# Recapitulation of the evolution of biosynthetic gene clusters reveals hidden chemical diversity on bacterial genomes

Pablo Cruz-Morales[1,*], Christian E. Martínez-Guerrero[1], Marco A. Morales-Escalante[1], Luis Yáñez-Guerra[1], Johannes Florian Kopp[2], Jörg Feldmann[2], Hilda E. Ramos-Aboites[1] & Francisco Barona-Gómez[1,*]

[1] Evolution of Metabolic Diversity Laboratory, Langebio, Cinvestav-IPN. Irapuato, Guanajuato, México.

[2] Trace Element Speciation Laboratory (TESLA), College of Physical Sciences. Aberdeen, Scotland, UK.

Authors for correspondence:

Francisco Barona-Gómez (fbarona@langebio.cinvestav.mx)

Pablo Cruz-Morales (pcruz@langebio.cinvestav.mx)

## Abstract

Natural products have provided humans with antibiotics for millennia. However, a decline in the pace of chemical discovery exerts pressure on human health as antibiotic resistance spreads. The empirical nature of current genome mining approaches used for natural products research limits the chemical space that is explored. By integration of evolutionary concepts related to emergence of metabolism, we have gained fundamental insights that are translated into an alternative genome mining approach, termed EvoMining. As the founding assumption of EvoMining is the evolution of enzymes, we solved two milestone problems revealing unprecedented conversions. First, we report the biosynthetic gene cluster of the 'orphan' metabolite leupeptin in *Streptomyces roseus*. Second, we discover an enzyme involved in formation of an arsenic-carbon bond in *Streptomyces coelicolor* and *Streptomyces lividans.* This work provides evidence that bacterial chemical repertoire is underexploited, as well as an approach to accelerate the discovery of novel antibiotics from bacterial genomes.

## Introduction

The concept of genome mining can be defined as both an extension of the central dogma of molecular biology, situating metabolites at the downstream end, and a novel approach that promises to turn the discovery of natural products (NPs) drugs into a chance-free endeavor (1–3). Within the context of increased antibiotic resistance and emergence of modern diseases, and given the strong track of NPs (which are the result of specialized metabolism) in providing useful molecules to human welfare (3, 4), genome mining has revitalized the investigation into NP biosynthesis and their mechanisms of action (3, 5). Evidence of this resurgence has steadily increased since the first NP that was discovered using genome mining approaches, i.e the farnesylated benzodiazepinone ECO4601, entered into human clinical trials more than a decade ago (6–8).

In contrast to experimental hurdles, which have been acknowledged elsewhere (9), *in silico* genome mining has enjoyed a relatively higher success. Early genome mining approaches built up from the merger between a wealth of genome sequences and an accumulated biosynthetic empirical knowledge, mainly surrounding Polyketide Synthases (PKS) and Non-Ribosomal Peptide Synthetases (NRPSs) (10, 11). These approaches, which rely on high quality genome sequences due to the modularity and repetitive nature of PKSs and NRPSs, can be classified as: (i) chemically-driven, where the structure of a metabolite is linked to potential enzymes, such that the biosynthetic genes of an ‘orphan’ metabolite that has been isolated and structurally characterized, are identified (12); or (ii) genetically-driven, where known sequences of protein domains (13) or active-site motifs (14) help to identify putative biosynthetic gene clusters (BGCs) and their products. The latter relates to the term ‘cryptic’ BGC, defined as a genetic locus that has been predicted to direct the synthesis of a NP, but which remains to be experimentally confirmed (15).

Genome mining of NPs has also helped to prioritize strains and metabolites on which to focus for further investigation. During this process, based on *a priori* biosynthetic insights, educated guesses surrounding PKS and NRPS can be put forward, increasing the likelihood of discovering interesting chemical and mechanistic variations. Moreover, biosynthetic logics for a growing number of NP classes, such as phosphonates (16–18) and ribosomally synthesized post-translationally modified peptides (RiPPs) (19) are complementing early NRPS/PKS-centric approaches. Nevertheless, the current pace of deciphering novel biosynthetic logics, which can only be achieved after long periods of research, hampers our ability to cope with the rate of appearance of antibiotic resistance. Focusing on known chemical scaffolds with the concomitant high rate of rediscovering the same classes of NPs, although useful under certain circumstances and specific cases (20), seems also sub-optimal for the discovery of much-needed novel drugs. From these observations it becomes apparent that more efficient approaches that will lead to the discovery of novel chemistry are needed.

In this work we have developed an alternative method for current NP genome mining, which is guided by evolutionary theory. By means of integrating three evolutionary concepts related to emergence of specialized metabolism, we have gained fundamental insights that are translated into the discovery of novel NPs. First, we embraced the concept that new enzymatic functions evolve by retaining their reaction mechanisms, while expanding their substrate specificities (21). In consequence, this process expands enzyme families. Second, evolution of contemporary metabolic pathways frequently occurs through recruitment of existing enzyme families to perform new metabolic functions (22). Indeed, in the context of NP biosynthesis, cases of functional overlap driven by promiscuous enzymes that have been expanded and recruited have been reported (23, 24). The correspondence of enzymes to either central or specialized metabolism

is typically solved through detailed experimental analyses, but we argue here that these could also be achieved through phylogenomics. Third, BGCs are rapidly evolving metabolic systems, consisting of smaller biochemical sub-systems or 'sub-clusters', which may have their origin in central metabolism (25–27).

Integration of these three evolutionary principles was formalized as a bioinformatics pipeline, termed EvoMining, which can be defined as a functional phylogenomics tool for identification of expanded, repurposed enzyme families, with the potential to catalyze new conversions in specialized metabolism (Figure 1A). As this process does not rely on sequence similarity searches of previously identified NP biosynthetic enzymes, but rather on recapitulation of an evolutionary process, the predictive power of evolutionary theory is fully embraced. Moreover, given that predictions are done at the single-gene level, rather than looking at large PKS, NRPS or BGC sequence assemblies, low quality draft genome and metagenome sequences are compatible with this approach. Indeed, we demonstrate that EvoMining can predict biosynthetic genes for orphan molecules, as well as new NP biosynthetic pathways in model strains, both involving novel enzymatic conversions.

Experimentally, we focused in the phylum *Actinobacteria*, which includes renowned NP-producing genera, such as *Streptomyces*, that have provided a plethora of useful NPs (3). Experimental evidence for two key cases using *Streptomyces* species is reported to advance the field of genome mining of NPs. First, the BGC for the biosynthesis of the orphan small peptide aldehyde (SPA) leupeptin, of high economic importance, is identified in *Streptomyces roseus* ATCC 31245. Second, a novel NP synthesized by the model organisms *Streptomyces coelicolor* A3(2) and *Streptomyces lividans* 66 is experimentally characterized. Both of these case studies include novel chemical conversions catalyzed by enzymes that were blindly targeted by

EvoMining. Therefore, these results validate EvoMining as an alternative and complementary method for the discovery of potential drug leads. Moreover, the insights gained during this integrative approach suggest that bacterial genomes encode a larger chemical diversity yet-to-be discovered that can be untapped by means of using evolutionary theory.

## Results & Discussion

The EvoMining bioinformatics pipeline, in its current version 1.0, uses three input databases (green cylinders, **Figure 1A**). First, a genome database that contains the annotated genomes of 230 members of the phylum *Actinobacteria*, as retrieved from the GenBank database (**Table S1**). Second, a database containing the amino acid sequences of enzymes belonging to nine 'precursor supply central metabolic pathways' (PSCP), defined as previously (28). This dataset provides a universe of 103 enzyme families, to be used as query sequences (**Table S2**), which were extracted from genome-scale metabolic network reconstructions (GSMR) of model *Actinobacteria*. Third, a NP seed database consisting of 226 experimentally characterized BGCs (mainly from *Actinobacteria*), including: (i) NRPSs and PKSs biosynthetic systems extracted from specialized databases (10, 11); and (ii) other classes of well-described NPs biosynthetic pathways, e.g. terpenes, phosphonates and RiPPs, extracted from the literature (**Table S3**).

The sequences in the PSCP database were used as queries to retrieve homologous sequences contained in the genome database. The threshold used for defining homology was non-conservative, such that expansion events resulting from both gene duplication and horizontal gene transfer could be retrieved. When propagated through the genomes database, after homology searches, these query sequences gave rise to an enzyme family internal database. After a heuristic approach, an organism's enzyme expansion was called by statistical measure when the number of homologs on its genome was larger than the average of each enzyme family plus its standard

129 deviation. The enzymes that complied with this criterion were stored on the enzyme expansion
130 internal database (yellow cylinders, **Figure 1A**; **Table S4**). The expansion of each enzyme family
131 was sorted throughout a phylogenetic species tree (**Tree S1**), allowing taxonomic resolution of
132 expansions, as previously observed (28). With this approach we found that 98 enzyme families,
133 out of 103 enzymes from the PSCP database, had expansion events.

134 A critical function for the EvoMining approach is identification of enzyme families
135 expanded in concert with NP biosynthesis clusters. To accomplish this, the expanded enzyme
136 families were then mined for recruitment of their members within the context of NP biosynthesis.
137 In the cases where an expanded enzyme family could be connected via sequence homology to one
138 or more proteins within the NP seed database, their sequences were stored on the enzyme
139 recruitment internal database (yellow cylinders, **Figure 1A**, **Table S5**). It should be noted that
140 only a fraction of all sequences in the NP seed database have been characterized in the context of
141 NP biosynthesis. Often, not even functional annotation is provided. Therefore, the functional
142 association between recruited enzymes and this relatively large sequence space is supported by the
143 occurrence of the expanded enzymes within BGCs that have been linked to a known metabolite.
144 The enzyme recruitment internal database consisted of 23 enzyme families, including both known
145 recruitments, e.g. aconitase in phosphinothricin biosynthesis (16), and all related sequences
146 codified by the analyzed genomes. Thus, the functional potential of the NP seed and the genome
147 databases, together, is fully exploited.

148 The sequences of the recruited enzymes, together with those from the expanded enzyme
149 families, were used to make multiple sequence alignments (MSA) and Bayesian phylogenetic
150 reconstructions (**Figure 1A**). Moreover, in order to provide useful functional annotation for
151 interpretation of EvoMining phylogenetic trees (**Figure 1B**), a bidirectional best-hit analysis,

between the enzyme recruitment and the PSCP databases, was used for directing the labeling of central metabolic orthologs (red branches). We adopted this simple strategy, as it is safe to assume that NP biosynthetic enzymes will diverge significantly from central metabolic homologs. Homologs related to the very few known recruitments (blue branches) were therefore considered to be NP biosynthetic homologs identified by EvoMining. This proof-of-concept analysis provided 515 recruitment events, which are called EvoMining hits, and their gene identifiers (GIs) were used as queries to retrieve contigs (12-109 Kbp, 71.3 Kbp in average, 19.9 Kbp standard deviation). The retrieved contigs were then analyzed for putative NP BGCs using antiSMASH and ClusterFinder (29, 30). When an NP positive hit was obtained after this process, this was also noted in the phylogenetic tree (cyan branches).

