Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes

Pablo Cruz-Morales[1,*], Johannes Florian Kopp[2], Christian Martínez-Guerrero[1], Luis Alfonso Yáñez-Guerra[1], Nelly Selem Mojica[1], Hilda Ramos-Aboites, Jörg Feldmann[2] & 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)
Evolution of Metabolic Diversity Laboratory,
Langebio, Cinvestav-IPN,
Irapuato, Guanajuato, México

Keywords: Natural Products Genome Mining, Phylogenomics, Arseno-organic metabolites, *Streptomyces*

Abstract

Natural products from microbes have provided humans with beneficial antibiotics for millennia. However, a decline in the pace of antibiotic discovery exerts pressure on human health as antibiotic resistance spreads, a challenge that may better faced by unveiling chemical diversity produced by microbes. Current microbial genome mining approaches have revitalized research into antibiotics, but the empirical nature of these methods limits the chemical space that is explored.

Here, we address the problem of finding novel pathways by incorporating evolutionary principles into genome mining. We recapitulated the evolutionary history of twenty-three enzyme families previously uninvestigated in the context of natural product biosynthesis in *Actinobacteria,* the most proficient producers of natural products. Our genome evolutionary analyses where based on the assumption that expanded -repurposed enzyme families- from central metabolism, occur frequently and thus have the potential to catalyze new conversions in the context of natural products biosynthesis. Our analyses led to the discovery of biosynthetic gene clusters coding for hidden chemical diversity, as validated by comparing our predictions with those from state-of-the-art genome mining tools; as well as experimentally demonstrating the existence of a biosynthetic pathway for arseno-organic metabolites in *Streptomyces coelicolor* and *Streptomyces lividans,* using gene knockout and metabolite profile combined strategy.

As our approach does not rely solely on sequence similarity searches of previously identified biosynthetic enzymes, these results establish the basis for the development of an evolutionary-driven genome mining tool that complements current platforms. We anticipate that by doing so real ‘chemical dark matter’ will be unveiled.

Introduction

Natural products (NPs) are a diverse group of specialized metabolites with adaptive functions, which include antibiotics, metal chelators, enzyme inhibitors, signaling molecules, amongst others (Traxler and Kolter, 2015). In bacteria, their biosynthesis is directed by groups of genes, referred to as Biosynthetic Gene Clusters (BGCs), usually encoded within a single locus, which allows for the concerted expression of biosynthetic enzymes, regulators, transporters and resistance genes (Diminic et al, 2014). Biosynthetic gene cluster formation has actually eased cloning of complete pathways and proposing experimentally verifiable *in silico* predictions of BGCs. Indeed, a fundamental principle of current bioinformatics approaches used for the discovery of novel NPs, commonly referred to as genome mining of NPs, rely in the assumption that once an enzyme is unequivocally linked to the production of a given metabolite, genes in the surroundings of its coding sequence are associated with its biosynthesis (Medema and Fischbach, 2015). This functional annotation approach, from genes to metabolites, has led to comprehensive catalogs of putative BGCs directing the synthesis of an ever-growing universe of metabolites (Medema et al, 2015; Hadjithomas et al, 2015).

NPs are also a rich source of compounds that have found pharmacological applications, as highlighted by the latest Nobel Prize in Physiology or Medicine awarded to researchers due to their contributions surrounding the discovery and use of NPs to treat infectious diseases. Indeed, in the context of increased antibiotic resistance, genome mining has revitalized the investigation into NP biosynthesis and their mechanisms of action (Demain, 2014; Harvey et al, 2015). In contrast with pioneering studies based on activity-guided screenings of NPs, current efforts based on genomics approaches promise to turn the discovery of NP drugs into a chance-free endeavor (Schreiber, 2005; Bachmann et al, 2014; Demain, 2014). Evidence supporting this possibility has steadily increased since ECO4601, a farnesylated benzodiazepinone discovered using genome mining approaches, which entered into human clinical trials more than a decade ago (Gourdeau et al, 2007).

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) (Conway and Boddy, 2013; Ichikawa et al, 2013). These approaches can be classified as: (i) chemically-driven, where the discovery of the biosynthetic gene cluster is elucidated based on a fully chemically characterized 'orphan' metabolite (Barona-Gómez et al, 2004); or (ii) genetically-driven, where known sequences of protein domains (Lautru et al, 2005) or active-site motifs (Udwary et al, 2007) help to identify putative BGCs and their products. The latter relates to the term 'cryptic' BGC, defined as a locus that has been predicted to direct the synthesis of a NP, but which remains to be experimentally confirmed (Challis, 2008).

Decreasing costs of sequencing technologies has dramatically increased the number of putative BGCs. In this context, genome mining of NPs can help to prioritize strains on which to focus for further investigation (Rudolf et al, 2015; Shen et al, 2015). During this process, based on a priori biosynthetic insights, educated guesses surrounding PKS and NRPS can be put forward. In turn, such efforts increase the likelihood of discovering interesting chemical and biosynthetic variations. Moreover, biosynthetic logics for a growing number of NP classes, such as phosphonates (Metcalf and van der Donk, 2009; Ju et al, 2013), are complementing early NRPS/PKS-centric approaches. In contrast, finding novel chemical scaffolds, expected to be synthesized by cryptic BGCs, remains a challenging task. Therefore, with the outstanding exception of ClusterFinder (Cimmermancic et al, 2014), which uses Pfam domain pattern-based predictions, most genome mining methods are focused in known classes of NPs, hampering our ability to discover chemical novelty (Medema & Fischbach, 2015).

In this work we address the problem of finding novel pathways by genome mining, by means of integrating three evolutionary concepts related to emergence of NP biosynthesis. First, we assume that new enzymatic functions evolve by retaining their reaction mechanisms, while expanding their substrate specificities (Gerlt and Babbitt, 2001). 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 (Caetano-Anollés et al, 2009). In the context of NP biosynthesis, the canonical example for this would be fatty acid synthetases as the ancestor of PKSs (Jenke-Kodama et al, 2005). Consequently, the correspondence of enzymes to either central or specialized metabolism, typically solved through detailed experimental analyses, could also be determined 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 (Vining, 1992; Firn and Jones, 2009; Medema et al, 2015).

These three evolutionary principles were formalized as a functional phylogenomics approach (Figure 1), which has been previously referred to as EvoMining (Medema & Fischbach, 2015). This approach leads to the identification of expanded, repurposed enzyme families, with the potential to catalyze new conversions in specialized metabolism. Using this approach we predicted several new potential pathways including the first ever reported family of BGCs for arseno-organic metabolites. Experimental evidence for arseno-organic metabolites produced by model actinomycetes *Streptomyces coelicolor* A3(2) and *Streptomyces lividans* is provided. As our approach does not rely solely on sequence similarity searches of previously identified NP biosynthetic enzymes, these results establish the basis for the development of an evolutionary-driven genome mining tool that complements current platforms.

