several genes in the *trcB-trcG* operon were detected at higher intensities than in the parent MBT27, despite the fact that the regulatory gene *trcR* had also been removed. Based on this, it appears that besides TrcR, other activating mechanisms exist. No other regulators were encoded on the contig that contains the *trc* cluster. Only one predicted regulator was detected at significantly higher levels in the proteome of both MBT27-CR and MBT27-*gaptrcR* was WP_167161651.1 (Figure S4). The gene encoding this protein was found in a terpene BGC, which suggests that it would function as a cluster-specific regulator for that BGC. Unfortunately, no other gene products of this BGC were detected by proteomics, so the involvement of this regulatory protein in the regulation of either the *trc* cluster or the terpene BGC could not be determined.

The proteins corresponding to the genes *trcI-trcH4* were still expressed in the mutant at levels comparable to those found in the parental strain MBT27. The only protein that was detected at significantly lower levels in the mutant was the threonine ligase TrcB, which also showed the strongest increase in MBT27-*gaptrcR* (~17-fold at 72 h). Taken together, these data show that TrcR functions as an activator of the *trc* cluster. However, the fact that the *trc* cluster was readily expressed in the wild-type strain indicates that TrcR it is not required *per se* for its expression. Interestingly, when *trcR* was removed alongside *trcA1*, *trcA2* and *trcH3*, most other gene products were more highly expressed. We cannot explain this based on the available data. Removal of these genes also changed the upstream regions of both putative operons, which could affect their transcriptional regulation. In addition, the product of the *trc* cluster itself may play a role in the regulation, as has been previously reported for other RiPPs [251]. Further experiments are required to unravel the exact regulatory mechanism of the *trc* cluster.

Analysis of the secreted metabolome by LC-MS

We then set out to identify the product synthesized by the *trc* cluster. We first removed the apramycin resistance cassette from the mutant strain via pUWLCRE-mediated Cre-Lox recombination [200], creating the in-frame deletion mutant MBT27-IFD. This strain together with MBT27 WT, MBT27-

pSET152 and MBT27-*gaptrcR* were cultured in liquid minimal medium and extracted with *n*-butanol, which has been previously used in the wide-scale extraction of different types of RiPPs simultaneously [252]. A total of six replicates were taken of each modified strain, while four replicates were taken for the wild-type strain. The crude extracts were analysed using LC-MS (Figure 4A and B), and mass features were extracted from the chromatograms with MzMine for comparative metabolomics using MetaboAnalyst (Materials and Methods).

Analysis of the heatmap based on the detected mass features revealed large differences among all the extracts (Figure S5). ANOVA analysis showed that 315 masses were differentially expressed among all the groups. For each of these masses, each possible pair of strains was compared with Fisher's LSD ($p \leq 0.05$) to find which masses were differentially expressed between which strains. Surprisingly, 25 different mass features were detected in higher intensities in the extracts of MBT27-IFD, as compared to the other groups. Apparently, removing part of the *trc* cluster altered the secondary metabolite profile of MBT27 in many ways. These masses may correspond to shunt products of the biosynthetic pathway of the *trc* cluster, as it was still partially expressed. Alternatively, abolishing production of the *trc* cluster may have freed up more resources for other BGCs, boosting their production, as previously reported [253]. In contrast, only two peaks were detected at higher intensities in extracts from MBT27-pSET152 compared to all other extracts. In total, 85 masses were detected in significantly higher levels in the extracts of MBT27-*gaptrcR*, compared to the extracts of all other strains (Fisher's LSD, $p \leq 0.05$).

Considering that the *trc* cluster was partially deleted, we selected candidate mass features which were not detected in the extracts of MBT27-IFD (peak area <= 3000), building on the assumption that secondary metabolite production of this gene cluster had been abolished. These criteria applied to 113 mass features, of which more than half (66) were detected at higher levels in the extracts of strain MBT27-*gaptrcR*. To further organize the results, we used molecular networking, through the Global Natural Products Social Molecular Networking (GNPS) platform to find relations between the identified mass features (Figure 4C) [254]. Using GNPS, mass features are represented as nodes,

