Genome mining for the identification of secondary metabolites and alternative metabolic pathway modelling in *Actinobacteria* isolates

Diego Andrés Otero Rodríguez

Advisor: María Mercedes Zambrano PhD; Co-advisor: Alejandro Reyes PhD

Abstract: *Actinobacteria* produce a broad range of natural products (NP), encoded by biosynthetic gene clusters (BCG), with diverse biological activities and special pharmaceutical potential. However, a link between BGC prediction and metabolic pathway design has not been proposed. Genome mining of six Actinobacterial strains, isolated from Colombian environments, revealed 502 gene clusters encoding the biosynthesis of a variety of secondary metabolites, including PKS, NRPS, lanthipeptides and siderophores. Several optimization algorithms where then applied to find alternative, more thermodynamically efficient pathways for NP production. Despite the application of promising new tools, the results indicate that information in databases and modelling still requires improvement for future prediction and optimization of pathways regarding secondary metabolite production.

Keywords: Natural products, biosynthetic gene clusters, secondary metabolites genome mining, metabolic engineering, alternative metabolic pathways

I. INTRODUCTION

Natural products (NP) are a diverse family of compounds with a wide variety of uses due to their biological activities. They can be the result of either primary or secondary metabolism of an enormous range of living beings such as bacteria, fungi, plants and even insects (Katz *et al*., 2016). Notably, NPs have been used in human health, as antibiotics, antifungals and immunosuppressant, animal health and agriculture (Giddings *et al.,* 2013; Newman *et al.,* 2000).

It is estimated that nearly 40% of the New Chemical Entities approved every year by the Food and Drug Administration (FDA) are NPs or NP-related (Kinch *et al.,* 2014). In fact, using NPs as drug candidates is a primary interest of many research projects (Gomez-Escribano *et al.,* 2016). To this day, many massively used antibiotics, and by extension their new generation derivatives, were originally isolated as bacterial NPs. Erythromycin, clavulanic acid and tobramycin are few examples of molecules produced by *Streptomyces* strains which continue to have medical relevance.

Since the discovery of penicillin in 1940, NP discovery has been traditionally carried out using a simple analysis pipeline. First, several soil samples would be collected from different locations, then they would be processed and batch fermented; finally these extracts would be tested against some common human pathogens or cancerous cell lines and hopefully some biological activity would be noticed (Katz *et al*., 2016). This process was later referred to as phenotypic screening and had a huge impact on NP discovery, with over 1000 new molecules described between 1940 and 1970 (Katz *et al*., 2016). Current estimates account for over 23,000 NPs discovered to the present date. Other recent approaches such as target based, mutasynthesis and, recently, genome mining, have also contributed significantly to this number (Katz *et al*., 2016). With the first investigations, it became clear that a large majority of NPs are produced by bacteria, mainly by the

46 *Actinomycetaceae* family within the phylum Actinobacteria, and particularly the *Streptomyces*
47 genus (Berdy, 2012).

48 Based on several studies it has been possible to elucidate NP enzyme structural information
49 (Strieker *et al.,* 2010; Weissman, 2015), distinguishing two main groups: non-ribosomal peptides
50 synthetases (NRPS) and polyketide synthases (PKS). NRPS synthetize NPs by the incorporation of
51 one selected amino acid into a growing polypeptide chain. Both of these BGCs consist of modular
52 enzymes with each module being responsible for the addition of a different domain to a growing
53 amino acid or polyketide chain.

54 Also, it is important to distinguish other groups: Bacteriocins and ribosomally- synthesized and
55 post-translationally modified peptides or RiPPs. Bacteriocins are simple peptic toxins produced by
56 certain bacteria that can inhibit the growth of closely related bacterial strains, because of this;
57 bacterioicns are of interest in the medical field. Some bacteriocins have been classified as RiPPs,
58 especially the ones having lanthionine as part of their structure, however, there have been efforts
59 to create a universal nomenclature for RiPP NPs leading to a differentiation between them and
60 bacteriocins based on encoding gene clusters (Arnison *et al.,* 2013).

61 Early cloning and sequencing experiments revealed that genes responsible for entire NPs were
62 clustered together in the genome. Genes for regulation, resistance and biosynthesis were
63 organized consecutively in groups that often are larger than 50 Kbp (Cundliffe *et al.,* 2008), leading
64 to coinage of the term Biosynthetic Gene Clusters (BGC).

65 By the end of the 80s, the discovery rate of NPs decreased substantially; screening methods were
66 no longer effective at detecting new compounds and resulted in an overall lethargy in the
67 pharmaceutical industry. However, upon sequencing of the first *Streptomyces* strain, *S. coelicolor*
68 A3 (2), in 2002 the situation changed (Bentley *et al.,* 2002). The *S. coelicolor* A3 (2) was found to

contain genes encoding far more NPs than predicted from the known metabolome or production of secondary metabolites. Similar studies also concluded that less than 10% of the genome encoded BGCs were expressed under normal laboratory conditions (Nett *et al.,* 2009, Baranasic *et al.,* 2013), indicating *Streptomyces* as one of the most promising genera for the identification of novel BGCs, which could range from 20 to 50 per genome (Ikeda *et al.,* 2014; Ohnishi *et al.,* 2008).