Materials and methods

Database integration

Genome DB: Complete and draft genomes of 230 members of the *Actinobacteria phylum* (supplementary table 1) were retrieved from 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. Central metabolism DB: 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 (Barona-Gómez et al, 2012). The databases included a total of 339 queries for nine pathways, including amino acid biosynthesis,

glycolysis, pentose phosphate pathway and tricarboxylic acids cycle (supplementary table 2) these sequences were obtained from the genome scale metabolic reconstructions of three actinobacterial model strains *S. coelicolor* (Borodina et al, 2005) *Mycobacterium tuberculosis* (Jamshidi and Palsson, 2007) and *Corynebacterium glutamicum* (Kjeldsen and Nielsen, 2009). Seed NP database: NRPS and PKS BGCs were obtained from the DoBISCUIT and ClusterMine 360 databases (Conway and Boddy, 2013; Ichikawa et al, 2013). BGCs for other NP classes were collected from available literature (supplementary table 3). Annotated GenBank formatted files were downloaded from the GenBank database to assemble a database that included 226 BGCs.

Phylogenomics analysis pipeline

The sequences Central Metabolism DB were used as queries to retrieve members of each enzyme family from the Genomes DB using BlastP (Altschul et al 1990), with an e-value cutoff of 0.0001 and a bit core cutoff of 100. 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 (supplementary table 4). An enzyme family expansion was scored if the number of homologs in at least one genome was higher than the average number of homologs identified from each BLAST search. The next step was to BLAST sequences in the expanded families using amino acid sequences from the Seed NP DB as queries using an e-value cutoff of 0.0001 and a bit score cutoff of 100. When matches were identified, these represented candidate recruitment events.

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. The bidirectional best hits with the sequences in the Central Metabolism DB were identified and tagged using in-house scripts to distinguish central metabolic orthologs from other homologs that resulted from expansion events. These sets of sequences were aligned using Muscle version 3.8.31 (Edgar, 2004). The alignments were inspected and curated manually using JalView 2.8 (Waterhouse et al, 2009). The curated alignments were used for phylogenetic reconstructions, which were estimated using MrBayes 3.2.3 (Ronquist et al, 2012) with the

following parameters: aamodelpr=mixed, samplefreq=100, burninfrac=0.25 in four chains and for 1000000 generations. The identification of new recruitments (hits) was done by visual inspection of each phylogentic tree. The homologs resulting from expansion events located in clades where at least one homolog from the Seed NP DB was found were selected and the regions of approximately 80 Kbs flanking their coding sequences were retrieved from GenBank formatted files using a perl Script. These contigs were annotated using the stand alone version of antiSMASH 3.0 (Weber et al, 2015) with the following options: -inclusive –cf_threshold 0.7 and RAST (Aziz et al, 2008). Further analysis of annotated contigs was done manually using the Artemis Genome Browser (Carver et al, 2012).

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) (Gust et al, 2003). 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 supplementary table 9. 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 (Redenbach et al, 1996). 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 (Gust et al, 2003). 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).

RT-PCR 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. 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.

LC-MS metabolite profile analysis

The SLI_1096 and SCO6818 minus mutants were grown on modified R5 medium (K_2SO_4 0.25 g; $MgCl_2$-$6H_20$ 10.12 g; glucose 10 g; casamino acids 0.1 g; TES buffer 5.73 g; trace element solution (Kieser et al, 2000) 2 mL; agar 20 g) 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 (Amayo et al, 2011). The ESI-Orbitrap-MS was set to positive ion mode in a scan range from 250-1100 amu. Also, MS2-spectra for the major occurring ions were generated.

Results and discussion

Evolutionary recapitulation of expanded and repurposed enzyme families

The proposed model for the evolution of NPs BGCs, based in the relationships between central metabolic enzymes and NP BGCs, was investigated through systematic and comprehensive phylogenomics (Figure 1A). The inputs for this analysis were formalized as three databases: (i) a genomes database consisting of 230 genome sequences belonging to the phylum *Actinobacteria,* which includes the most proficient producers of NPs, e.g. the genus *Streptomyces* (Genomes DB; Supplementary table 1); (ii) the amino acid sequences of central metabolic enzymes, belonging to nine selected pathways known to provide precursors for the synthesis of NPs (Central Metabolism DB, Supplementary table 2). These pathways, which have been shown to have enzymes that suffer from expansions with taxonomic resolution in *Actinobacteria* (Barona-Gómez et al, 2012), include glycolysis, tricarboxylic acid cycle, pentose phosphate pathway and the biosynthetic pathways for the proteinogenic amino acids. In total, these pathways consist of 106 enzyme families that were retrieved from well curated and experimentally validated genome-scale metabolic model reconstructions of model *Actinobacteria*, e.g. *Streptomyces coelicolor*; and (iii) a NP-related enzymes database including amino acid sequences from 226 known actinobacterial BGCs manually extracted from the literature, which provided a proof of concept limited universe (Seed NP DB, Supplementary table 3).

For each enzyme family, enzyme expansions were identified when the number of orthologs detected for a query sequence was greater than the average number detected in all genomes plus one standard deviation (Supplementary table 4). 101 enzyme families fulfilled this criterion, and all homologous sequences from each expanded enzyme family were retrieved from the Genomes DB. The expanded enzyme families were classified into enzymes devoted to central metabolism, and expanded enzymes that may perform other functions, including NP biosynthesis. The latter classification was done using bidirectional best hits analyses, under the assumption that orthologous relationships imply the same

function, and *vice versa*. However, at this stage, involvement of these enzymes in other metabolic functions, such as catabolism or resistance mechanisms, could not be ruled out.

We therefore, as our next step, aimed at identifying expanded enzymes that may have been recruited to catalyze conversions during NP biosynthesis. For this purpose, we performed a sequence similarity search between the Seed NP DB and the central metabolism DB obtained from the previous analyses. Recruitment events in NP BGCs were called when a sequence in an expanded enzyme family had a homologous sequence in the Seed NP DB. This search revealed that a total of 23 out of 101 expanded enzyme families had recognizable recruitment events by NP BGCs included in our Seed NP DB, which by no means represents an exhaustive catalog of NP pathways (Supplementary table 4). The remaining 75 expanded enzyme families may also include recruitment events by other NP BGCs not included in our Seed NP DB, or by other functional forms of metabolism such as degradation of xenobiotics. Nevertheless, the 23 enzyme families with recruitments analyzed in the following steps served as proof of concept for the discovery of new BGCs using phylogenomics.

To identify new recruitment events, we used phylogenetic analysis to define the evolutionary relationships in the 23 expanded-then-recruited enzyme families, and to infer possible functional fates for each homolog. In most of the 23 phylogenetic reconstructions (provided as online supplementary material), the central metabolic homologs formed clades with topologies that recapitulate speciation events. This could be confirmed by comparing the obtained topologies with a species phylogenetic tree constructed using neutral markers, such as the RNA Polymerase Beta subunit or RpoB (supplementary figure 1). In clear contrast, expansion events grouped in one or more independent clades, and the most divergent clades often included the recruited homologs from the Seed NP database (Figure 1B). We then assumed that high sequence divergence reflects rapid evolution, as expected for adaptive metabolism such as NP biosynthesis. Therefore, divergence of the clades that included homologs from the Seed NP DB was used to call positive hits (Figure 1B). Using these criteria we identified 515 enzyme hits contained in phylogenetic trees, which were visually inspected.