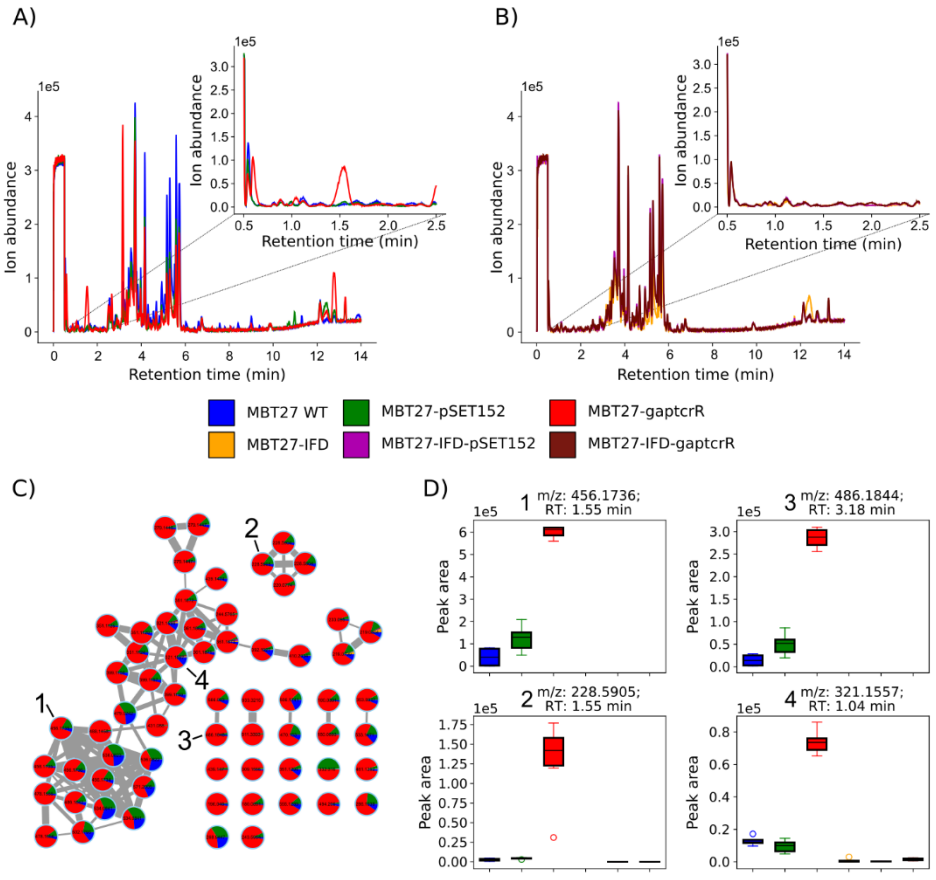

Figure 4. Comparative LC-MS analysis reveals large differences between the extracted strains. A) Overlay LC-MS chromatograms of MBT27 WT, MBT27-pSET152 and MBT27-*gaptrcR*. B) Overlay LC-MS chromatograms of MBT27-IFD, MBT27-IFD-pSET152 and MBT27-IFD-*gaptrcR*. C) Molecular families of the ions which were significantly enhanced in the extracts from a strain over-expressing TrcR as compared to the extracts of all other strains. At the same time, their production was very low or abolished in the strains lacking the core region of the *trc* cluster, regardless of whether trcR was over-expressed or not. Data of the complete network can be found in Data S1, and of this network in Data S2. D) Boxplot of selected peaks in the highlighted molecular families in C, compared among all six extracted strains.

which are connected to each other by edges due to similarities in their MS/MS spectra, or fragmentation patterns. When molecules share fragmentation patterns, it indicates that they have structural similarities. For 73 mass features

that were not detected in the extracts of MBT27-IFD, MS/MS spectra were obtained and the masses were compared. Most masses appeared as nodes connected into a single molecular family, suggesting that they belong to a related group of metabolites. The most highly expressed of these was a node with *m/z* 456.1736 (Figure 4D - 1). Several other highly abundant masses that were not linked with GNPS molecular families were still identified as related to this node, like it's doubly charged ion (*m/z* 228.5905, Figure 4D - 2), or nodes where the mass difference could be related to a number of atoms (*m/z* 486.1844, 456.1736 + mass of CH_2O group, figure 4D – 3). To look for further evidence that these peaks did not originate from effects of the enhanced expression of *trcR* that were unrelated to the *trc* cluster, we introduced the empty vector pSET152 and construct pAK10 into MBT27-IFD using conjugation, creating the strains MBT27-IFD-pSET152 and MBT27-IFD-*gaptrcR*, respectively. The mass features of interest were not detected in extracts of either of these strains, which makes it more plausible that they indeed originate from the *trc* cluster, and did not arise due to secondary effects from expressing the regulator. Further evidence for this could be gained by heterologously expressing the BGC in a different strain.

Of the masses described in the network, most had an *m/z* below 500 Da. While these masses are small for RiPPs, several RiPPs have been reported with a core peptide of only a few amino acids. These include the RiPPs modified by radical SAM enzymes that are closely related to TrcD (see above) [104, 105]. Similarly, the masses detected here may have been derived from the TTGWQ core sequence, which has a monoisotopic mass of 591.27 Da. However, we were unable to match the predicted mass to any of the identified masses in the LC-MS run, or identify amino acid residues from their respective MS/MS spectra, even when considering dehydration and deoxygenation of the precursor peptide. Additionally, it was not possible to identify a known structural class for these masses when the data were analysed using the MS2LDA tool [255]. Given the large number of enzymes encoded by the *trc* cluster, many different modifications to the precursor peptide are likely. Head-to-tail cyclization, for example, would make structure elucidation from MS/MS spectra difficult. Further structure resolution is required to completely resolve the structure of these metabolites and to determine whether or not they are RiPPs.