The abovementioned functional annotation provides information that validates NP-related phylogenetic clades that consist of EvoMining hits (**Figure 2B**). Subtraction of the known (blue branches) and antiSMASH / ClusterFinder predicted (cyan branches) NP lineages, within the NP-related clades, reveals putative BGCs coding for repurposed enzymes only accessible by EvoMining (green branches). Henceforth, we refer to these homologs as EvoMining predictions, which we define as unknown NP biosynthetic enzymes supported by phylogenetic evidence encoded within previously undetected BGCs. Chemically, an implication of an EvoMining hit is that it uncovers enzymatic conversions, mainly involving diverging substrate specificities (but potentially also mechanistic variations), which in turn can lead to alternative biosynthetic logics and therefore chemical scaffolds. Thus, an EvoMining prediction, as a concept, is here applied to entire BGCs rather than to specific enzymes. In addition to **Figure 1**, EvoMining version 1.0 can be explored as online supplementary material at http://148.247.230.39/newevomining/new/evomining_web/index.html

## Evolutionary insights and performance of EvoMining

Of the 515 EvoMining hits we successfully retrieved contigs containing their cognate enzyme coding genes for 448 of them (71.3 Kbp on average; 87 %) (**Table S5**). Among these, many EvoMining hits (20 %) were included in contigs with internal gaps, which hampers sequence annotation. This subset, together with the remaining 13 % of the total hits whose contigs could not be retrieved, account for one third of contigs that come from highly fragmented genomes. So, the advantage of EvoMining in this respect is that predictions can be made, early on during analysis, in draft genomes (and metagenomes) that can be further improved. This provides an opportunity to prioritize in cost-effective manner large strain collections during genome-driven drug discovery efforts. Focusing on an EvoMining hit related to the enolase enzyme family (**Tree S2**), which was found in the genome of *Streptomyces sviceus* (1 scaffold of 9 Mbp with 552 gaps and 8X coverage, GI: 297196766), illustrates the benefit of EvoMining in this respect. The contig containing this recruited enolase (GI: 297146550) had 6 gaps including missing sequence at its 5' end. After closing of these gaps by sequencing PCR products, the complete sequences for several phosphonate-related enzymes, namely, alcohol dehydrogenase (*phpC*), phosphonopyruvate decarboxylase (*ppd*), nicotinamide mononucleotide adenyl transferase (*phpF*), carboxy-phosphonoenolpyruvate synthase (*phpH*, EvoMining hit) and aldehyde dehydrogenase (*phpJ*), could be annotated. Further sequence analysis suggested that indeed this locus encodes for a putative phosphonate BGC related to phosphinothricin (31–33) (**Text S1**, **Figures S1** and **S2**).

We further asked the question of whether EvoMining enzymes are indeed encoded within BGCs potentially directing the synthesis of NPs. The retrieved contigs were mined for BGCs of known classes of NPs using antiSMASH (29) as well as for putative new BGCs using ClusterFinder (30). From this analysis we found that these tools could predict BGCs for 62.5 %

and 10.5 %, respectively, of the contigs harboring EvoMining hits (73 % together). The remaining 27 % of contigs are unique EvoMining hits, and therefore potentially EvoMining predictions related to emerging BGCs and chemical scaffolds (**Figure 2A**). The enzymes included in this 27 % represent the core of EvoMining, and due to the lack of functional information subsequent analysis is experimentally and bioinformatically challenging. However, manual analysis of the green branches within the NP-related clade provided as an example in **Figure 1B**, allowed us to identify a highly conserved BGC-like locus present in the genus *Streptomyces* (see below discussion related to **Figure S7**, **Table S7** and **S8**).

As a prerequisite for embarking on characterization of completely unprecedented BGCs we first determined whether EvoMining hits are associated to particular BGC classes. For this purpose, we used antiSMASH to classify and count for the number of BGCs contained in each contig. Globally, BGCs for 22 out of the 24 categories used by antiSMASH (34) could be detected. Only aminoglycosides and indoles could not be detected, which may be due to the limited enzymatic repertoire explored by our seed NP and PSCP input databases. Among the 22 antiSMASH categories, type I PKSs and NRPSs represent the most abundant classes (**Figure S3**), confirming that EvoMining can identify well-known NP biosynthetic systems following a unique, non-biased strategy. Despite this convergence, at least with this limited analysis, it was also found that the number of EvoMining hits increased in parallel to the number of BGC classes (**Figure 2B**). The implication of this observation is that novel classes may be discovered for any given enzyme recruitment, as long as enough sequence space is explored.

For instance, among the 23 recruited enzyme families, only that of indole-3-glycerolphosphate synthase was linked to a single class of BGCs. It should be noted, however, that this family has the smallest number of EvoMining hits. At the other end, the enzyme families with

the larger number of EvoMining hits showed the highest BGC diversity, namely, 17 BGC classes for the asparagine synthetase enzyme family, followed by 10 BGC classes for both 3-phosphoshikimate-1-carboxyvinyl-transferase and 2-dehydro-3-deoxyphosphoheptonate aldolase families (**Figure 2B**). It seems, therefore, that some of the recruitments have high predictive potential when used as beacons for detection of BGCs. The latter may be due to the evolvability of enzymes towards specificities for common precursors of NPs, driven by enzyme promiscuity (23, 24), or because these enzymes catalyse recurrent reactions in NP biosynthesis with less mechanistic restrictions (27).

After our *in silico* analysis, we aimed to demonstrate that EvoMining could indeed be used to predict enzymes of unknown function involved in the synthesis of NPs. More specifically, we focused in providing experimental evidence to support two critical cases, which could not be solved with current methods: (i) the discovery of a BGC of an orphan metabolite that has been extensively investigated; and (ii) the discovery of a BGC driving the synthesis of an NP produced by well-studied model strains. Remarkably, and in agreement with the definition of an EvoMining hit, the enzymes identified by this approach are proposed to catalyse unprecedented chemical conversions.

## Discovery of leupeptin BGC in *Streptomyces roseus*: the orphan metabolite problem

Leupeptin is the first member of a large family of NPs generically known as small peptide aldehydes (SPA), and it is widely used in industry and bioresearch due to its potent anti-proteolytic activity (35). Despite the fact that leupeptin was discovered during the golden age of antibiotic research (35) its BGC remained elusive until now (**Figure 3**). The structure of leupeptin

includes a C-terminal aldehyde group in a peptide chain that includes an acyl group at its N-terminal end. Early biochemical studies of leupeptin revealed that: (i) this peptide is produced from acetate, leucine and arginine; (ii) an ATP-dependent synthetase is responsible for the condensation of an acetyl-leucine-leucine intermediary; (iii) this intermediary is released from an enzymatic complex before its condensation with an arginine residue; and (iv) the enzymatic complex responsible for the synthesis of acetyl-leucine-leucine-arginine, called leupeptidic acid, has a molecular mass of approximately 320 kDa (36–39).

A more recent study on the biosynthesis of the SPA flavopeptin produced by *Streptomyces* sp. NRRL-F6652 (40), suggested that synthesis of leupeptin by producing strains, such as *S. roseus* ATCC 31245 (35), may occur via a distinctive NRPS complex. Key features of this putative NRPS were proposed to be: (i) an acyl transfer domain typically found as an N-terminal starter C-domain, responsible for initial acylation of non-ribosomal peptides (41); (ii) three complete modules for peptide synthesis, two for leucine and one for arginine residues; and (iii) a reductase domain at its C-terminal end, responsible for reductive peptide release, leading to an aldehyde. Based in this modern proposal, after sequencing the genome of *S. roseus* ATCC 31245 (BioProject: PRJNA287805, Accession: LFML00000000), we searched for flavopeptin-like NRPSs. However, this approach proved to be unsuccessful.

Mining of the *S. roseus* genome sequence with EvoMining, by contrast, produced a hit to an enzyme annotated as argininosuccinate lyase (ASL), typically involved in arginine biosynthesis. ASLs condense succinate and arginine via an amidic bond at the guanidine group of the arginine in a reversible reaction (42). Indeed, two ASL homologs sharing 25 % amino acid sequence identity, which could be phylogenetically resolved, were found in the genome of *S. roseus* (**Figure 3A** and **Tree S3**). The first homolog, as expected, is located within a clade that includes central metabolic homologs (ASL1 or *argH* gene); whereas the second homolog is

267 located in a clade together with enzymes previously related to the biosynthesis of uridyl-peptides,
268 namely, napsamycin and pacidamycin (43, 44). Thus, the members of the latter clade, termed here
269 ASL2, are predicted to include recruited enzymes involved in NP biosynthesis, and that these
270 enzymes are performing a related but different chemistry than that catalyzed by ASL1.

271 After detailed annotation of the region surrounding the recruited ASL2 gene, a NRPS
272 gene was found (**Figure 3A**). The product of this gene was predicted to have an N-terminal
273 condensation domain ($C_1$), followed by an adenylation domain predicted to bind threonine ($A_1$); a
274 peptidyl carrier protein domain ($PCP_1$); a second condensation domain ($C_2$); an adenylation
275 domain predicted to bind serine ($A_2$); a second peptidyl carrier protein ($PCP_2$); and a thioesterase
276 domain (TE) (**Figure 3B**). On the basis of this annotation, which drastically differs from the
277 prediction based in the flavopeptin system, we predicted that this NRPS would produce an
278 acylated dipeptide, which is released upon the action of the thioesterase domain. This biosynthetic
279 logic is indeed consistent with the early biochemical data available for leupeptin (36–39).

280 We therefore renamed the EvoMining hit as *leupB*, and the NRPS as *leupA*, and a
281 functional association between LeupA and LeupB was assumed. Specifically, we speculated that
282 LeupB is capable of condensing the dipeptide produced by LeupA, possibly acyl-Leu-Leu, with an
283 arginine residue, leading to leupeptidic acid or acyl-Leu-Leu-Arg peptide (**Figure 3C**). In
284 accordance with this hypothesis the total predicted mass of LeupA and LeupB is 318 kDa, which
285 is strikingly close to that mass of 320 kDa early on estimated by Umezawa and co-workers for the
286 complex directing leupeptidic acid formation (39). Additional genes downstream of *leupB*,
287 transcribed in the same direction and therefore possibly involved in leupeptin biosynthesis include
288 *leupC* and *leupD*, are annotated as a threonine kinase and cysteine synthase, respectively (**Figure**
289 **3B**). The relevance of this functional annotation remains to be further investigated, but may have
290 to do with reduction of the leupeptidic acid.

To demonstrate that the predicted locus is involved in leupeptin biosynthesis, we used insertional mutagenesis to disrupt the *leupA* gene. We chose this rather simple approach due to the lack of genetic tools for manipulation of *S. roseus*, and because this gene could be considered to be essential for synthesis of leupeptin. We did not target *leupB*, as the central ASL homolog may complement its function via enzyme promiscuity, masking the expected phenotype (unless a double mutant is obtained), as previously reported in other BGCs (23). The *S. roseus leupA* mutant, termed PCMSr1, was obtained, and after comparative LC-MS analysis of this mutant with the parental wild-type strain, it was found that mutation of *leupA* renders PCMSr1 unable to produce leupeptin. The peak absent from the LC chromatograms obtained from PCMSr1, present in the wild type strain, corresponds with leupeptin authentic standard, as confirmed after electrospray ionization (ESI) mass spectrometry (m/z 427) and fragmentation pattern ($ms^2$ of 367 and 409) (**Figures S4A** and **S5**).