To further support the functional association of these enzyme's hits with NP biosynthesis, we retrieved up to 80 Kbp of surrounding genomic sequence harboring the targeted genes (71.3 Kbp in average, 19.9 Kbp standard deviation). This process yielded 423 contigs out of the 515 enzyme hits, as some of these hits (almost 18%) were located in short contigs that could not be properly annotated. The 423 retrieved contigs were mined for BGCs of known classes of NPs using antiSMASH 3.0 (Weber et al, 2015), as well as for putative BGCs using ClusterFinder with a probability cut-off of 0.70 (Cimermancic et al, 2014). This analysis allowed us to "validate" our predictions after confirming that a hit enzyme was located within the boundaries of BGCs that could be predicted and annotated with these tools. For instance, 15% of the enzyme hits came from the Seed NP DB, whilst 59% coincided with AntiSMASH (43%), ClusterFinder (9%), or both (7%) (Figure 2A and supplementary table 5).

The remaining 26% of the enzyme hits were not included within the boundaries of BGCs predicted by neither AntiSMASH nor ClusterFinder, and therefore we refer to as "not validated". Detailed analysis, however, revealed that 11% of the total hits, belonging to this category (not validated in Figure 2A), are at 20 Kbp or less away from predicted BGCs. This observation suggests that these enzymes escaped from the detection limits of the algorithms used, either due to their location at few Kbp beyond their boundary detection rules, or for having scores below the cut-off used for ClusterFinder predictions (0.70). The remaining 15% of the total hits, belonging to this category, could only be predicted by our approach (Figure 2A), and may represent truly novel BGCs. As a first step to look into this possibility, we analyzed the distribution and annotation of these hits, and we noticed that some enzyme families seem to have more predictive value as discussed further.

For instance, 43% of the total hits were found within three enzyme families, namely, asparagine synthase, 3-phosphoshikimate-1-carboxivinyl transferase and 2-dehydro-3-deoxyphosphoheptanoate aldolase (Figure 2B). This may be due to the limitations of our input databases. However, it is also tempting to speculate that enzyme recruitment is biased for certain enzyme families, at least more than currently acknowledged (Medema et al., 2015). Along these lines, the 3-phosphoshikimate-1-carboxivinyl

transferase family, or AroA, comprise 38% of the enzyme hits that could not be validated, and includes the largest number of hits not associated to BGCs (Figure 2C). The not validated hits may represent the lowest confidence predictions, but also those with the highest potential for discovery of unprecedented chemistry (Medema and Fischbach, 2015).

The experimental characterization of this type of BGCs, i.e. not validated predictions, is challenging. Therefore, we focused on providing experimental evidence to validate our approach by selecting a predicted BGC as supported by other well-known NP enzymes, but which also includes an AroA hit that is unprecedented within the synthesis of NPs (Supplementary table 6). The selected prediction included known NP enzymes, namely, a two-gene PKS assembly, but also a 3-phosphoshikimate-1-carboxivinyl transferase homolog that is different to the AroA enzymes present in the BGCs of the polyketide asukamycin (Rui et al, 2010) and phenazines (Seeger et al, 2011), as discussed further.

Discovery of arseno-organic metabolites related to a novel AroA enzyme

S. coelicolor A3(2) and *S. lividans* 66 are closely related model streptomycetes that have been thoroughly investigated both, before the genomics era and with genome mining approaches (Nett et al 2009, Bentley et al, 2002, Cruz-Morales et al, 2013). Presumably, most of their NPs biosynthetic repertoire has been elucidated (Challis 2014). Furthermore, several methods for genetic manipulation of these two strains are available (Kieser et al, 2000), making them ideal for proof of concept experiments. Specifically, our phylogenomics analysis led to five hits related to three BGCs in the *S. coelicolor* and *S. lividans* genomes (Supplementary table 5). Four of these hits (in both strains), included known recruitments that are associated with the BGC responsible for the synthesis of the calcium-dependent antibiotic or CDA (Hojati et al, 2002). The remaining hit, which caught our attention, is related to the AroA enzyme family, which 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 (Zhang and Berti, 2006).

The phylogenetic reconstruction of the actinobacterial AroA enzyme family shows a major clade associated with central metabolism; this clade includes SLI_5501 from *S. lividans* and SCO5212 from *S. coelicolor*. More importantly, the phylogeny also includes a divergent clade with at least three sub-clades, two that include family members linked to the known BGCs of the polyketide asukamycin (Rui et al, 2010) and phenazines (Seeger et al, 2011); as well as AroA homologs from 26% of the genomes of our database. In *S. coelicolor* and *S. lividans*, these recruited homologs are encoded by SLI_1096 and SCO6819, and they are marked with green crosses in Figure 2C. These genes are situated only six genes upstream of the two-gene PKS system (SCO6826-7 and SLI_1088-9, respectively) used by AntiSMASH and ClusterFinder to "validate" these BGCs in the previous section (supplementary table 6). The PKS was identified in *S. coelicolor* since the early genome mining efforts in this organism, but often it was referred to as a cryptic BGC, which did not include these *aroA* homologs (Bentley et al, 2002; Nett et al, 2009).

Given that the aroA genomic context is highly conserved between the genomes of *S. lividans* and *S. coelicolor*, for simplicity we will refer to the *S. lividans* genes only (Supplementary table 6; MIBIG accession BGC0001283). The syntenic region spans from SLI_1077 to SLI_1103, including several biosynthetic enzymes, regulators and transporters, suggesting a functional association. Further annotation 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 the 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 (Metcalf and van der Donk 2009), but that were not be detected by AntiSMASH or ClusterFinder (Supplementary table S6).

Other non-enzymatic functions within this BGC could be annotated, including a set of ABC transporters, originally annotated as phosphonate transporters (SLI_1100 and SLI_1101), and four arsenic tolerance-related genes (SLI_1077-1080) located upstream. These genes are paralogous to the main arsenic tolerance system encoded by the *ars* operon (Wang et al, 2006), 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 (Figure 3A).

Prior to functional characterization, 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 (Tawfik and Viola, 2011; Elias et al, 2012). Second, previous studies have demonstrated that AroA is able to catalyze a reaction in the opposite direction to the biosynthesis of aromatic amino acids with poor efficiency, namely, the formation of PEP and 3-phosphoshikimate from enolpyruvyl shikimate 3-phosphate and phosphate (Zhang and Berti, 2006). Indeed, arsenate and enolpyruvyl shikimate 3-phosphate can react to produce arsenoenolpyruvate (AEP), a labile analog of PEP, which is spontaneously transformed into pyruvate and arsenate (Zhang and Berti, 2006). 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 (Metcalf and van der Donk, 2009), has been reported (Chawla et al, 1995).

Altogether, the previous evidence was used to postulate a novel biosynthetic pathway encoded by SLI_1077-SLI_1103, which may direct the synthesis of a novel arseno-organic metabolite (Figure 3A) (A detailed functional annotation, and biosynthetic proposal is provided in the supplementary table 6 and supplementary figure 2).

To determine the product of the predicted BGC, in both *S. lividans* and *S. coelicolor*, we used expression analysis, as well as comparative metabolic profiling of wild type and mutant strains. Using RT-PCR analysis, we first determined the transcriptional expression profiles of one of the PKS genes (SLI_1088), *aroA* (SLI_1096), one of the *arsR*-like regulator (SLI_1103), and the periplasmic-binding protein of the ABC-type transporter (SLI_1099). As expected for a cryptic BGC, the results of these experiments demonstrate that the proposed pathway is repressed under standard laboratory conditions. We then analyzed the potential role of arsenate as the BCG inducer upon the addition of a gradient concentration of arsenate (0 – 500 µM) and phosphate (3-300µM). Indeed, we found that the analyzed genes were induced when both *S. lividans* and *S. coelicolor* were grown in the presence of 500 µM of arsenic and 3 µM of phosphate (Figure 3B).