Conclusions and final perspectives

We have found a novel candidate RiPP BGC using decRiPPter, called the *trc* cluster, which was partially characterized via mutational analysis, expression, proteomics and mass spectrometry. The gene cluster contains several genes that relate to RiPP BGCs, such as a gene encoding an ATP-grasp ligase closely related to PGM1 and a gene encoding a radical SAM. The exact combination of enzymes has not been identified before, suggesting a novel natural product is specified by the *trc* cluster. We have not yet elucidated the natural product produced from the BGC. However, our bioinformatic analysis suggests that the gene cluster specifies a RiPP, whereby in particular the multiple TTGWQ repeats in the putative precursor peptides are suggestive of a RiPP, as short repeats are found in the precursors peptides of various RiPP sublasses. Enhanced expression of the regulatory gene *trcR* resulted in increased expression of the *trc* cluster, which could be correlated with the increased production of several secondary metabolites within the range of 400-600 Da. A few of these compounds were no longer produced when the cluster was partially removed, suggesting that these masses were products of the *trc* gene cluster. The mutant also lacked the genes *trcA1* and *trcA2*, suggesting that they may be involved in the biosynthesis of these products. We have so far been unable to confirm whether or not these products originated directly from the *trcA1* or *trcA2* precursors. Future studies will have to unveil the exact nature of the candidate masses, their relatedness, and the final product of the *trc* cluster.

Materials and Methods

Bioinformatics

Phylogenetic trees

For the generation of the phylogenetic trees, proteins were aligned with MUSCLE [153], and trees were generated with FastTree V2.1 [154] and visualized using iTOL [256].

Radical SAM

To create the radical SAM dataset, all proteins from the MIBiG database V2.0 [29] were scanned with hmmsearch [65, 134] against the Pfam [75] model of the radical SAM enzyme (PF04055), using the trusted cutoffs. The resulting proteins were mapped to radical SAM families by looking for the best hit among representatives of these families previously outlined [100] (Table S3). Phylogenetic trees were created as described above.

PGM1 homologs

The antiSMASH database [257] was queried for all BGCs containing a PGM1 homolog using the built-in query system. All BGCs were downloaded and parsed with BioPython [258] to detect PGM12 and PGM1 homologs. Phylogenetic trees were created as described above.

Threonyl-tRNA synthetases

Threonyl-tRNA synthetases were detected in 1,295 *Streptomyces* genomes analyzed previously with decRiPPter (Chapter 3). Protein homologs were detected with NCBI BLAST v.2.6 [56, 127] using a bit score cutoff of 300. Phylogenetic trees were created as described above.

Experimental procedures

Bacterial strain and growth conditions

Streptomyces sp. MBT27 was obtained from the Leiden University strain collection, which had been previously isolated from the Qingling Mountains [229]. Media components were purchased from Thermo Fisher Scientific, Sigma-Aldrich or Duchefa Biochemie. For strain cultivation on solid media, *Streptomyces* spores were spread on mannitol soya flour agar (SFM; 20 g/L Agar, 20 g/L mannitol, 20 g/L soya flour, supplemented with tap water) prepared as described previously [210], and incubated at 30°C. Spores were harvested after 4-7 days of growth when the strain started to produce a grey pigment, by adding water directly to the plate and releasing the spores with a cotton swab. Spores were centrifuged and stored in 20% glycerol.

For cultivation in liquid media, 20-50 µL of a dense spore stock was inoculated into 100 mL shake flasks with coiled coils containing 20 mL of the medium of interest. For extractions, NMMP was used (0.60 mg/L $MgSO_4$, 5 mg/L NH_4SO_4, 5 g/L Bacto casaminoacids, 1 mL trace elements (1 g/L $ZnSO_4.7H_2O$, 1 g/L $FeSO_4.7H_2O$, 1 g/L $MnCl_2.4H_2O$, 1 g/L $CaCl_2$, anhydrous)), while for genomic DNA isolation, a 1:1 mixture of TSBS: YEME with 0.5% glycine and 5 mM $MgCl_2$ was used (TSBS: 30 g/L Bacto Tryptic Soy Broth, 100 g/L sucrose; YEME: Bacto Yeast Extract: 3 g/L, Bacto Peptone 5 g/L, Bacto Malt Extract 3 g/L, glucose 10 g/L, sucrose 340 g/L).

E. coli strains JM109 and ET8 were used for general cloning purposes and demethylation, respectively. Strains were cultivated in liquid LB and on LB-agar plates at 37°C.

Molecular biology

All materials and primers were purchased from Sigma-Aldrich or Thermo Fisher Scientific unless stated otherwise. Restriction enzymes and T4 ligase were purchased from NEB. Restriction and

ligation protocols were followed as per manufacturer's description. For amplification of DNA fragments with PCR, Pfu polymerase was used. Primers were designed with T_m of the annealing region roughly equal to 60°C. Standard PCR protocols consisted of 30 cycles (45s DNA melting @ 95 °C, 45s primer annealing @55°C-65°C, 60s-180s primer elongation @ 72°C), but PCR protocols were optimized where necessary.

Following construction of the vectors (see below for specifics), constructs were transferred to MBT27 by conjugation [210]. Briefly, 50 µL of a dense MBT27 spore stock was added to 500 µL 2xYT, and a heat shock was applied at 50 °C for 10 minutes to trigger spore germination. In parallel, *E. coli* ET8 containing the construct of interest was grown until the OD_{600} measured 0.6 – 0.8 in 10 mL LB containing 50 µg/mL kanamycin, 50 µg/mL chloramphenicol and, as required, 20 µg/mL thiostrepton and 50 µg/mL apramycin. *E. coli* cultures were centrifuged, washed twice with LB to remove any remaining antibiotics, mixed with the germinated MBT27 spores and plated out on SFM plates containing 10 mM $MgCl_2$ and 10 mM $CaCl_2$. The plates were incubated at 30°C for 14-18 hours, and overlaid with 1.2 mL H_2O containing 417 µg/mL chloramphenicol, and, as required, 417 µg/mL thiostrepton and 1.04 mg/mL apramycin.