To further establish a link between leupeptin biosynthesis and the postulated locus, currently involving *leupA-D* (**Figure 3**), we constructed a genomic library from which two clones, containing at least these four genes, were isolated. Both constructs were introduced into *E. coli* DH10B, and the resulting transformants were used for comparative fermentation experiments. LC-MS analysis of these cultures revealed that extracts of supernatants from both strains presented fractions with peaks with same retention times and expected mass as authentic leupeptin standard (**Figures S4B** and **S5**). Therefore, we concluded that this locus, as predicted using EvoMining, is indeed directing the synthesis of leupeptin in *S. roseus* ATCC31245. However, given that the constructs bearing the *leupABCD* genes are 65 to 70 Kbp in size (clone 8_10B and 9_18N, respectively), the involvement of other genes in leupeptin biosynthesis, which are included in these constructs, cannot be ruled out at current time.

## Discovery of an arseno-organic metabolite in *Streptomyces lividans* and *Streptomyces coelicolor*: novel chemistry in model organisms.

For the second case, we aimed to identify a novel NP in the model organisms *S. coelicolor* A3(2) and *S. lividans* 66, which have been mined thoroughly and presumably most of their NP repertoire has been elucidated (45). Furthermore, several methods for genetic manipulation for these two strains are available (46), making these organisms ideal for these proof-of-concept experiments. EvoMining hits for these two strains include recruitments belonging to the 3-phosphoshikimate-1-carboxyvinyltransferase enzyme family, which was used as an example in Figure 1B.This enzyme, or AroA, catalyzes the transfer of a vinyl group from phosphoenolpyruvate (PEP) to 3-phosphoshikimate forming 5-enolpyruvylshikimate-3-phosphate and releasing phosphate. The reaction is part of the shikimate pathway, a common pathway for the biosynthesis of aromatic amino acids and other metabolites (47).

The phylogenetic reconstruction of the actinobacterial AroA enzyme family (**Fig 1B** and **Tree S4**), as expected, shows a major clade associated with central metabolism; this clade includes SLI_5501 from *S. lividans* and SCO5212 from *S. coelicolor*. The phylogeny also has a divergent clade that includes two family members linked to the BGCs of the polyketide asukamycin (48) and phenazines (49), as well as AroA homologs from 26 % of the genomes in the database. In *S. coelicolor* and *S. lividans* these recruited homologs are encoded by SLI_1096 and SCO6819, respectively. Moreover, in *S. lividans*, these orthologous genes are located within a large genomic island, SliGI-1, which has been functionally linked to metal homeostasis (50), and they are situated only six genes upstream of a two-gene PKS system spanning SCO6826-7 and SLI_1088-9, respectively. This PKS was identified in *S. coelicolor* since the early days of the genome mining of this organism, but often it was referred to as a cryptic BGC (51, 52). Indeed,

the divergent AroA homologs have not been associated with it. Furthermore, these homologs are classified as "other genes" when both genomes are mined using antiSMASH.

The gene neighborhood of these *aroA* genes is highly conserved between the genomes of *S. lividans* and *S. coelicolor*. Thus, from this point onwards we will refer to the *S. lividans* genes only. The syntenic region spans from SLI_1077 to SLI_1103, including several biosynthetic enzymes, regulators, transporters and the PKS, suggesting that these genes, together with other biosynthetic genes in this locus, are functionally linked and form a single BGC (**Figure 4A**). Detailed annotation of this BGC (**Table S5**) revealed the presence of a 2,3-bisphosphoglycerate-independent phosphoenolpyruvate mutase enzyme (SLI_1097; PPM), downstream and possibly transcriptionally coupled to the *aroA* homolog. Thus, a functional link between these genes, as well as with phosphonopyruvate decarboxylase gene (PPD; SLI_1091) encoded in this BGC, was proposed. The combination of mutase-decarboxylase enzymes is a conserved biosynthetic feature of NPs containing Carbon-Phosphate bonds (16).

Other non-enzymatic functions were found encoded within this BGC, including a set of ABC transporters, originally annotated as phosphonate transporters (SLI_1100 and SLI_1101). Four arsenic tolerance-related genes (SLI_1077-1080) located upstream of the PKS could also be annotated. These genes are paralogous to the main arsenic tolerance system encoded by the *ars* operon (53), which is located at the core of the *S. lividans* chromosome (SLI_3946-50). This BGC also codes for regulatory proteins, mainly arsenic responsive repressor proteins (SLI_1078, SLI_1092, SLI_1102 and SLI_1103). Thus, overall, our detailed annotation suggests a link between arsenic and phosphonate biosynthetic chemistry. Accordingly, in order to reconcile the presence of phosphonate-like biosynthetic, transporter and arsenic resistance genes, within a BGC,

we postulated a biosynthetic logic analogous to that of phosphonate biosynthesis, but involving arsenate as the driving chemical moiety.

The abovementioned hypothesis was further supported by the three following observations. First, arsenate and phosphate are similar in their chemical and thermodynamic properties, causing phosphate and arsenate utilizing enzymes to have overlapping affinities and kinetic parameters (54, 55). Second, previous studies have demonstrated that AroA is able to inefficiently catalyze a reaction in the opposite direction to the biosynthesis of aromatic amino acids, namely, the formation of PEP and 3-phosphoshikimate from enolpyruvyl shikimate 3-phosphate and phosphate (47). Indeed, arsenate and enolpyruvyl shikimate 3-phosphate can react to produce arsenoenolpyruvate (AEP), a labile analog of PEP, which is spontaneously broken down into pyruvate and arsenate (47). Third, it has been demonstrated that the phosphoenolpyruvate mutase, PPM, an enzyme responsible for the isomerization of PEP to produce phosphonopyruvate, is capable of recognizing AEP as a substrate. Although at low catalytic efficiency, the formation of 3-arsonopyruvate by this enzyme, a product analog of the phosphonopyruvate intermediate in phosphonate NPs biosynthesis (16), has been reported (56).

The previous evidence was used to postulate a novel biosynthetic pathway encoded by SLI_1077-SLI_1103. A putative arseno-organic product synthesized by this pathway may resemble the structural characteristics and properties of a phospolipid. A detailed functional annotation, and biosynthetic proposal, is provided as supplementary information (**Figure S6** and **Table S6**). To determine the product of the predicted BGC we used expression analysis, as well as comparative metabolic profiling of wild type and mutant strains, in both *S. lividans* and *S. coelicolor*. Using RT-PCR analysis, we first determined the transcriptional expression profiles of the PKS (SLI_1088), *aroA* (SLI_1096), one of the *arsR*-like regulator (SLI_1103), and the

383 periplasmic-binding protein of the ABC-type transporter (SLI_1099). As expected for a cryptic
384 BGC, the results of these experiments demonstrate that the proposed pathway is repressed under
385 standard laboratory conditions. We then analyzed the potential role of arsenate as an inducer of the
386 expression of this BGC, either alone or in combination with phosphate deprivation. Indeed, we
387 found that the analyzed genes were induced when *S. lividans* 66 was grown in the presence of 500
388 µM of arsenic and 3 µM of phosphate (**Figure 4B**).

389 In parallel, we used PCR-targeted gene replacement to produce mutants of the SLI_1096
390 and SCO6819 genes, and analyzed the phenotypes of the mutant and wild type strains on a
391 combined arsenate/phosphate gradient, i.e. low phosphate and high arsenate, and *vice versa*. After
392 this, we cultivated the wild type and mutant strains in liquid cultures, with and without arsenic,
393 during 14 days to obtain enough biomass for chemical analysis. Organic extracts from the pellets
394 of these cultures were analyzed using HPLC coupled with an ICP-MS calibrated to detect arsenic-
395 containing molecular species. Simultaneously, a high-resolution mass spectrometer determined the
396 mass over charge of the ions detected by the ICP. This set up allows for high-resolution detection
397 of arseno-organic metabolites (57). Using this approach, we detected the presence of arseno-
398 organic metabolites in the organic extracts from the pellets of both wild-type *S. coelicolor* and *S.*
399 *lividans*, with m/z of 331.1248, 333.1041, and 351.1147 (**Figure 4C** and **Figure S6**). These
400 metabolites could not be detected in either identical extracts from wild type strains grown in the
401 absence of arsenate or in the mutant strains deficient for the SLI_1096/SCO6819 genes. Thus, the
402 product of this pathway may be a relatively polar arsenolipid (**Figure S6**). The actual structures of
403 these products are still subject to further investigation and will be discussed in detail in a future
404 publication.

405

406

407 **EvoMining and its future impact into NP genome mining**

408 On one hand, confirmation of a link between SLI_1096/SCO6819 and the synthesis of an
409 arseno-organic metabolite provides an example on how genome-mining efforts, based in novel
410 enzyme sequences, can be advanced. For instance, co-occurrence of divergent SLI_1096
411 orthologs, now called arsenoenolpyruvate synthases (AEPS); arsenopyruvate mutase (APM) and
412 arsonopyruvate decarboxylase (APD), can be used as beacons to mine publically available
413 bacterial genomes. Indeed, thirteen BGCs with the potential to synthesize arseno-organic
414 metabolites, all of them encoded in genomes of myceliated *Actinobacteria*, were identified after
415 sequence similarity searches using the non-redundant GenBank database. The divergence and
416 potential chemical diversity within these arseno-related BGCs was characterized after a
417 phylogenetic analysis, using as matrix all conserved genes of these BGCs (**Figure 5**). This
418 analysis suggests three possible sub-classes with distinctive features, PKS-related, PKS-
419 independent and PKS/NRPS hybrid, which warrant further investigation.

420 On the other hand, once an EvoMining hit is validated as an enzyme with an
421 unprecedented function in the context of a hidden BGC, i.e. an EvoMining prediction, this could
422 lead to identification of novel classes of conserved BGCs. To illustrate this, we focused in the
423 AroA EvoMining tree (Figure 1B). The green branches within the NP-related clade of this tree
424 were manually curated in search for a conserved BGC (**Figure 1B**). One particular case was found
425 to appear frequently, and thus its genes were annotated in detail. This BGC was found to be
426 conserved in at least sixty-three *Streptmyces* genomes, and in the genome of *Microtetraspora*
427 *glauca* NRRL B-3735, included in the non-redundant GenBank database (**Table S8**).

For these analyzes, fifteen genes upstream and downstream the EvoMining AroA hit were extracted as before, and annotated on the basis of the locus from *S. griseolus* NRRL B-2925, as this organism provides a condensed version of this locus (**Table S7**). Indeed, this locus has some of the expected features for an NP BGC, including: (i) gene organization suggesting an assembly of operons, most of them transcribed in the same direction; (ii) genes encoding for enzymes, regulators and potential resistance mechanisms; and (iii) enzymes that have been found in other known NP BGCs. The latter observation, which actually includes seven genes out of thirty-one, present in an equal number of NP BGCs, actually confirms the NP nature of this locus. The reason why EvoMining did not lead to these homologs as expanded and recruited enzymes has to do with the fact that none of these enzyme families are included within the limited space explored by our PSCP database.

Thus, we conclude that EvoMining has great potential to ease natural product and drug discovery by means of accelerating the conceptual genome-mining loop that goes from novel enzymatic conversions, their sequences, and propagation after homology searches.