In parallel, we used PCR-targeted gene replacement to produce mutants of the SLI_1096 and SCO6819 genes, and analyzed the phenotypes of the mutant and wild type strains on a combined arsenate/phosphate gradient, i.e. low phosphate and high arsenate, and *vice versa*. Intracellular organic extracts were analyzed using HPLC coupled with an ICP-MS calibrated to detect arsenic-containing molecular species. Simultaneously, a high-resolution mass spectrometer determined the mass over charge (*m*/*z)* of the ions detected by the ICP. This set up allows for high-resolution detection of arseno-organic metabolites (Amayo et al, 2011). Using this approach, we detected the presence of an arseno-organic metabolite in the organic extracts of both wild-type *S. coelicolor* and *S. lividans*, with *m*/*z* value of 351.1147. These metabolites could not be detected in either identical extracts from wild type strains grown in the absence of arsenate or in the mutant strains deficient for the SLI_1096/SCO6819 genes (Figure 3C). Thus, the product of this pathway may be a relatively polar arseno-polyketide. The actual structures of

these products are still subject to further chemical investigation and will be discussed in detail in a future publication.

Closure of the conceptual loop: Genome mining for arseno-organic BGCs

Confirmation of a link between SLI_1096/SCO6819 and the synthesis of an arseno-organic metabolite provides an example on how genome-mining efforts, based in novel enzyme sequences, can be advanced. For instance, co-occurrence of divergent SLI_1096 orthologs, now called arsenoenolpyruvate synthases (AEPS); SLI_1097 arsenopyruvate mutase (APM) and arsonopyruvate decarboxylase (APD), can be now confidently used beacons to mine bacterial genomes. Indeed, 26 BGCs with the potential to synthesize arseno-organic metabolites, all of them encoded in genomes of myceliated *Actinobacteria*, were identified after sequence similarity searches using the non-redundant GenBank database (Figure 4). The divergence and potential chemical diversity within these arseno-related BGCs was analyzed using phylogenetic reconstructions with a matrix of three conserved genes of these BGCs. This analysis suggests three possible sub-classes with distinctive features, PKS-NRPS independent, PKS-NPRS-hybrid dependent and PKS-dependent biosynthetic systems.

In addition, once a hit predicted to perform an unprecedented function is positively linked to a new BGC and its resulting metabolite, this could lead to identification of novel classes of conserved BGCs (Figure 2C). To illustrate this, the non-validated AroA hits within the NP-related clade of this phylogenetic tree were manually curated in search for a conserved BGC. One particular case, representing a previously unnoticed sub-clade that belongs to the not validated category was found to appear frequently. After detailed annotation, it was found that this BGC is conserved in at least sixty-three *Streptomyces* genomes, and in the genome of *Microtetraspora glauca* NRRL B-3735, which are included in the non-redundant GenBank database (supplementary table 7). For these analyzes, 15 genes upstream and downstream the AroA hit were extracted as before and annotated on the basis of the locus from

Streptomyces griseolus NRRL B-2925 (Supplementary table 8 and supplementary figure 3). 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 31, present in an equal number of NP BGCs, actually supports the NP nature of this locus.

Overall, we conclude that functional predictions based upon phylogenomics analysis 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. The proposed approach now makes the retrieval of arseno-organic BGCs possible that were previously not detected by AntisMASH or ClusterFinder (for instance the one from *Streptomyces ochraceiscleroticus*). The current work provides the proof of concept and the fundamental insights to develop a bioinformatics pipeline based upon evolutionary principles, to be referred as EvoMining.

Acknowledgements

We are indebted with 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 Service of Unidad Irapuato, Cinvestav, and Araceli Fernandez for technical support in high-performance computing. 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 posdoctoral fellowship. JF and JFK acknowledge funding from the College of Physical Sciences, University of Aberdeen, UK.

Competing interests

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

References.

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.
2. Amayo KO, et al. 2011 Identification and Quantification of arsenolipids using reversed-phase HPLC coupled simultaneously to high-resolution ICPMS and high-resolution electrospray SM without species-specific standards. Anal. Chem. 83: 3589-3595.

3. Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC genomics 9:75.
4. 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. 41:175–184.
5. 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. J. Am. Chem. Soc. 126:1628216283.
6. Barona-Gómez F, Cruz-Morales P, Noda-García L. 2012. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? Antonie van Leeuwenhoek 101:35–43.
7. Bentley SD, Chater KF, Cerdeño-Tárraga A-MM, et al. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417:141–147.
8. Borodina I, Krabben P, Nielsen J. 2005. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. Genome Res. 15(6):820-9.
9. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE. 2009. The origin and evolution of modern metabolism. Int. J. Biochem. Cell Biol. 41:285–297.
10. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 28(4):464-9.
11. Challis G. 2008. Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology (Reading, England) 154:1555–1569.
12. Challis GL. 2014. Exploitation of the Streptomyces coelicolor A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. J. Ind. Microbiol. Biotechnol. 41:219–232.
13. Chawla S, Mutenda EK, Dixon HB, Freeman S, Smith AW. 1995. Synthesis of 3-arsonopyruvate and its interaction with phosphoenolpyruvate mutase. Biochem. J. 308 (Pt 3):931–935.

14. Cimermancic P et al. 2014. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell 158(2):412-2.
15. Conway K, Boddy C. 2013. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 41:D402–D407.
16. 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 Biol. Evol. 5:1165–1175.
17. Demain AL. 2014. Importance of microbial natural products and the need to revitalize their discovery. J. Ind. Microbiol. Biotechnol. 41:185–201.
18. Diminic J, et al 2014. Evolutionary concepts in natural products discovery: what actinomycetes have taught us. J Ind Microbiol Biotechnol. 41(2):211-7.
19. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids Res. 32:1792–1797.
20. Elias M, et al. 2012. The molecular basis of phosphate discrimination in arsenate-rich environments. Nature 491:134–137.
21. Firn R, Jones C. 2009. A Darwinian view of metabolism: molecular properties determine fitness. J. Exp. Bot 60:719–726.
22. Gerlt JA, Babbitt PC. 2001. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu. Rev. Biochem. 70:209–246.
23. Gourdeau H, et al. 2007. Identification, characterization and potent antitumor activity of ECO-4601, a novel peripheral benzodiazepine receptor ligand. Cancer Chemother. Pharmacol. 61:911921.
24. 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. Proc. Natl. Acad. Sci. USA 100:1541–1546.