MBT27-CR knockout mutants were created by replacing the gene cluster with an *aac(3)IV* apramycin resistance cassette via homologous recombination. The -1553/-209 and +18/+1561 regions upstream and downstream of the *trc* cluster were amplified by PCR with the trc_LF_F/trc_LF_R and trc_RF_F/trc_RF_R primer pairs (table S1), respectively, and inserted into the pWHM3-oriT vector (Table S2) into the EcoRI/HindIII sites. The *aac(3)IV* apramycin resistance cassette was inserted into the created XbaI site, creating pAK9. pAK9 was transformed to *E.coli ET8* for DNA demethylation, which was used as a donor for transfer to MBT27 by conjugation [210]. Three colonies were picked after 4 days of growth and spread onto SFM plates without added antibiotic to allow for homologous recombination. Colonies containing the correct phenotype (apramycin-resistant, thiostrepton-sensitive) were picked and the homologous recombination was confirmed by PCR, using the trc_del_check_F/trc_del_check_R primer pair.

The strain MBT27-IFD was created by removal of the apramycin cassette from the strain MBT27-CR using the vector pUWLCRE [200]. This vector was conjugated to the strain MBT27-CR, and three separate colonies were picked and grown separately on SFM without antibiotics. After one round of growth, fresh spores were collected and plated at diluted concentrations to allow the spores to grow as individual colonies. From these, colonies were selected with the correct antibiotic resistance phenotype (apramycin-sensitive, thiostrepton-sensitive). Deletion was confirmed by PCR using the trc_del_check_F/trc_del_check_R primer pair.

Constructs for the overexpression of the *trcR* regulator were constructed as follows: the -0/+30 region of the *trcR* gene was amplified from the genomic DNA of MBT27 using the trcR_F/trcR_R primer pair, and placed into the EcoRI/XbaI site of the pSET152 vector. The -0/-457 upstream region of glyceraldehyde 3-phosphate dehydrogenase amplified from the genome of *S. coelicolor*, was obtained from previous studies [212, 213]. The promoter region was inserted into the EcoRI site and the engineered NdeI site, placing it directly upstream of the *trcR* gene. The resulting vector was named pAK10.

Extractions

Strains were cultured in 100 mL shake flasks containing 20 mL NMMP, with coiled coils at 30°C for 7 days. The entire culture was extracted by adding an equivalent volume of n-butanol and shaking overnight at 4°C. The mixture was collected and centrifuged at 4°C, after which the top butanol

layer was collected. The crude extracts were dried and weighed, and dissolved in methanol at a concentration of 1 mg/mL for LC-MS analysis.

LC-MS analysis

LC-MS/MS acquisition was performed using Shimadzu Nexera X2 UHPLC system, with attached PDA, coupled to Shimadzu 9030 QTOF mass spectrometer, equipped with a standard ESI source unit, in which a calibrant delivery system (CDS) is installed. The dry extracts were dissolved in MeOH to a final concentration of 1 mg/mL, and 2 µL were injected into a Waters Acquity Peptide BEH C_{18} column (1.8 µm, 100 Å, 2.1 × 100 mm). The column was maintained at 30 °C, and run at a flow rate of 0.5 mL/min, using 0.1% formic acid in H_2O as solvent A, and 0.1% formic acid in acetonitrile as solvent B. A gradient was employed for chromatographic separation starting at 5% B for 1 min, then 5 – 85% B for 9 min, 85 – 100% B for 1 min, and finally held at 100% B for 4 min. The column was re-equilibrated to 5% B for 3 min before the next run was started. The LC flow was switched to the waste the first 0.5 min, then to the MS for 13.5 min, then back to the waste to the end of the run. The PDA acquisition was performed in the range 200–400 nm, at 4.2 Hz, with 1.2 nm slit width. The flow cell was maintained at 40 °C.

The MS system was tuned using standard NaI solution (Shimadzu). The same solution was used to calibrate the system before starting. Additionally, a calibrant solution made from Agilent API-TOF reference mass solution kit was introduced through the CDS system, the first 0.5 min of each run, and the masses detected were used for post-run mass correction of the file, ensuring stable accurate mass measurements. System suitability was checked by including a standard sample made of 5 µg/mL paracetamol, reserpine, and sodium dodecyl sulfate, which was analyzed regularly in between the batch of samples.

All the samples were analyzed in positive polarity, using data dependent acquisition mode. In this regard, full scan MS spectra (*m/z* 100 – 1700, scan rate 10 Hz, ID enabled) were followed by two data dependent MS/MS spectra (*m/z* 100 – 1700, scan rate 10 Hz, ID disabled) for the two most intense ions per scan. The ions were selected when they reach an intensity threshold of 1500, isolated at the tuning file Q1 resolution, fragmented using collision induced dissociation (CID) with fixed collision energy (CE 20 eV), and excluded for 1 s before being re-selected for fragmentation. The parameters used for the ESI source were: interface voltage 4 kV, interface temperature 300 °C, nebulizing gas flow 3 L/min, and drying gas flow 10 L/min. The parameters used for the CDS probe were: interface voltage 4.5 kV, and nebulizing gas flow 1 L/min.