## Methods

### Bioinformatics

*Seed NP database*: NRPS and PKS BGCs were obtained from the DoBISCUIT and ClusterMine 360 databases (10, 11). BGCs for other NP classes were collected from available literature (**Table S1**). The database included amino acid fasta sequences from GenBank, DoBISCUIT and ClusterMine360. Annotated GenBank formatted files were downloaded from the GenBank database to assemble a database that included 226 BGCs. *Genome database:* Complete and draft genomes of 230 members of the *Actinobacteria* family (**Table S1**) were retrieved from the

GenBank either as single contigs or as groups of contigs in GenBank format, amino acid and DNA sequences were extracted from these files using in-house made scripts. *Precursor supply central pathway (PSCP) database*: the amino acid sequences from the proteins involved in central metabolism were obtained from a database that we assembled for a previous enzyme expansion assessment, published elsewhere (28). The final database for this work included a total of 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway and tricarboxylic acids cycle (**Table S2**).

*EvoMining pipeline*: The PSCPs database was used as query to retrieve PSCP enzyme families from the genome database using BlastP (65), with an e-value cutoff of 0.0001 and a score cutoff of 100. At least three query sequences, representing each of the GSMRs used as sources of these sequences (28), were used for Blast searches. The average number of homologs of each enzyme family per genome and the standard deviation were calculated to establish a cutoff to identify and highlight significant expansion events (**Table S4**). An enzyme family expansion was scored if the number of homologs in at least one genome was higher than the average number of homologs plus a standard deviation unit. Enzyme families with expansions were used for the next step of the analysis. To identify enzyme recruitments, the amino acid sequences of the seed NP database were used as queries for BlastP searches against expanded enzyme families identified in the previous step using an e-value cutoff of 0.0001 and a bit score cutoff of 100. These parameters were consistent with the approach used by EvoMining, as confirmed heuristically.

The homologs found in known BGCs were added as seeds to the sequences from the expanded enzyme families with recruitments for future clade identification and labeling. These sets of sequences were aligned using Muscle version 3.8.31 (58). The alignments were inspected and curated manually using JalView (59). The curated alignments were used for phylogenetic reconstructions, which were estimated using MrBayes (60) with the following parameters:

aamodelpr=mixed, samplefreq=100, burninfrac=0.25 in four chains and for 1000000 generations. The bidirectional best hits with the sequences in the PSCP database were identified and tagged using in-house scripts to distinguish PSCP orthologs from other homologs that result from expansion events. The gene identifiers (GIs) of the EvoMining hits were used as queries to retrieve their genome context as DNA regions of approximately 80 Kbs including the EvoMining hit coding genes. These contigs were retrieved from the genomes in GenBank format using an in-house script. These contigs were annotated using antiSMASH (29) through its web interface. The whole process was executed semi-automatically using in-house scripts written in Perl.

**Mutagenesis analysis**

*S. coelicolor* (SCO6819) and *S. lividans* 66 (SLI_1096) knock-out mutants were constructed using in-frame PCR-targeted gene replacement of their coding sequences with an apramycin resistance cassette (*acc(3)IV*) (61). The plasmid pIJ773 was used as template to obtain a mutagenic cassette containing the apramycin resistance marker by PCR amplification with the primers reported in **Table S9**. The mutagenic cassettes were used to disrupt the coding sequences of the genes of interest from the cosmid clone 1A2 that spans from SCO6971 to SCO6824 (62). Given the high sequence identity between the regions covered by cosmid 1A2 with the orthologous region in *S. lividans*, this cosmid clone was also used for disruption of SLI_1096. The gene disruptions were performed using the Redirect system reported elsewhere (61). Double cross-over ex-conjugants were selected using apramycin resistance and kanamycin sensitivity as phenotypic markers. The genotype of the clones was confirmed by PCR. The strains and plasmids of the Redirect system were obtained from the John Innes Centre (Norwich, UK).

The *S. roseus leupA* mutant was constructed following an insertional mutagenesis strategy. For this purpose, a fragment of 640 bp was amplified by PCR and cloned in the vector pCR2.1-TOPO (ampicillin/ kanamycin resistance) using a TA cloning kit from Invitrogen (Carlsbad, USA), to produce the suicide plasmid pLEUPA that cannot be replicated in *S. roseus*. This plasmid was introduced into *S. roseus* via protoplasts, generated following standard protocols. The transformants were selected using kanamycin (50µ/mL) and the genotype of the insertional mutants was confirmed by PCR.

**Transcriptional analysis**

The *S. lividans* 66 wild type strain was grown on 0 and 3; 0 and 300; 500 and 3; 500 and 300 µM of Arsenate and $KH_2PO_4$ respectively in solid modified R5 media for eight days. The complete culture conditions used in this work are further detailed in a following section. Mycelium collected from plates was used for RNA extraction with a NucleoSpin RNA II kit (Macherey-Nagel). The RNA samples were used as template for RT-PCR using the one step RT-PCR kit (Qiagen) (2ng RNA template for each 40 µl reaction). The housekeeping sigma factor *hrdB* (SLI_6088) was used as a control.

***S. roseus* genome sequencing and library construction**

*S. roseus* ATCC31245 was obtained from the ATCC collection, and its genomic DNA was extracted using common protocols (46) and sequenced at the genomic sequencing facilities of Langebio, Cinvestav-IPN (Irapuato, Mexico), using an Illumina MiSeq platform in paired-end format with read lengths of 250 bases and insert length of 800 bases. In total, 721 Mbp of

sequence was obtained. The raw reads were filtered using Trimmomatic (63) and assembled with velvet (64), obtaining a 7.8 Mb assembly in 165 contigs with a coverage of 95 X and a GC content of 72 %. This assembly was annotated using RAST (65) antiSMASH (29) and EvoMining. A genomic library of *S. roseus* ATCC31245 was further obtained for cloning into the pESAC13A vector with an average insert length of 70 Kbps (Bio S&T, Montreal, Canada). pESCA13A is a derivative from pPAC-S1 (66), which has an apramycin resistance as selection marker. This library was screened for the *leup* locus by PCR, leading two clones named 9_18N and 8_10B, containing the desired region. As described further, these constructs were used for heterologous expression in *E. coli*.

**LC-MS metabolite profile analysis**

The SLI_1096 and SCO6818 minus mutants were grown on modified R5 medium ($K_2SO_4$ 0.25 gr; $MgCl_2$-$6H_20$ 10.12 gr; glucose 10 gr; casamino acids 0.1 gr; TES buffer 5.73 gr; trace element solution (46) 2 ml; agar 20gr) supplemented with a gradient of $KH_2PO_4$ and $Na_3AsO_4$ ranging from 3 to 300 µM and 0 to 500 µM, respectively. Induction of the arseno-organic BGC in both strains was detected in the condition were phosphate is limited and arsenic is available. Therefore, modified R5 liquid media supplemented with 3µM $KH_2PO_4$ and 500 µM $Na_3AsO_4$ was used for production of arseno-organic metabolites, and the cultures were incubated for 14 days in shaken flasks with metal springs for mycelium dispersion at 30 C. The mycelium was obtained by filtration, and the filtered mycelium was washed thoroughly with deionized water and freeze-dried. The samples were extracted overnight twice with MeOH/DCM (1:2). The extracts were combined and evaporated to dryness, and the dry residues were re-dissolved in 1 mL of MeOH (HPLC-Grade) and injected to the HPLC. The detection of organic arsenic species was achieved

by online-splitting of the HPLC-eluent with 75% going to ESI-Orbitrap MS (Thermo Orbitrap Discovery) for accurate mass analysis and 25% to ICP-QQQ-MS (Agilent 8800) for the detection of arsenic. For HPLC, an Agilent Eclipse XDB-C18 reversed phase column was used with a $H_2O$ /MeOH gradient (0-20 min: 0-100% MeOH; 20-45 min: 100% MeOH; 45-50 min: 100% $H_2O$). The ICP was set to oxygen mode and the transition $^{75}As^{+}$ -> $(^{75}As^{16}O)^{+}$ (Q1: m/z = 75, Q2: m/z = 91) was observed. The correction for carbon enhancement from the gradient was achieved using a mathematical approach as described previously (57). The ESI-Orbitrap-MS was set to positive ion mode in a scan range from 250-1100 amu. Also, $MS^2$-spectra for the major occurring ions were generated.

For native leupeptin production wild type *S. roseus* and *Leup*A, mutants were grown in leupeptin production media (67) containing: Glucose 3gr; $NH_4NO3$ 0.5gr; MgSO4 ($7H_2O$) 0.5gr; KCl 0.05 gr; L-leucine 0.75 gr; L-arginine 0.75 gr; glycine 0.75 gr; casaminoacids 0.1 gr; yeast extract 0.4 gr per liter. These cultures were set up in shaken flasks with metal springs for 48 hours at 30 C. For heterologous production of leupeptins, *E. coli* DH10B transformants carrying the 9_18N and 8_10B PAC clones were grown in Luria-Bertani media (LB) in shaken flasks with 50 µg per mL of apramycinm at 37 C for 48 hours. The cultures were centrifuged and the supernatants freeze-dried to obtain 10X concentrates. The crude extracts were analyzed using a C18-218TP vydac column (Grace Healthcare; Columbia, USA) or Restek C18 column (Restek Chromatography; Bellefonte, US), with a 0-100% gradient of [triflouroacetic acid 0.01% in water]-acetonitrile, and detected by diode array (DAD) at λ=210 nm. Leupeptin authentic standard (L2884, Sigma-Aldrich, St Louis, USA) was used as reference and the peaks with equivalent retention times from the extracts were collected for MS analysis performed on an ion trap LTQ-

563 VELOS equipment in positive mode (Thermo scientific, Waltham, USA) at the MS Unit of
564 Unidad Irapuato Cinvestav-IPN (Irapuato, Mexico).

## Acknowledgements

We are indebted with Angélica Cibrián-Jaramillo, Marnix Medema, Paul Straight and Sean Rovito, for useful discussions and critical reading of the manuscript, as well as with Alicia Chagolla and Yolanda Rodriguez of the MS Unit of unidad Irapuato, Cinvestav, for analytical services. This work was funded by Conacyt Mexico (grants No. 179290 and 177568) and FINNOVA Mexico (grant No. 214716) to FBG. PCM was funded by Conacyt scholarship (No. 28830) and a Cinvestav postdoctoral fellowship. JF and JFK acknowledge funding from the College of Physical Sciences, University of Aberdeen, UK.

## Competing interests

PCM, CEMG, MAME, HERA and FBG have filed patent applications related to this work. The other authors declare that no competing interests exist.

## References


1. Schreiber S (2005) Small molecules: the missing link in the central dogma. *Nature Chemical Biology* 1:64–66. DOI: 10.1038/nchembio0705-64

2. Bachmann BO, Van Lanen SG, Baltz RH (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol* 41:175–84. DOI: 10.1007/s10295-013-1389-9

3. Demain AL (2014) Importance of microbial natural products and the need to revitalize their discovery. *J Ind Microbiol Biotechnol* 41:185–201. DOI: 10.1007/s10295-013-1325-z

4. *Antimicrobial resistance: global report on surveillance 2014* (World Health Organization-UN). ISBN**:** 9789241564748.

5. Harvey A, Edrada-Ebel R, Quinn R (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* 14:111–129. DOI:10.1038/nrd4510

6. Bachmann BO, McAlpine JB, Zazopoulos E (2006) Farnesyl dibenzodiazepinone, and processes for its production. *US Patent CA2466340* A.