25. Hadjithomas M, et al. 2015. IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. MBio. 6(4):e00932.
26. Harvey A, Edrada-Ebel R, Quinn R. 2015. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 14:111–129.
27. Hojati Z, et al. 2002. Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. Chem Biol. 9(11):1175-87.
28. Ichikawa N, et al. 2013. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic acids Res. 41:D408–14.
29. Jamshidi N, Palsson BØ. 2007. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol. 8;1:26
30. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E. 2005. Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol. 22(10):2027-39.
31. Ju K-S, Doroghazi J, Metcalf W. 2013. Genomics-enabled discovery of phosphonate natural products and their biosynthetic pathways. J. Ind. Microbiol. Biotechnol. 41(2):345-56.
32. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. 2000. Practical Streptomyces genetics. Norwitchultsch UK: The John Innes Foundation
33. Kjeldsen KR, Nielsen J. 2009. In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network. Biotechnol Bioeng. 102(2):583-97.
34. Lautru S, Deeth R, Bailey L, Challis G. 2005. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat. Chem. Biol. 1:265–269.
35. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. 2015. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput. Biol. 10(12):e1004016..
36. Medema MH, et al, 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 11(9):625-31

37. Medema MH, Fischbach MA. 2015. Computational approaches to natural product discovery. Nat Chem Biol. 11(9):639-48.

38. Metcalf W, van der Donk W. 2009. Biosynthesis of phosphonic and phosphinic acid natural products. Biochemistry 78:65–94.

39. Nett M, Ikeda H, Moore BS. 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat Prod Rep 26:1362–1384.

40. Redenbach M, et al. 1996. A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb Streptomyces coelicolor A3(2) chromosome. Mol. Microbiol. 21:77–96.

41. Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

42. Rudolf JD, Yan X, Shen B. 2015. Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. J Ind Microbiol Biotechnol. 43(2-3):261-76

43. Rui Z, et al. 2010. Biochemical and genetic insights into asukamycin biosynthesis. J. Biol. Chem. 285:24915–24924.

44. Schreiber S. 2005. Small molecules: the missing link in the central dogma. Nan Chem Biol 1:64–66.

45. Shen B, et al. 2015. Enediynes: Exploration of microbial genomics to discover new anticancer drug leads. Bioorg Med Chem Lett. 25(1):9-15.

46. Seeger K, et al. 2011. The biosynthetic genes for prenylated phenazines are located at two different chromosomal loci of Streptomyces cinnamonensis DSM 1042. Microb Biotechnol. 4:252–262.

47. Tawfik D, Viola R. 2011. Arsenate replacing phosphate: alternative life chemistries and ion promiscuity. Biochemistry 50:1128–1134.

48. Traxler MF, Kolter R. 2015. Natural products in soil microbe interactions and evolution. Nat Prod Rep. 32(7):956-70.

49. Udwary D, et al. 2007. Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. Proc. Natl. Acad. Sci. USA. 104:10376–10381.

50. Vining LC. 1992. Secondary metabolism, inventive evolution and biochemical diversity - a review. Gene 115:135–140.

51. Wang L, et al. 2006. arsRBOCT arsenic resistance system encoded by linear plasmid pHZ227 in Streptomyces sp. strain FR-008. Appl. Environ. Microbiol. 72:3738–3742.

52. 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) 25:1189–1191.

53. Weber T, et al. 2015. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 43(W1):W237-43.

54. Zhang F, Berti PJ. 2006. Phosphate analogues as probes of the catalytic mechanisms of MurA and AroA, two carboxyvinyl transferases. Biochemistry 45:6027–6037.