LC-MS based comparative metabolomics

All raw data obtained from LC-MS analysis were converted to mzXML centroid files using Shimadzu LabSolutions Postrun Analysis. The converted files were imported and processed MZmine 2.5.3 [214]. Throughout the analysis, *m/z* tolerance was set to 0.002 *m/z* or 10.0 ppm, retention time (RT) tolerance was set to 0.05 min, noise level was set to 2.0E2 and minimum absolute intensity was set to 5.0E2 unless specified otherwise. Features were detected (polarity: positive, mass detector: centroid) and their chromatograms were built using the ADAP chromatogram builder [215] (minimum group size in number of scans: 10; group intensity threshold: 2.0E2). The detected peaks were smoothed (filter width: 9), and the chromatograms were deconvoluted (algorithm: local minimum search; Chromatographic threshold: 90%; search minimum in RT range: 0.05; minimum relative height: 1%; minimum ratio of peak top/edge: 2; peak duration 0.03 – 3.00 min). The detected peaks were deisotoped (maximum charge: 5; representative isotope: lowest *m/z*). Peak lists from different extracts were aligned (weight for RT = weight for *m/z*; compare isotopic

pattern with a minimum score of 50%). Missing peaks detected in at least one of the sample were filled with the gap filling algorithm (RT tolerance: 0.1 min). Among the peaks, we identified fragments (maximum fragment peak height: 50%), adducts ($[M+Na]^+$, $[M+K]^+$, $[M+NH_4]$, maximum relative adduct peak height: 3000%) and complexes (Ionization method: $[M+H]^+$, maximum complex height: 50%). Duplicate peaks were filtered. Artifacts caused by detector ringing were removed (*m/z* tolerance: 1.0 *m/z* or 1000.0 ppm) and the results were filtered down to the retention time of interest. The aligned peaks were exported to a MetaboAnalyst file. From here, peaks were additionally filtered to keep only peaks present in all three replicates, using in-house scripts. The resulting peak list was uploaded to MetaboAnalyst [216], log transformed and normalized with Pareto scaling without prior filtering. Missing values were filled with half of the minimum positive value in the original data. Heatmaps and volcano plots were generated using default parameters.

Molecular networking

Raw LC-MS data were processed first in MZmine 2as described above, with added steps for MS2 mass detection (polarity: positive, mass detector: centroid, noise level: 0), and MS2 pairing (m/z range 0.05 Da, RT range 0.2 min).. The processed data were then exported to GNPS-FBMN as a .mgf spectra file and a .csv quantification table, with the following parameters: merge MS/MS enabled - spectra to merge across sample, m/z merge - mode most intense, intensity merge mode - maximum intensity, mass deviation - 0.005 or 20 ppm, cosine threshold - 60%, peak count threshold - 0%, Filter rows - only with MS2. The exported files, together with a metadata file describing the samples, were submitted to the Global Natural Products Social Molecular Networking (GNPS) tool for molecular networking [254]. The Feature-Based Molecular Networking (FBMN) workflow [259] was used adopting the default parameters apart from the maximum connected component size changed to 200, and disabling of filtration of peaks around precursor ion mass and peaks in 50Da window. The molecular networking job can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1699fc78a3b84cefb36271cf673d3b83.
Additionally, the data were analyzed using the MS2LDA tool for the identification of likely sub-structures in the extracts based on the obtained fragmentation pattern of the molecules [260]. The default parameters were used for TOF data apart from LDA free motifs being set to 300, and databases for urine and plant motifs being excluded in the analysis. The MS2LDA job can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2886d1b2edc44c31ba693f5eab7cec73. The results obtained from the MS2LDA workflow were mapped to the molecular network previously generated using the MolNetEnhancer workflow [255], and the obtained network was visualized using Cytoscape [133].

Mass spectrometry-based quantitative proteomics

20 µL of dense spore stocks were inoculated in NMMP and grown for 7 days as described above. 1 mL samples were taken after 2 and 7 days. Mycelium was gathered by centrifugation and washed with disruption buffer (100 mM Tris-HCl, pH 7.6, 0.1 M dithiothreitol). The samples were sonicated for 5 minutes (in cycles off 5s on, 5s off) to disrupt the cell wall, and centrifuged at max speed for 10 minutes to collect the proteins. Proteins were then precipitated using chloroform-methanol [217]. The dried proteins were dissolved in 0.1% RapiGest SF surfactant (Waters) at 95°C. Protein digestion steps were done according to van Rooden et al [218]. After digestion, formic acid was added for complete degradation and removal of RapiGest SF. Peptide solution containing 8 µg peptide was then cleaned and desalted using the STAGETipping technique [219].

Final peptide concentration was adjusted to 40 ng/µL with 3% acetonitrile, 0.5% formic acid solution. 200 ng of digested peptide was injected and analysed by reverse-phase liquid chromatography on a nanoAcquity UPLC system (Waters) equipped with HSS-T3 C18 1.8 µm, 75 µm X 250 mm column (Waters). A gradient from 1% to 40% acetonitrile in 110 min was applied, [Glu^{1}]-fibrinopeptide B was used as lock mass compound and sampled every 30 s. Online MS/MS analysis was done using Synapt G2-Si HDMS mass spectrometer (Waters) with an UDMS^{E} method set up as described [218].