7. McAlpine J et al. (2008) Biosynthesis of diazepinomicin/ECO-4601, a Micromonospora secondary metabolite with a novel ring system. *Journal of natural products* 71:1585–90. DOI: 10.1021/np800376n

8. Gourdeau H et al. (2007) Identification, characterization and potent antitumor activity of ECO-4601, a novel peripheral benzodiazepine receptor ligand. *Cancer Chemotherapy and Pharmacology* 61:911921. DOI: 10.1007/s00280-007-0544-2

9. Jensen PR, Chavarria KL, Fenical W, Moore BS, Ziemert N (2014) Challenges and triumphs to genomics-based natural product discovery. *J Ind Microbiol Biotechnol* 41:203–9. DOI: 10.1007/s10295-013-1353-8

10. Conway K, Boddy C (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Research* 41:D402–D407. DOI: 10.1093/nar/gks993

11. Ichikawa N et al. (2013) DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. *Nucleic acids research* 41:D408–14. DOI: 10.1093/nar/gks1177

12. Barona-Gómez F, Wong U, Giannakopulos A, Derrick P, Challis G (2004) Identification of a cluster of genes that directs desferrioxamine biosynthesis in Streptomyces coelicolor M145. *Journal of the American Chemical Society* 126:1628216283. DOI: 10.1021/ja045774k

608 13. Lautru S, Deeth R, Bailey L, Challis G (2005) Discovery of a new peptide natural product by
609 Streptomyces coelicolor genome mining. *Nature Chemical Biology* 1:265–269. DOI:
610 10.1038/nchembio731

611 14. Udwary D et al. (2007) Genome sequencing reveals complex secondary metabolome in the
612 marine actinomycete Salinispora tropica. *Proceedings of the National Academy of Sciences*
613 104:10376–10381. DOI: 10.1073/pnas.0700962104

614 15. Challis G (2008) Mining microbial genomes for new natural products and biosynthetic
615 pathways. *Microbiology (Reading, England)* 154:1555–69. DOI:
616 10.1099/mic.0.2008/018523-0

617 16. Metcalf W, Donk W (2009) Biosynthesis of phosphonic and phosphinic acid natural products.
618 *Biochemistry* 78:65–94. DOI: 10.1146/annurev.biochem.78.091707.100215

619 17. Yu X et al. (2013) Diversity and abundance of phosphonate biosynthetic genes in nature.
620 *Proceedings of the National Academy of Sciences* 110:20759–20764. DOI:
621 10.1073/pnas.1315107110

622 18. Ju K-S, Doroghazi J, Metcalf W (2013) Genomics-enabled discovery of phosphonate natural
623 products and their biosynthetic pathways. *Journal of Industrial Microbiology &*
624 *Biotechnology*. 41:345-56 doi: 10.1007/s10295-013-1375-2

625 19. Arnison P et al. (2012) Ribosomally synthesized and post-translationally modified peptide
626 natural products: overview and recommendations for a universal nomenclature. *Natural*
627 *Product Reports* 30:108–160. DOI: 10.1039/c2np20085f

628 20. Thaker M et al. (2013) Identifying producers of antibacterial compounds by screening for
629 antibiotic resistance. *Nature Biotechnology* 31:922–927. DOI: 10.1038/nbt.2685

630 21. Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically
631 diverse superfamilies and functionally distinct suprafamilies. *Annual review of biochemistry*
632 70:209–46. DOI: 10.1146/annurev.biochem.70.1.209

633 22. Caetano-Anollés G et al. (2009) The origin and evolution of modern metabolism. *Int J*
634 *Biochem Cell Biol* 41:285–97. DOI: 10.1016/j.biocel.2008.08.022

635 23. Tahlan K, Park HU, Wong A, Beatty PH, Jensen SE (2004) Two sets of paralogous genes
636 encode the enzymes involved in the early stages of clavulanic acid and clavam metabolite
637 biosynthesis in Streptomyces clavuligerus. *Antimicrob Agents Chemother* 48:930–9. DOI:
638 10.1128/AAC.48.3.930-939.2004

639 24. Verdel-Aranda K, López-Cortina ST, Hodgson DA, Barona-Gómez F (2015) Molecular
640 annotation of ketol-acid reductoisomerases from Streptomyces reveals a novel amino acid
641 biosynthesis interlock mediated by enzyme promiscuity. *Microb Biotechnol* 8:239–52. DOI:
642 10.1111/1751-7915.12175

643 25. Vining LC (1992) Secondary metabolism, inventive evolution and biochemical diversity--a
644 review. *Gene* 115:135–40. DOI: 10.1016/0378-1119(92)90551-Y

645 26. Firn R, Jones C (2009) A Darwinian view of metabolism: molecular properties determine
646 fitness. *Journal of Experimental Botany* 60:719–726. DOI: 10.1093/jxb/erp002

647 27. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA (2014) A systematic
648 computational analysis of biosynthetic gene cluster evolution: lessons for engineering
649 biosynthesis. *PLoS Comput Biol* 10:e1004016.

650 28. Barona-Gómez F, Cruz-Morales P, Noda-García L (2012) What can genome-scale metabolic
651 network reconstructions do for prokaryotic systematics? *Antonie Van Leeuwenhoek* 101:35–
652 43. DOI: 10.1007/s10482-011-9655-1

653 29. Medema M et al. (2011) antiSMASH: rapid identification, annotation and analysis of
654 secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences.
655 *Nucleic Acids Research* 39:W339–W346. DOI: 10.1093/nar/gkr466

656 30. Cimermancic P et al. (2014) Insights into secondary metabolism from a global analysis of
657 prokaryotic biosynthetic gene clusters. *Cell* 158:412–21. DOI: 10.1016/j.cell.2014.06.034

658 31. Blodgett JA et al. (2007) Unusual transformations in the biosynthesis of the antibiotic
659 phosphinothricin tripeptide. *Nat Chem Biol* 3:480–5. DOI:10.1038/nchembio.2007.9

660 32. Blodgett J, Zhang J, Metcalf W (2005) Molecular Cloning, Sequence Analysis, and
661 Heterologous Expression of the Phosphinothricin Tripeptide Biosynthetic Gene Cluster from
662 Streptomyces viridochromogenes DSM 40736. *Antimicrobial Agents and Chemotherapy*
663 49:230–240. DOI: 10.1128/AAC.49.1.230-240.2005

664 33. Schwartz D et al. (2004) Biosynthetic Gene Cluster of the Herbicide Phosphinothricin
665 Tripeptide from Streptomyces viridochromogenes Tü494. *Applied and Environmental*
666 *Microbiology* 70:7093–7102. DOI: 10.1128/AEM.70.12.7093-7102.2004

667 34. Blin K et al. (2013) antiSMASH 2.0-a versatile platform for genome mining of secondary
668 metabolite producers. *Nucleic acids research* 41:W204–12. DOI: 10.1093/nar/gkt449

669 35. Aoyagi T, Takeuchi T, Matsuzaki A, Kawamura K, Kondo S (1969) Leupeptins, new protease
670 inhibitors from Actinomycetes. *The Journal of antibiotics* 22:283–6. DOI:
671 http://doi.org/10.7164/antibiotics.22.283

672 36. Suzukake K, Fujiyama T, Hayashi H, Hori M (1979) Biosynthesis of leupeptin. II. Purification and
673 properties of leupeptin acid synthetase. *The Journal of antibiotics* 32:523-30. DOI:
674 http://doi.org/10.7164/antibiotics.22.283

675 37. Hori M, Hemmi H, Suzukake K, Hayashi H (1978) Biosynthesis of leupeptin. *The Journal of*
676 *antibiotics* 31:95-8. DOI: http://doi.org/10.7164/antibiotics.31.95

677 38. Suzukake K, Hori M, Tamemasa O, Umezawa H (1981) Purification and properties of an enzyme
678 reducing leupeptin acid to leupeptin. *Biochim Biophys Acta*. 661:175-81. DOI: 10.1016/0005-
679 2744(81)90001-2

680 39. Suzukake K, Hayashi H, Hori M (1980) Biosynthesis of leupeptin. III. Isolation and properties of an
681 enzyme synthesizing acetyl-L-leucine. *The Journal of antibiotics* 33:857-62. DOI:
682 http://doi.org/10.7164/antibiotics.33.857

683 40. Chen Y, McClure R, Zheng Y, Thomson R, Kelleher N (2013) Proteomics guided discovery
684 of flavopeptins: anti-proliferative aldehydes synthesized by a reductase domain-containing
685 non-ribosomal peptide synthetase. *Journal of the American Chemical Society* 135:10449–56.
686 DOI: 10.1021/ja4031193

687 41. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH (2007) Phylogenetic analysis of
688 condensation domains in NRPS sheds light on their functional evolution. *BMC evolutionary*
689 *biology* 7:78. DOI: 10.1186/1471-2148-7-78

690 42. Sampaleanu LM, Vallée F, Thompson GD, Howell PL (2002) Three-dimensional structure of
691 the argininosuccinate lyase frequently complementing allele Q286R. *Biochemistry* 40:15570–
692 80. DOI: 10.1021/bi011525m

693 43. Kaysser L et al. (2011) Identification of a napsamycin biosynthesis gene cluster by genome
694 mining. *Chembiochem* 12:477–87. DOI: 10.1002/cbic.201000460

695 44. Zhang W, Ostash B, Walsh CT (2010) Identification of the biosynthetic gene cluster for the
696 pacidamycin group of peptidyl nucleoside antibiotics. *Proceedings of the National Academy*
697 *of Sciences of the United States of America* 107:16828–33. DOI: 10.1073/pnas.1011557107

698 45. Challis GL (2014) Exploitation of the Streptomyces coelicolor A3(2) genome sequence for
699 discovery of new natural products and biosynthetic pathways. *J Ind Microbiol Biotechnol*
700 41:219–32. DOI: 10.1007/s10295-013-1383-2

701 46. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA (2000) *Practical Streptomyces*
702 *genetics* (The John Innes Foundation, Norwich UK). ISBN 0-7084-0623-8

703 47. Zhang F, Berti PJ (2006) Phosphate analogues as probes of the catalytic mechanisms of MurA
704 and AroA, two carboxyvinyl transferases. *Biochemistry* 45:6027–37. DOI:
705 10.1021/bi0601914

706 48. Rui Z et al. (2010) Biochemical and genetic insights into asukamycin biosynthesis. *The*
707 *Journal of biological chemistry* 285:24915–24. DOI: 10.1074/jbc.M110.128850

708 49. Seeger K et al. (2011) The biosynthetic genes for prenylated phenazines are located at two
709 different chromosomal loci of Streptomyces cinnamonensis DSM 1042. *Microbial*
710 *biotechnology* 4:252–62. DOI: 10.1111/j.1751-7915.2010.00234.x

50. Cruz-Morales P et al. (2013) The genome sequence of Streptomyces lividans 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island. *Genome Biology and Evolution* 5:1165–1175. DOI: 10.1093/gbe/evt082

51. Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* 26:1362–84. DOI: 10.1039/B817069J

52. Bentley SD et al. (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 417:141–7. DOI: 10.1038/417141a

53. Wang L et al. (2006) arsRBOCT arsenic resistance system encoded by linear plasmid pHZ227 in Streptomyces sp. strain FR-008. *Applied and environmental microbiology* 72:3738–42. DOI: 10.1128/AEM.72.5.3738-3742.2006

54. Elias M et al. (2012) The molecular basis of phosphate discrimination in arsenate-rich environments. *Nature* 491:134–137. DOI: 10.1038/nature11517

55. Tawfik D, Viola R (2011) Arsenate replacing phosphate: alternative life chemistries and ion promiscuity. *Biochemistry* 50:1128–34. DOI: 10.1021/bi200002a

56. Chawla S, Mutenda EK, Dixon HB, Freeman S, Smith AW (1995) Synthesis of 3-arsonopyruvate and its interaction with phosphoenolpyruvate mutase. *The Biochemical journal* 308 ( Pt 3):931–5.