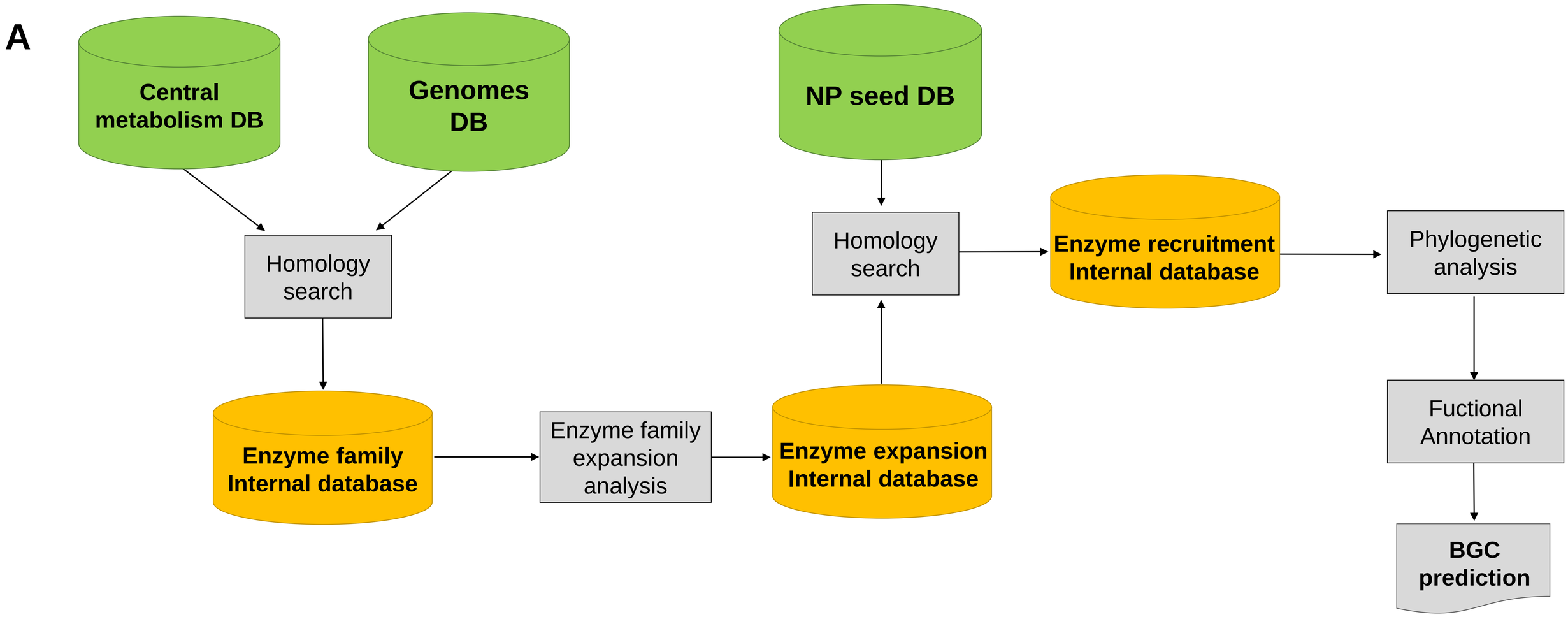
Figure legends

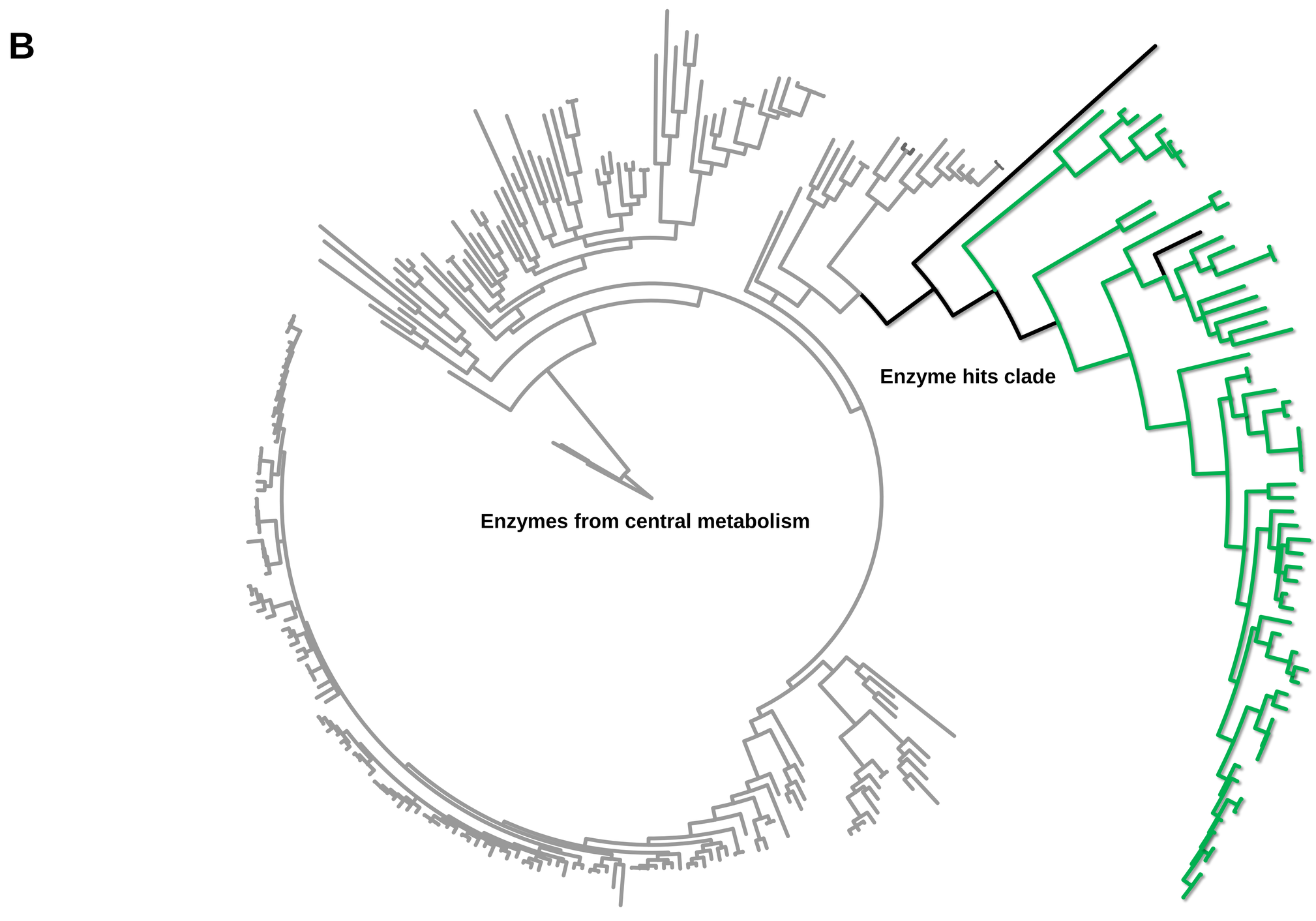
Figure 1. Phylogenomic analysis for the recapitulation of the evolution of NP biosynthesis. **A.** Bioinformatic workflow: the three input databases, as discussed in the text, are shown in green. Internal databases are shown in yellow, whereas grey boxes depict processes. **B.** An example of a phylogenetic tree for the recapitulation of the evolution of NP biosynthesis (**Tree S4**) using the case of 3-phosphoshikimate-1-carboxyvinyl transferase family. Gray branches include homologs related to central metabolism and their topology resembles that of a species guide tree (**Tree S1**). The Enzyme hits clade is highlighted: branches in black mark homologs from the Seed NP DB, and green branches indicate new enzyme hits. A detailed annotation of the clade is shown in figure 2C.

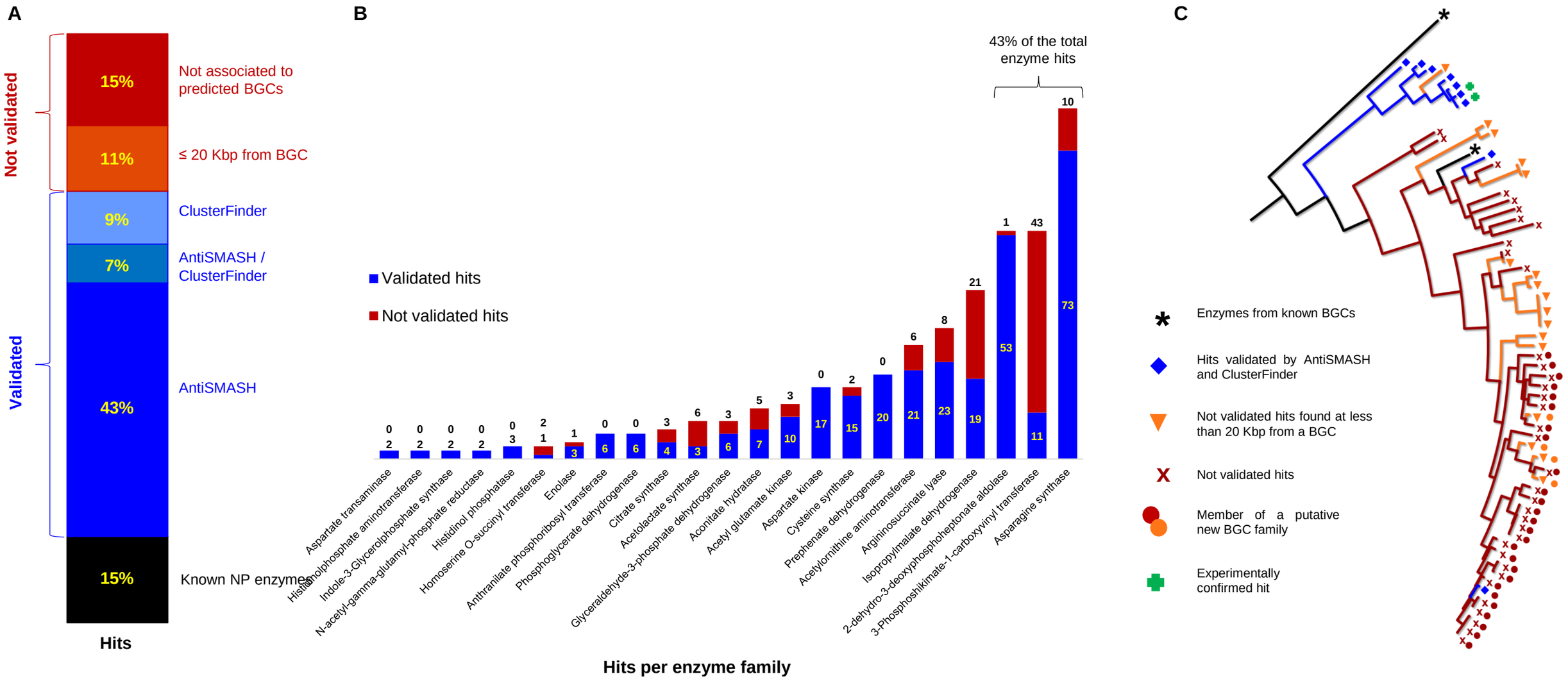
Figure 2. Overview of the BGC predictions obtained after the reconstruction of the evolution of NP biosynthesis. A. Distribution of validated and not validated enzyme hits. Most hits were validated by AntiSMASH and/or ClusterFinder. Refer to text for further details B. Distribution of hits per recruited enzyme family. The numbers show the total hits either validated or not validated for each enzyme family. C. Zoom-in of the sub-clade containing enzyme hits belonging to the 3-carboxyvinyl-phosphoshikimate transferase or AroA enzyme family. The complete tree, from which this sub-clade was extracted, can be seen in Figure 1B or Tree S4. Asterisks indicate enzymes recruited into known BGCs. blue branches and diamonds mark hits validated by other NP genome mining methods, orange triangles on orange branches mark not-validated hits located at less than 20 Kbp from a predicted BGC. Red X letters on red branches indicate not validated hits not related to predicted BGCs, red or orange circles indicate hits belonging to a family of new putative BGCs, finally green crosses indicates two experimentally characterized recruitment events discussed further in this article.