Mass spectrum data were generated using ProteinLynx Global SERVER (PLGS, version 3.0.3), with MS^{E} processing parameters with charge 2 lock mass 785.8426 Da. Reference protein database was downloaded from GenBank with the accession number GCA_001278075.1. The resulting data were imported to ISOQuant [220] for label-free quantification. The TOP3 quantification result from ISOQuant was used when further investigating the data.

Supplementary information for Chapter 5

Data S1. Molecular network of all the ions detected in the extracts of MBT27 and derived strains. The relative intensities of the ions are mapped to the nodes as pie charts, and the nodes are labelled by the monoisotopic mass of their precursor ions. The edge thickness represents the cosine score, which indicates the degree of the relatedness of the MS/MS spectra. Available upon request.

Data S2. Molecular network of the ions detected in the extracts of MBT27 and derived strains, as represented in Figure 4C. The relative intensities of the ions are mapped to the nodes as pie charts, and the nodes are labelled by the monoisotopic mass of their precursor ions. The edge thickness represents the cosine score, which indicates the degree of the relatedness of the MS/MS spectra. Available upon request.

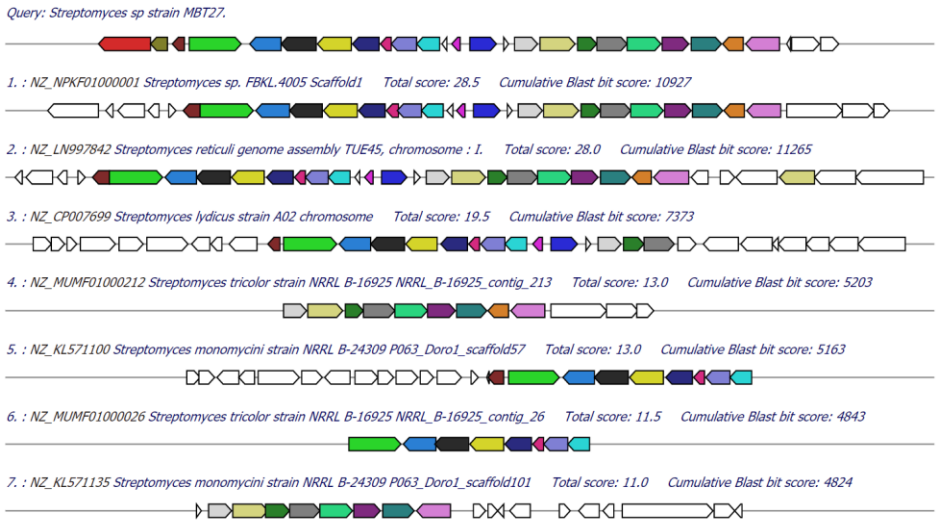

Figure S1. MultiGeneBlast analysis reveals five gene clusters closely related to the *trc* gene cluster. Homologous gene clusters all contained the same set of the genes, except for *Streptomyces lydicus* sp. A02, which a transporter (*trcT2*) and five other genes (*trcKLH4MN*). Homologs of *trcA1* and *trcA2* were also not always detected in genomes in which the cluster was detected on two different contigs.

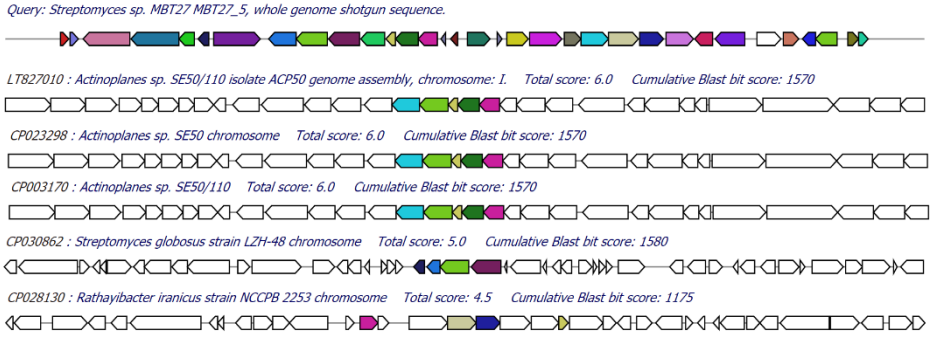

Figure S2. MultiGeneBlast shows several gene clusters distantly related to the *trc* cluster.

```
WP_059250216.1      MSFARRRTAKAVATLFLTSCAALAVTATTAQ--DTHHAAGRSTTGTH-----ALGATTGWQSPA----------PATVQDTTGWQ
WP_046926578.1      MPFARRRPAKALATIVLASCATLTATTTAQA-------AGHTTHRTATRAP-LAQHTTGWQTPA----------PIA-RSTTGWQ
TrcA1               MSFARRSPAKALAATLLVAGATLTVTAQTAQ----------ATTGWQSPAPAAAPATTGWQAPAPDTTGWQNPGPAAARATTGWQ
TrcA2               MSFARRSPAKALAATLLVAGATLTVTAQTAQ----------ATTGWQSPAPAAAPATTGWQAPA----------P----DTTGWQ
WP_030024344.1      MSFARRRPVKALAAISLAACAALTVTTTTAQTADTTTVAAPTTTGWQAPAP-TIQHTTGWQAPA----------P-----TTGWQ
                    *.**** ..**:*:  *.: *:*:.*: :           :*            *****:**          *     *****
```

Figure S3. Alignment of all detected precursors shows conserved TTGWQ motif. All predicted precursors varied between 61 and 75 amino acids in length. Alignment of the sequences showed a conserved N-terminal sequence, combined with a more varied C-terminal sequence. However, despite the variation, between two and four repeats of the TTGWQ motif were consistently detected.