57. Amayo KO, Raab A, Krupp EM, Gunnlaugsdottir H, Feldmann J (2013) Novel identification of arsenolipids using chemical derivatizations in conjunction with RP-HPLC-ICPMS/ESMS. *Analytical chemistry* 85:9321–7. DOI: 10.1021/ac4020935

58. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32:1792–7. DOI: 10.1093/nar/gkh340

59. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)* 25:1189–91. DOI: 10.1093/bioinformatics/btp033

60. Ronquist F et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61:539–42. DOI: 10.1093/sysbio/sys029

61. Gust B, Challis G, Fowler K, Kieser T, Chater K (2003) PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proceedings of the National Academy of Sciences* 100:1541–1546. DOI: 10.1073/pnas.0337542100

743 62. Redenbach M et al. (1996) A set of ordered cosmids and a detailed genetic and physical map
744 for the 8 Mb Streptomyces coelicolor A3(2) chromosome. *Molecular microbiology* 21:77–96.
745 DOI: 10.1046/j.1365-2958.1996.6191336.x

746 63. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
747 sequence data. *Bioinformatics (Oxford, England)* 30:2114–20. DOI:
748 10.1093/bioinformatics/btu170

749 64. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de
750 Bruijn graphs. *Genome research* 18:821–9. DOI: 10.1101/gr.074492.107

751 65. Aziz RK et al. (2008) The RAST Server: rapid annotations using subsystems technology.
752 *BMC genomics* 9:75. DOI: 10.1186/1471-2164-9-75

753 66. Sosio M et al. (2000) Artificial chromosomes for antibiotic-producing actinomycetes. *Nat*
754 *Biotechnol* 18:343–5. DOI:10.1038/73810

755 67. Takagi K, Yamamoto Y, Yamazaki T, Yamaguchi H, Umezawa H (1978) Process for
756 producing l-leupeptins. *US patent 4066507 A*

## Figures

**Figure 1. EvoMining pipeline for the recapitulation of the evolution of NP biosynthesis**. **A.** Bioinformatic workflow: the three input databases, as discussed in the text, are shown in green. Internal databases are shown in yellow, whereas grey boxes depict processes. **B.** An example of a typical EvoMining phylogenetic tree (**Tree S4**) using the case of 3-carboxyvinyl-phosphoshikimate synthase family. Red branches include homologs related to central metabolism and their topology resembles that of a species guide tree (**Tree S1**), while blue branches have been recruited into known BGCs. Cyan braches are EvoMining hits found within regions recognised as NP-related also by antiSMASH or ClusterFinder. Green branches are not classifiable by other methods and thus represent EvoMining predictions that may form part of BGCs for novel classes of NPs (see **Figure 2A** for further details).

**Figure 2. Analysis of EvoMining Hits. A**. Pie chart of the whole set of EvoMining hits as annotated using antiSMASH and ClusterFinder. **B.** Diversity of BGCs per recruited enzyme family. Top panel, the number of hits and BGC classes per family are compared. As the number of hits increases, more BGC classes are found. Bottom panel, a diversity plot for each enzyme family, showing the proportion and number of BGC classes as defined by AntiSMaSH. The label "Detected by ClusterFinder" means EvoMining hits that are also found by this algorithm, whereas the label "EvoMining predictions" includes all hits that could not be detected by antiSMASH or Cluster Finder.

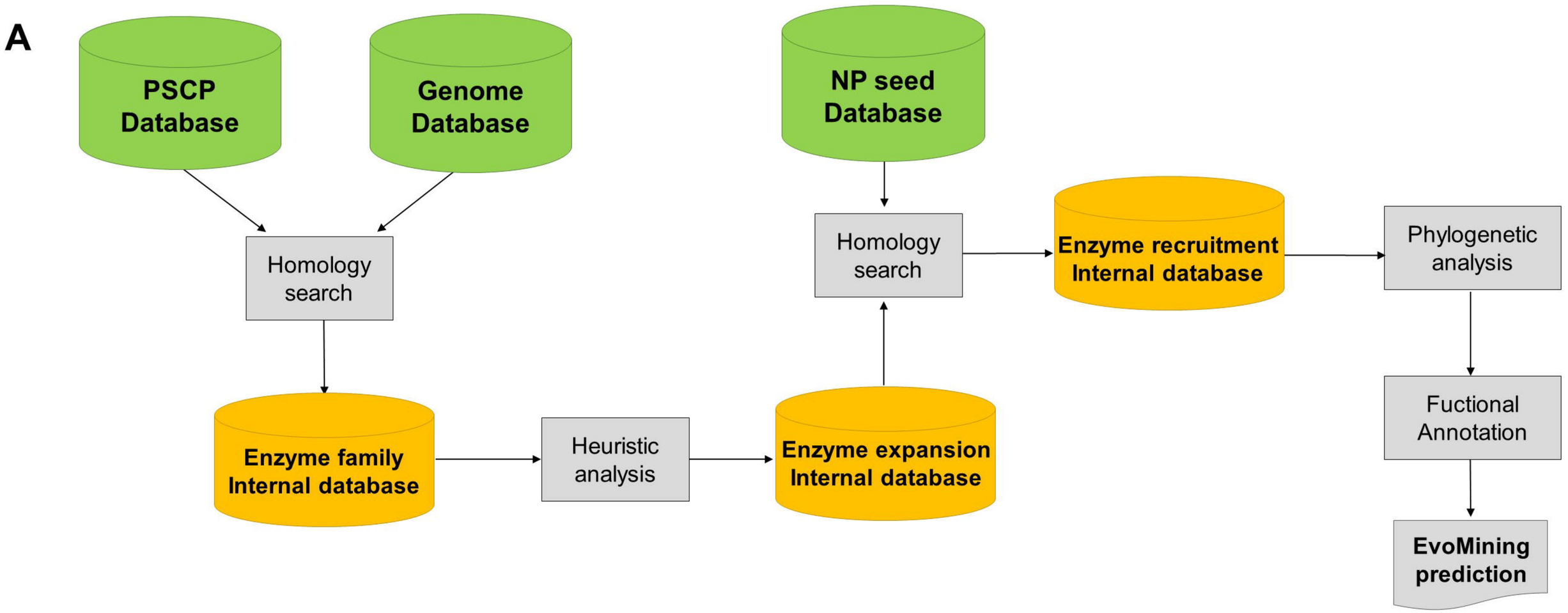
779 **Figure 3. Discovery of the BGC for leupeptin. A.** Left panel, phylogenetic reconstruction of the
780 actinobacterial argininosuccinate lyase enzyme family (**Tree S3**). Homologs related to central
781 metabolism are shown in red branches. A clade including recruited homologs is shown in cyan.
782 The LeupB homolog from *S. roseus*, a leupeptin producer, together with the known recruitments
783 for Pacidamycin and Napsamycin (blue branches), are indicated. **B.** Genome context of *leupB*,
784 including *leupA* (novel NRPS), *leupC* (annotated as threonine kinase) and *leupD* (annotated as
785 cysteine synthase). **C**. Biosynthetic proposal for leupeptin, based in the EvoMining prediction and
786 earlier biochemical data, as discussed within the text. LeupA is proposed to produce Acyl-Leu-
787 Leu, which is used by LeupB, together with arginine, for the synthesis of leupeptidic acid. A
788 reductase activity, that remains to be identified, is required for formation of the characteristic
789 aldehyde group.

790

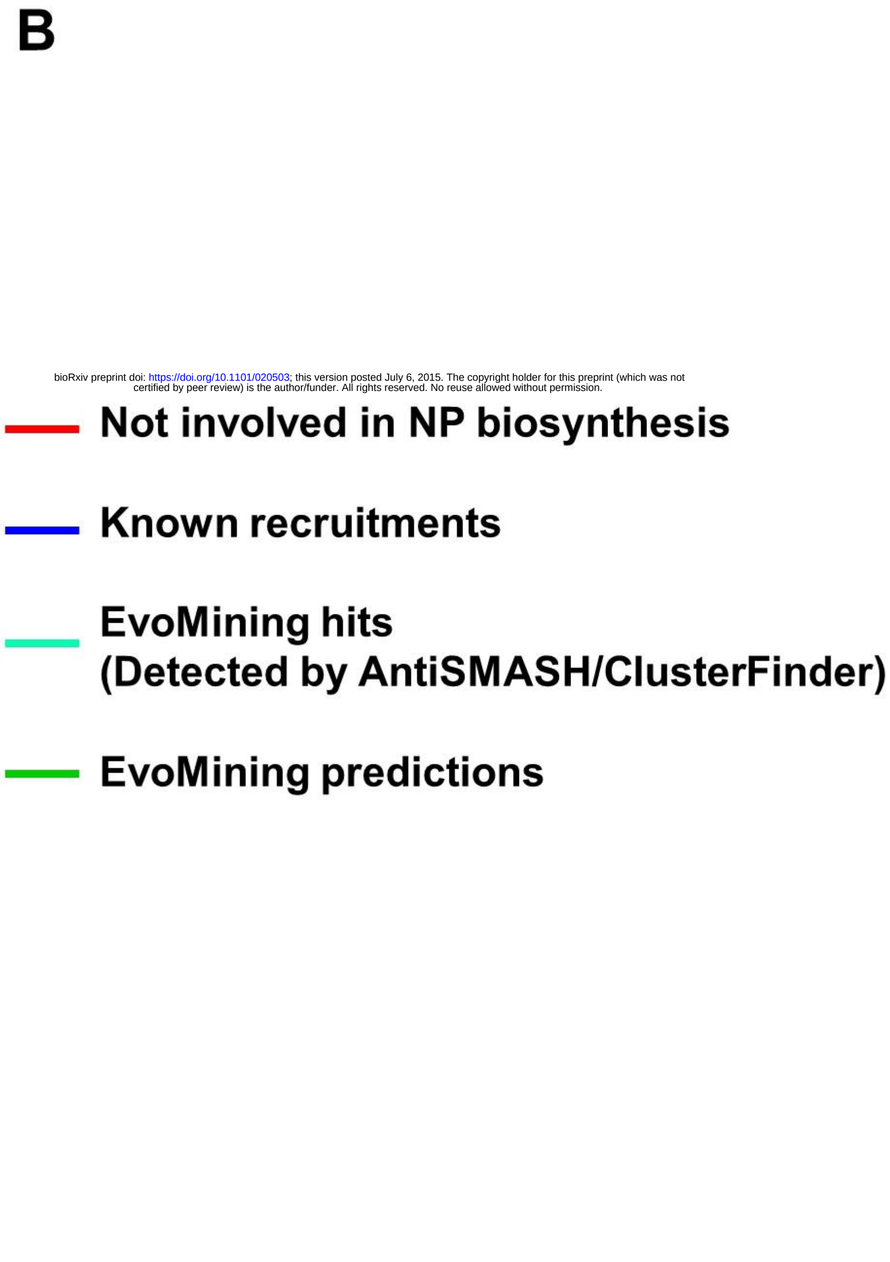
791 **Figure 4. Discovery of a BGC for arseno-organic NPs in *S. coelicolor* and *S. lividans***. **A.** The
792 conserved arseno-organic BGC in *S. lividans* 66 and *S. coelicolor* is shown. The proposed
793 biosynthetic logic for early intermediates in the biosynthesis of arseno-organic metabolites is
794 shown. **B.** Transcriptional analysis of selected genes within the arseno-organic BGC, showing that
795 gene expression is repressed under standard conditions, but induced upon the presence of arsenate.
796 **C.** HPLC-Orbitrap/QQQ-MS trace of organic extracts from mycelium of wild type and the
797 SLI_1096 mutant showing the detection of arsenic-containing species. Three m/z signals were
798 detected within the two peaks found in the trace from the wild type strain grown on the presence
799 of arsenate. These m/z signals are absent from the wild type strain grown without arsenate, and
800 from the SLI_1096 mutant strain grown on phosphate limitation and the presence of arsenate.
801 Identical results were obtained for *S. coelicolor* and the SCO6819 mutant