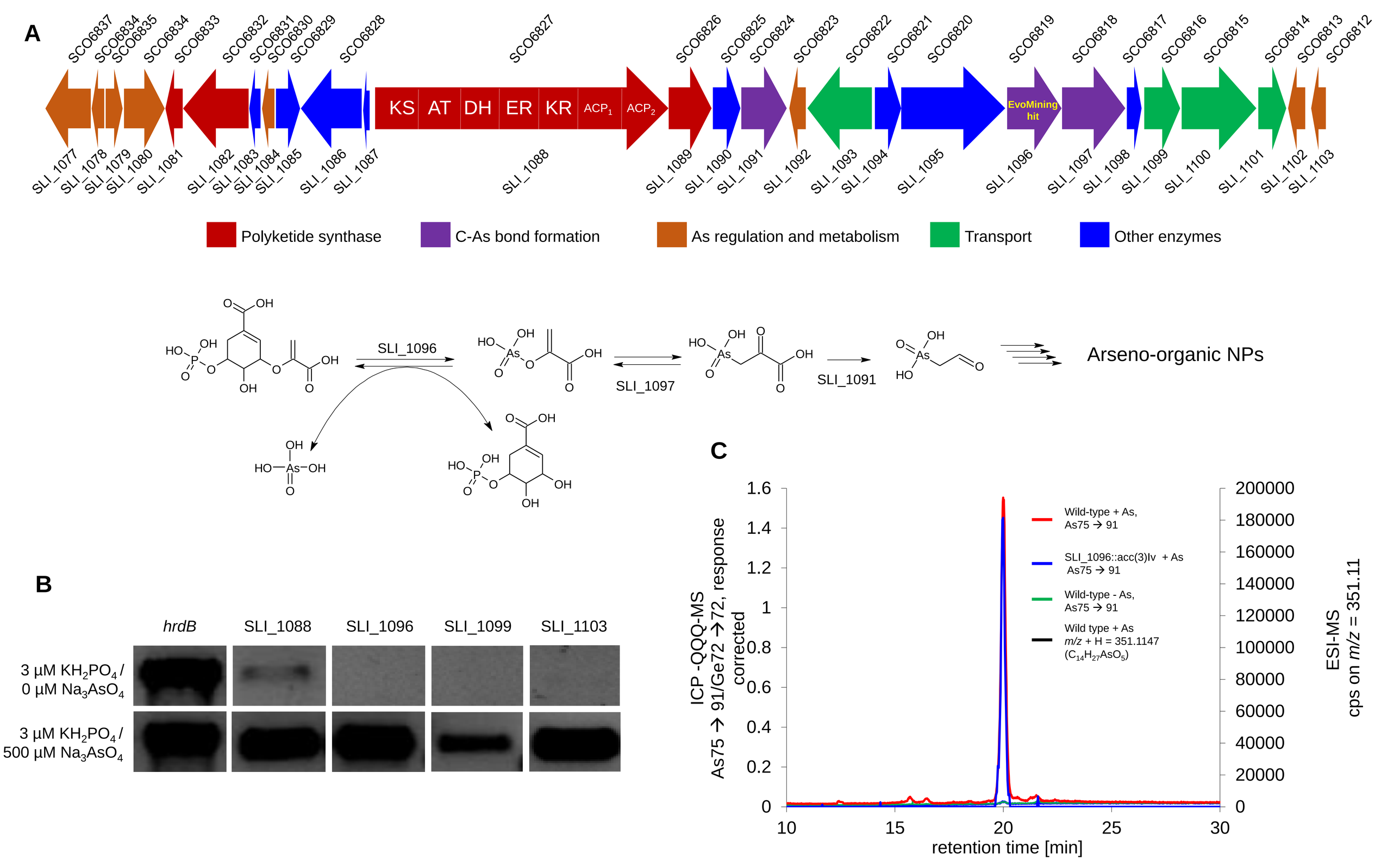
Figure 3. Discovery of a BGC for arseno-organic NPs in *S. coelicolor* and *S. lividans*. A. The conserved arseno-organic BGC in *S. lividans* 66 and *S. coelicolor* is shown. The proposed biosynthetic logic for early intermediates in the biosynthesis of arseno-organic metabolites is shown. **B.** Transcriptional analysis of selected genes within the arseno-organic BGC, showing that gene expression is repressed under standard conditions, but induced upon the presence of arsenate. **C.** HPLC-Orbitrap/QQQ-MS trace of organic extracts from mycelium of wild type and the SLI_1096 mutant showing the detection of arsenic-containing species. Three *m*/*z* signals were detected within the two peaks found in the trace from the wild type strain grown on the presence of arsenate. These *m/z* signals are absent from the wild type strain grown without arsenate, and from the SLI_1096 mutant strain grown on phosphate limitation and the presence of arsenate. Identical results were obtained for *S. coelicolor* and the SCO6819 mutant.