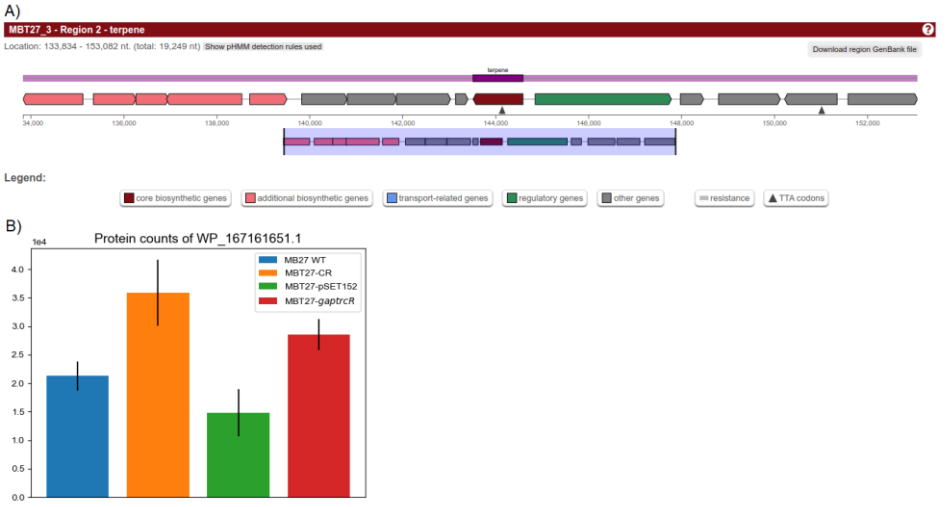

Figure S4. Protein levels of a differently expressed regulator. A regulatory protein encoded by a terpene BGC (A) is detected at significantly higher levels in the proteome of MBT27-CR and of MBT27-*gaptrcR* compared to the proteome of the control strains (B).

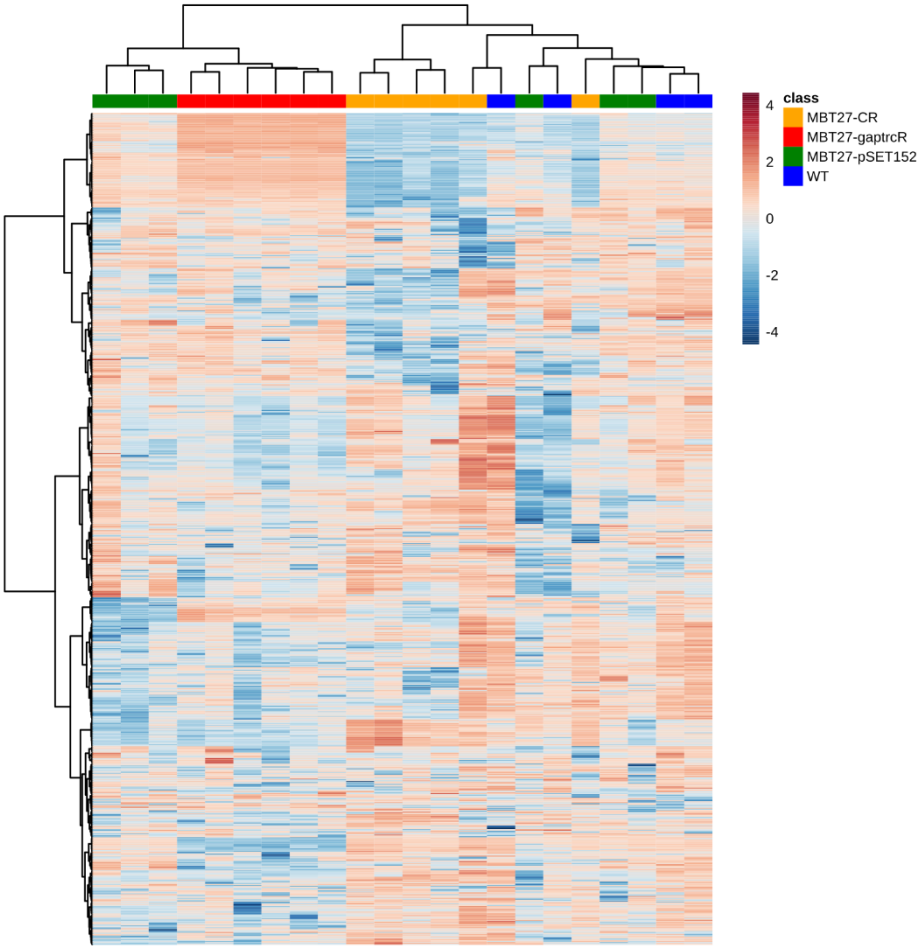

Figure S5. Heatmap of mass features detected by LC-MS and processed with MZmine shows large differences among the extracted strains. Each row represents a single mass feature and each column represents a single extract, while the colour scale indicates the $\log_{10}$-scaled intensity of the mass features for each extract. Large differences are even seen among replicates, arising from the different biological replicates of the modified strains which were extracted.