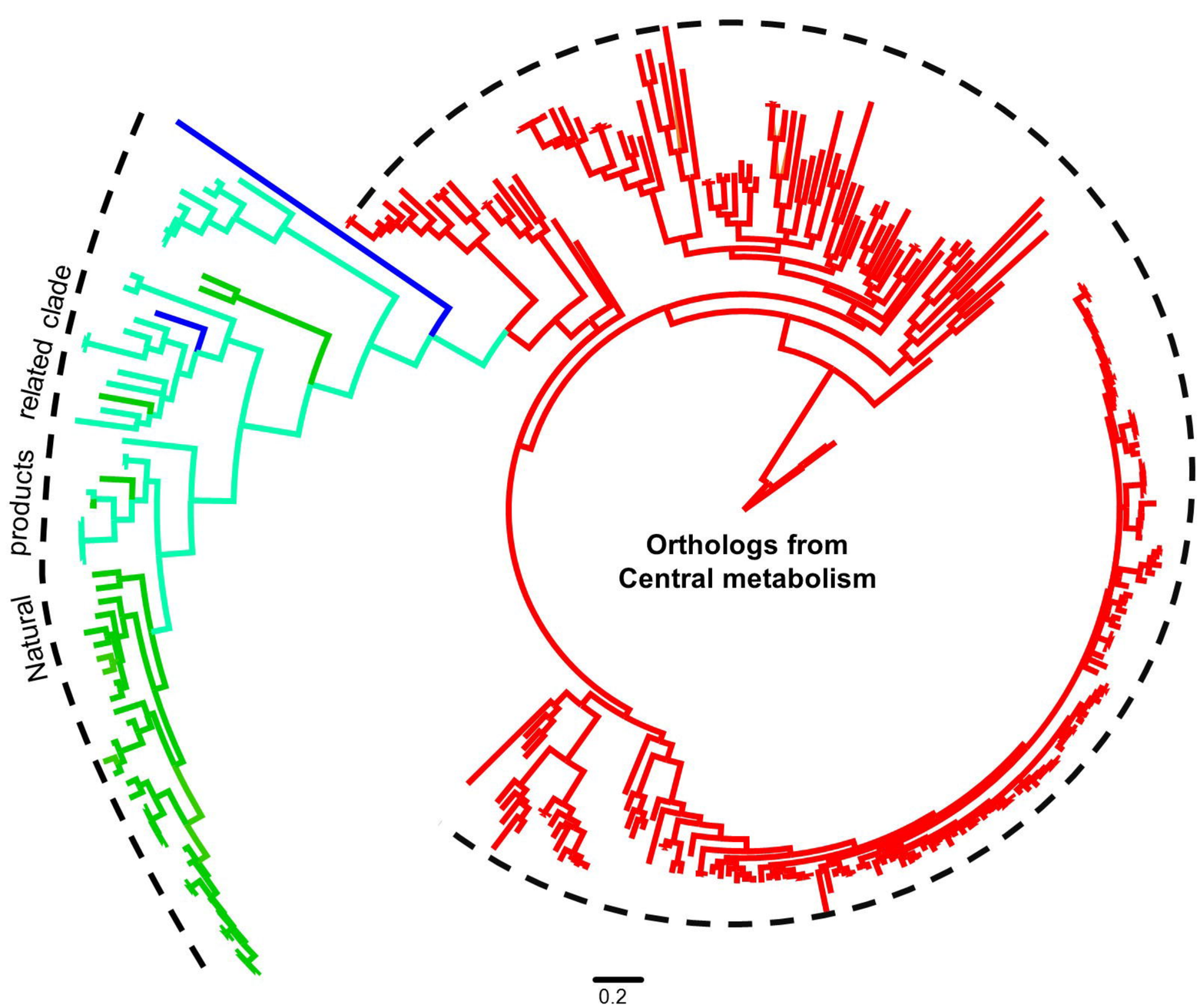
**Figure 5. Novel BGCs for arseno-organic metabolites found in *Actinobacteria.*** The BGCs were found by mining for the co-occurrence of arsenoenol pyruvate synthase (AEPS), arsenopyruvate mutase (APM), arsenoenolpyruvate decarboxylase (APD), in available bacterial genomes from the GenBank, as of November 2014. Related BGCs were only found in actinomycetes. The phylogeny was constructed with a concatenated matrix of conserved enzymes among the BGCs that included AEPS, APM, APD (purple arrows), plus CTP synthase and anaerobic dehydrogenase (shown as blue arrows together with other enzymes). Variations in the functional content of the BGCs are accounted by PKSs (red arrows), hybrid PKS-NRPS (yellow genes), arsenic regulation and metabolism proteins (brown arrows), and other regulators and transporters (green arrows). Three main classes or arseno-organic BGCs could be expected from this analysis: PKS-independent, PKS-NPRS-dependent and PKS-dependent biosynthetic systems. Dotted lines indicate sequence gaps, and an asterisk marks the sequence from *Nocardiopsis lucentensis*, which is assumed that have a missing PKS gene in one of the sequence gaps.