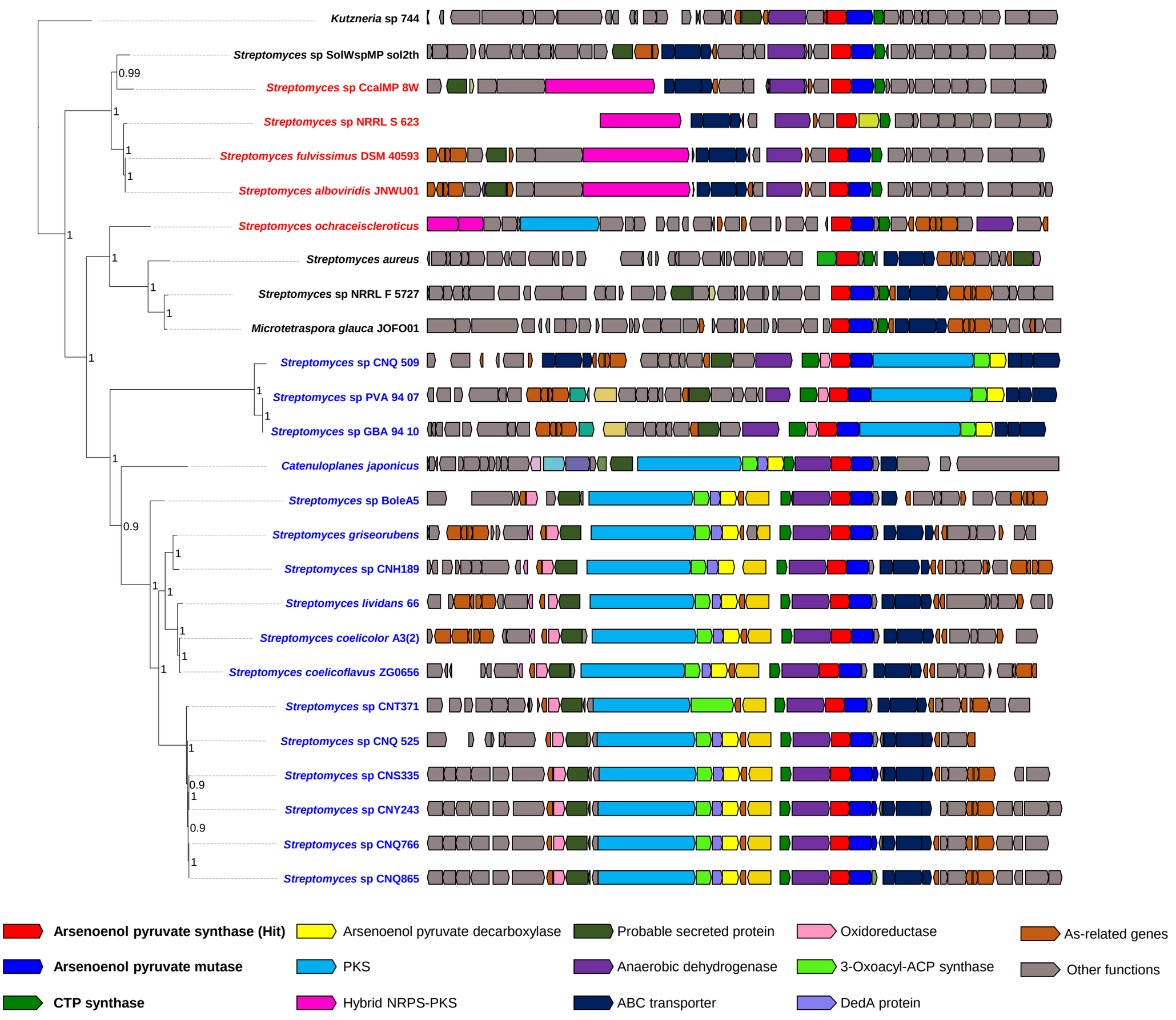
Figure 4. Novel BGCs for arseno-organic metabolites found in *Actinobacteria.* The BGCs were found by mining for the co-occurrence of arsenoenol pyruvate synthase (AEPS), arsenopyruvate mutase (APM). Related BGCs were only found in actinomycetes. The phylogeny was constructed with a concatenated matrix of conserved enzymes among the BGCs including AEPS, APM and CTP synthase. Variations in the functional content of the BGCs are accounted by genes indicated with different colors. Three main classes or arseno-organic BGCs could be expected from this analysis: PKS-NRPS independent (black names), PKS-NPRS-hybrid dependent (red names) and PKS-dependent biosynthetic systems (blue names).


A
Central metabolism DB
Genomes DB
NP seed DB
Homology search
Homology search
Enzyme recruitment Internal database
Phylogenetic analysis
Enzyme family Internal database
Enzyme family expansion analysis
Enzyme expansion Internal database
Fuctional Annotation
BGC prediction


B
Enzyme hits clade
Enzymes from central metabolism


A
Not validated
Validated
15%
Not associated to predicted BGCs
11%
≤ 20 Kbp from BGC
9%
ClusterFinder
7%
AntiSMASH / ClusterFinder
43%
AntiSMASH
15%
Known NP enzymes
Hits
B
43% of the total enzyme hits
Validated hits
Not validated hits
Aspartate transaminase
Histidinolphosphate aminotransferase
Indole-3-Glycerolphosphate synthase
N-acetyl-gamma-glutamyl-phosphate reductase
Histidinol phosphatase
Homoserine O-succinyl transferase
Enolase
Anthranilate phosphoribosyl transferase
Phosphoglycerate dehydrogenase
Citrate synthase
Acetolactate synthase
Glyceraldehyde-3-phosphate dehydrogenase
Aconitate hydratase
Acetyl glutamate kinase
Aspartate kinase
Cysteine synthase
Prephenate dehydrogenase
Acetylornithine aminotransferase
Argininosuccinate lyase
Isopropylmalate dehydrogenase
2-dehydro-3-deoxyphosphoheptonate aldolase
3-Phosphoshikimate-1-carboxyvinyl transferase
Asparagine synthase
Hits per enzyme family
C
Enzymes from known BGCs
Hits validated by AntiSMASH and ClusterFinder
Not validated hits found at less than 20 Kbp from a BGC
Not validated hits
Member of a putative new BGC family
Experimentally confirmed hit


A
SCO6837
SCO6834
SCO6835
SCO6834
SCO6833
SCO6832
SCO6831
SCO6830
SCO6829
SCO6828
SCO6827
SCO6826
SCO6825
SCO6824
SCO6823
SCO6822
SCO6821
SCO6820
SCO6819
SCO6818
SCO6817
SCO6816
SCO6815
SCO6814
SCO6813
SCO6812
KS
AT
DH
ER
KR
ACP1
ACP2
EvoMining hit
SLI_1077
SLI_1078
SLI_1079
SLI_1080
SLI_1081
SLI_1082
SLI_1083
SLI_1084
SLI_1085
SLI_1086
SLI_1087
SLI_1088
SLI_1089
SLI_1090
SLI_1091
SLI_1092
SLI_1093
SLI_1094
SLI_1095
SLI_1096
SLI_1097
SLI_1098
SLI_1099
SLI_1100
SLI_1101
SLI_1102
SLI_1103
Polyketide synthase
C-As bond formation
As regulation and metabolism
Transport
Other enzymes
SLI_1096
SLI_1097
SLI_1091
Arseno-organic NPs
B
hrdB
SLI_1088
SLI_1096
SLI_1099
SLI_1103
3 µM KH2PO4 / 0 µM Na3AsO4
3 µM KH2PO4 / 500 µM Na3AsO4
C
ICP -QQQ-MS
As75 → 91/Ge72 →72, response corrected
ESI-MS
cps on m/z = 351.11
Wild-type + As, As75 → 91
SLI_1096::acc(3)Iv + As As75 → 91
Wild-type - As, As75 → 91
Wild type + As m/z + H = 351.1147 (C14H27AsO5)
retention time [min]


Kutzneria sp 744
Streptomyces sp SolWspMP sol2th
Streptomyces sp CcalMP 8W
Streptomyces sp NRRL S 623
Streptomyces fulvissimus DSM 40593
Streptomyces alboviridis JNWU01
Streptomyces ochraceiscleroticus
Streptomyces aureus
Streptomyces sp NRRL F 5727
Microtetraspora glauca JOFO01
Streptomyces sp CNQ 509
Streptomyces sp PVA 94 07
Streptomyces sp GBA 94 10
Catenuloplanes japonicus
Streptomyces sp BoleA5
Streptomyces griseorubens
Streptomyces sp CNH189
Streptomyces lividans 66
Streptomyces coelicolor A3(2)
Streptomyces coelicoflavus ZG0656
Streptomyces sp CNT371
Streptomyces sp CNQ 525
Streptomyces sp CNS335
Streptomyces sp CNY243
Streptomyces sp CNQ766
Streptomyces sp CNQ865
Arsenoenol pyruvate synthase (Hit)
Arsenoenol pyruvate decarboxylase
Probable secreted protein
Oxidoreductase
As-related genes
Arsenoenol pyruvate mutase
PKS
Anaerobic dehydrogenase
3-Oxoacyl-ACP synthase
Other functions
CTP synthase
Hybrid NRPS-PKS
ABC transporter
DedA protein