Table S1. Primers used in this study.

Primer name	Primer sequence
trcR_F	gatc GGTACC CAT ATGCCGGAGAACCTGGGC
trcR_R	gatc TCTAGA AAGCTT CCGTTGCACTACATGGTCGAAGCC
trc_LF_F	cgta GAATTC GCGACTGACAGCACCACTGG
trc_LF_R	cgta TCTAGA CACCGATCCACACCAGTGG
trc_RF_F	cgta TCTAGA GGCCGTAGGGACAATCAATCACC
trc_RF_R	cgta AAGCTT CACTGTCAGCCACGCAATGATGG
trc_del_check_F	GGTGTCGAAATCGGACATGG
trc_del_check_R	CCTTTGGCAGGTCGTCGCTGACC

Table S2. Plasmids used in this study.

Plasmid	Description	Reference
pSET152	Integrative *E. coli* / *Streptomyces* shuttle vector.	Bierman *et al.* [199]
pWHM3	Unstable *E. coli* / *Streptomyces* shuttle vector with high copy number; used for homologous recombination	Vara *et al*. [198]
pUWLCRE	Unstable *E. coli/Streptomyces* shuttle vector containing the Cre recombinase enzyme. behind a constitutive promoter.	Fedorshyn *et al*. [200]
pAK9	pSET152 containing *trcR* behind GAP promoter from *S. coelicolor* (SCO1947).	This work.
pAK10	pWHM3 containing regions flanking the *trc* cluster.	This work.

Table S3. Strains used in this study.

Strain	Description	Reference
MBT27 wildtype	*Streptomyces* sp. MBT27, previously isolated from the Qingling mountains.	Zhu *et al.* [229]
MBT27-pSET152	*Streptomyces* sp. MBT27containing an empty pSET152 (empty vector).	This work.
MBT27-*gaptrcR*	*Streptomyces* sp. MBT27containing pAK4.	This work.
MBT27-CR	*Streptomyces* sp. MBT27in which the **c**entre region from *trcA1* to *trcA2* was **r**eplaced by the apramycin resistance cassette aac3(IV).	This work.
MBT27-IFD	MBT27-CR from which the apramycin cassette was removed with the Cre-Lox system.	This work.
MBT27-IFD-pSET152	MBT27-IFD containing an empty pSET152 (empty vector).	This work.
MBT27-IFD -*gaptrcR*	MBT27-IFD containing pAK9.	This work.

Table S4. Overview of UniProt accessions of representatives of radical SAM subfamilies used to classify radical SAM proteins from MIBiG. Adapted from Holliday *et al* [100].

Class name	Uniprot Accession IDs
tRNA wybutosine-synthesizing	Q57705
spectinomycin biosynthesis (SpcY-like)	Q9S1L5
antiviral proteins	O70600
AviX12-like	Q93KV6
lipoyl synthase like	P60716
FeMo cofactor biosynthesis protein	P11067
DesII-like	Q9ZGH1
anaerobic coproporphyrinogen-III oxidase like	A0A060PWX2, V0VQG0, P9WP73, P52062, P32131, Q796V8, C6FX53, Q9FB10, I3NN68
BATS domain containing	P12996, Q58195, Q9X0Z6, Q46E78, P30140, C6FX51, Q6PSL4
7-carboxy-7-deazaguanine synthase like	O31677, A0A0H3KB22, O54060
PLP-dependent	Q841K7, A4J6G2, Q9XBQ8, P39280
methylthiotransferase	Q9WZC1, Q96SZ6, P0AEI4, P54462
F420, menaquinone cofactor biosynthesis	Q58826, Q57888, Q9XAP2, Q5SK48
organic radical-activating enzymes	Q84F14, P0A9N8, O87941, Q8GEZ7, P0A9N4, P39409
methyltransferase	P36979, Q9FBG4, Q58036
spore photoproduct lyase like	A4IQU1, Q97L63
elongater protein-like	Q02908
B12-binding domain containing	Q3ME29, Q2MFI7, A0A095DNL6, Q1Q0N1, P26168, A8R0J7, A8R0J8, D2KTX8, F8JND9, F8JNE0, Q58275, Q8GHB6, O24770, Q70KE5, B3QHD1, B9ZUJ4, D2KTX6, Q60AV6, Q5IW50, A8R0J3, Q56184, Q50258, Q6QVU0, Q8KCU0, C0JRZ9, Q8KBK9
SPASM/Twitch domain containing	A0A1I5E523, P69848, A1B2Q7, D0QZJ5, Q841K9, B8J367, P27507, Q8RAM6, Q8G907, A0A0E2Q059, Q46CH7, Q0TTH1, O31423, Q51741, Q53U14, P9WJ79, Q9X758, P71011, A0A095EC78, E5KJ95, C2TQ82, Q6E3K8

5