A

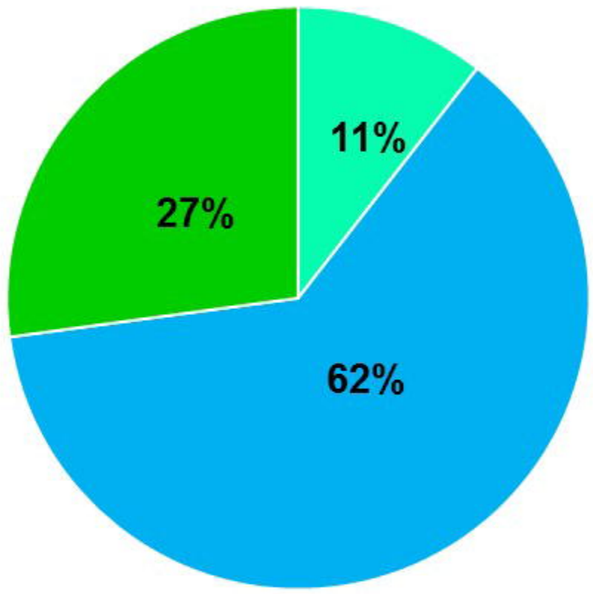



B



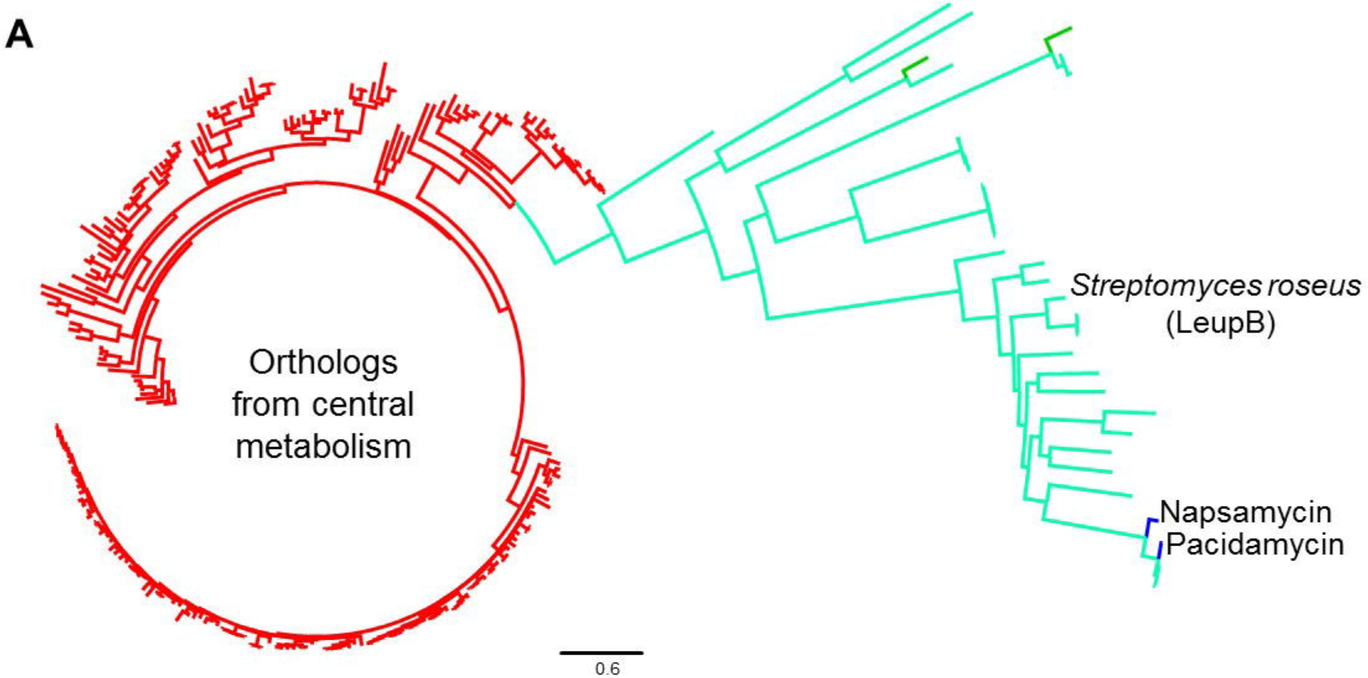
**A**

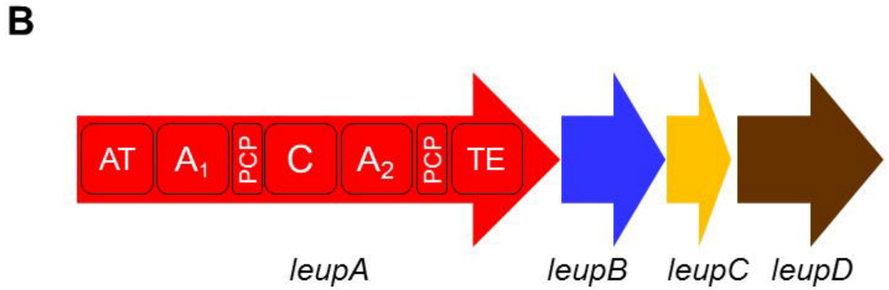


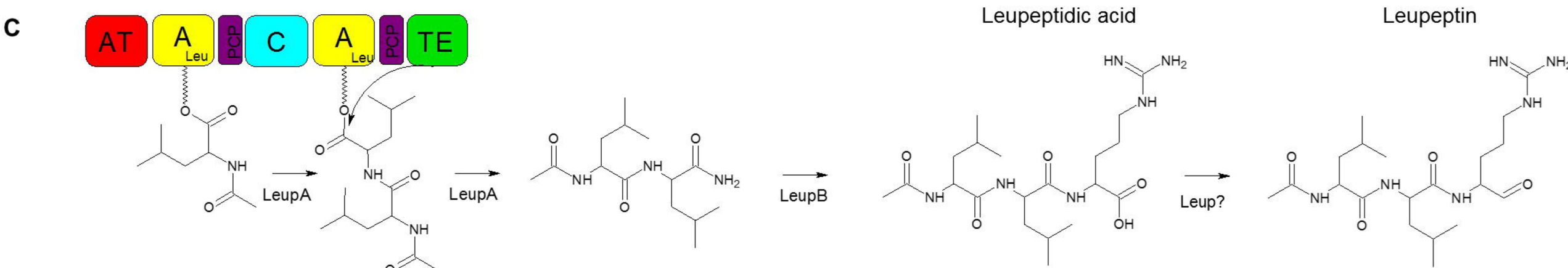

**EvoMining hits**
- **Detected by ClusterFinder**
- **Detected by AntiSMASH**
- **EvoMining predictions**

1. Indole-3-Glycerolphosphate synthase
2. *N*-acetyl-gamma-glutamyl-phosphate reductase
3. Aspartate transaminase
4. Histidinolphosphate aminotransferase
5. Homoserine-*O*-succinyl transferase
6. Enolase
7. Anthranilate phosphoribosyltransferase
8. Histidinol phosphatase
9. Citrate synthase
10. Acetolactate synthase
11. Glyceraldehyde-3-phosphate dehydrogenase
12. Phosphoglycerate dehydrogenase
13. Aconitate hydratase
14. Acetyl glutamate kinase
15. Aspartate kinase
16. Cysteine synthase
17. Prephenate dehydrogenase
18. Argininosuccinate lyase
19. Acetyl ornithine aminotransferase
20. Isopropylmalate dehydrogenase
21. 3-Phosphoshikimate-1-carboxyvinyl-transferase
22. 2-dehydro-3-deoxyphosphoheptonate aldolase
23. Asparagine synthase

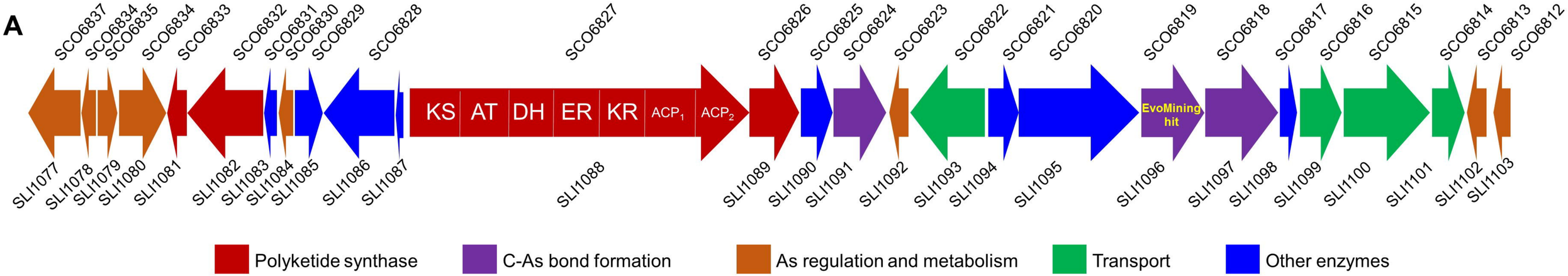
**B**

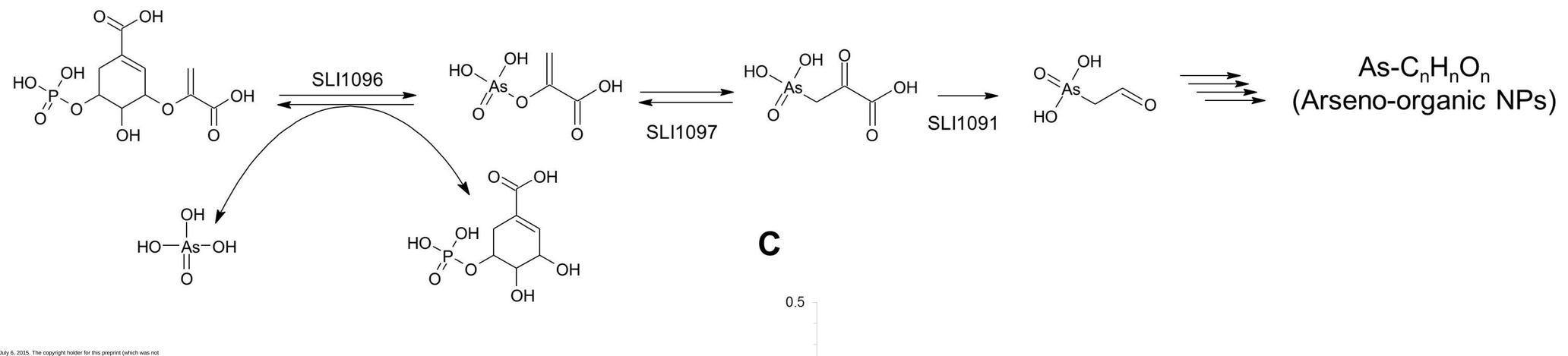

A
Orthologs
from central
metabolism
Streptomyces roseus
(LeupB)
Napsamycin
Pacidamycin
0.6


B
AT
A1
PCP
C
A2
PCP
TE
leupA
leupB
leupC
leupD
NRPS
Threonine kinase
Argininosuccinate lyase
Cysteine synthase

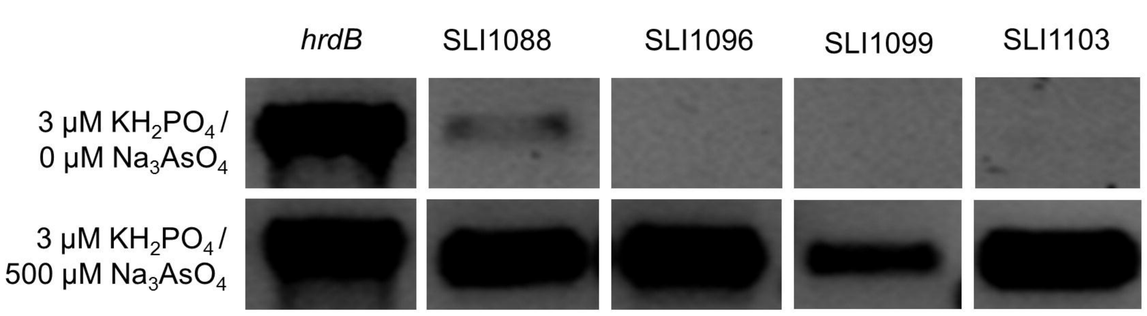

C
AT
A
Leu
PCP
C
A
Leu
PCP
TE
LeupA
LeupA
LeupB
Leupeptidic acid
Leup?
Leupeptin

**A**

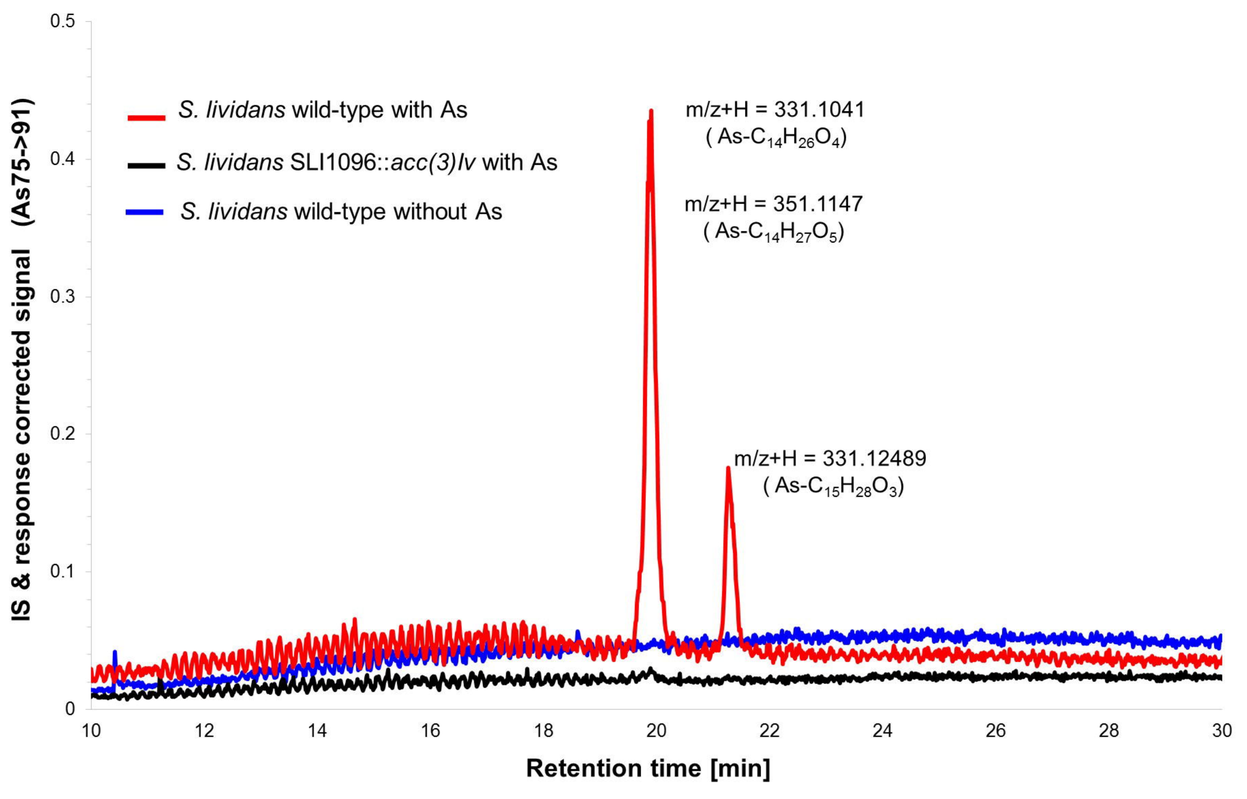



**B**




**C**



Hybrid PKS-NRPS dependent

PKS- dependent

PKS- Independent

30 Kb

*

*Kutzneria* sp 744

*Streptomyces* sp CcalMP8W

*Streptomyces fulvissimus* DSM 40593

*Streptomyces* sp PVA94-07

*Streptomyces* sp GBA94-07

*Saccharomonospora saliphila* YIM 90502

*Nocardiopsis lucentensis* DSM 44048

*Streptomycetaceae* bacterium MP113-05

*Streptomyces* sp BoleA5

*Streptomyces* sp CNQ766

*Streptomyces* sp CNY243

*Streptomyces* sp CNS335

*Streptomyces lividans* 66

*Streptomyces coelicolor* A3(2)

*Streptomyces coelicoflavus* ZG0656

1

0.99

1

1

1

0.98

0.99

1

0.96

0.99

1

0.99