Next Generation Sequencing (NGS) technologies like Illumina, and more recently PacBio and Nanopore, have allowed rapid and cheap sequencing of multiple strains at a time. These technologies dramatically boosted the number of BGCs discovered, which was in addition enhanced by recovery of strains from diverse sites like fresh and salt water locations (Jensen, 2016). The increase in sequence data has led bioinformatics to become a very important field to manage and analyze the large amounts of produced data. New and more efficient algorithms arose for genome assembly, annotation and prediction and they were used successfully to search for new BGCs. The bioinformatics approach is now the standard first step in this field of investigation; it provides a broad view of a strain's genomic potential and can guide screening efforts for a great number of isolates without the necessity of time-consuming laboratory procedures.

A typical bioinformatics workflow starts with a genome assembly. Most sequencing projects are based on a shotgun sequencing strategy, which generates a staggering amount of data: a simple microbial genome with 50X coverage means data files in the order of tens of gigabytes (Ekblom *et al*., 2014). Prior to assembly, the quality of the reads, GC content, and the proportions of duplicated reads should be assessed. Good starting points are tools like FastQC (Andrews, 2010), which provides helpful summary statistics. It is also necessary to remove low quality data and barcodes left over from the sequencing process, which can be achieved with software such as

ConDeTri (Smeds *et al.,* 2011) and Trimmomatic (Bolger *et al.,* 2014). Another common practice is to remove known vector contamination by using a short read aligner like BWA (Li *et al.,* 2009) or Bowtie2 (Langmead *et al.,* 2012) and then delete all reads matching to the known contamination sequence.

Once the reads have been filtered they can be assembled. Some assembly methods are clearly superior to others; however, it is not easy to know which tools will be better for a given project so it is advisable to use more than one (Ekblom *et al*., 2014). Since the first assemblers were built to work with long reads, mainly generated by Sanger sequencing, they use an algorithm called Overlay Layout Consensus (OLC) for the assembly. OLC builds an overlap graph as the first step, then it bundles stretches of this graph into contigs and finally picks the most likely nucleotide sequence for each segment of overlapping DNA or contig. Popular assemblers that use this strategy are Arachne (Batzoglou *et al.,* 2002), PCAP (Huang *et al.,* 2003) and Celera (Denisov *et al.,* 2008)

Yet, current NGS technologies produce shorter reads, so OLC is no longer a very useful algorithm. Instead, an approach using De Bruijn graphs is preferred. Here, the reads are divided in substrings of variable length called k-mers which form the edges of a directed graph while the nodes are the substrings with length k-mer-1. The assembly is based on finding an Eulerian walk through all the nodes. This is used for assemblers like Velvet (Zerbino *et al.,* 2008), SPAdes (Bankevich et al., 2012), and SOAPdenovo2 (Luo *et al.,* 2012).

How to choose the right k-mer length for each case is another issue. This can be done empirically or by using the software Kmergenie (Chikhi *et al.,* 2014) that estimates the best k-mer length for *de novo* genome assembly. However, assemblers such as SPAdes use a set of multiple k-mers for

its assemblies so they are generally run using their default parameters instead of testing Kmergenie results.

Ideally, the assembly process would generate one contig per bacterial genome, but this is hardly ever the case, in part due to limitation of current sequencing technologies and the difficulties faced by the assembler when there are repeats within the genome. A quality assessment is therefore often needed to ensure a high quality draft genome. BUSCO (Simao et al., 2015) provides quantitative measures based on evolutionary expectations of gene content using 40 almost universal single copy orthologues; that are found within the draft genome using Hidden Markov Models (HMM). The more complete the draft genome, the bigger the number of single copy orthologues found. Another useful tool for assembly metrics is QUAST (Gurevich *et al.,* 2013), which evaluates genome assemblies by computing various metrics, including N50 and L50.

With a high-quality draft genome, it is possible to perform genome annotation, a fundamental step before the identification of BGCs. Primarily, this process gathers evidence from other genomic data to create the initial gene predictions; this is usually done with algorithms trained on gene models from related species. One such program is Augustus (Stanke *et al.,* 2004), which predicts coding sequence (CDS) that can also be supported with protein alignment information if available. This information is then used for gene annotation, usually as an automated task and, following rules for each software. Tools like MAKER (Cantarel *et al.,* 2008) or PASA (Haas *et al.,* 2003) can weigh evidence extracted from several sources and make a unified single annotation. Also there are servers such as RAST (Aziz *et al.,* 2008) which have automated annotation tools and use a slightly different approach: defining subsystems as a set of proteins that perform related functional roles and doing a classification to assign functional roles to genes. The annotation is then performed comparing subsystems of different species. It is worth mentioning that the

success of the annotation procedure would depend heavily on the quality of the assembly, especially when working with larger genomes.

Finally, the BGC prediction is made. There are a number of tools for predicting secondary metabolites like antiSMASH (Medema *et al.,* 2011), Bagel3 (van Heel *et al.,* 2013), NaPDoS (Ziemert *et al.,* 2012) and ClustScan (Starcevic *et al.,* 2008)

antiSMASH is one of the most used computational tools. It performs its own genomic annotation through Glimmer 3 (Salzberg *et al.,* 1998), and then runs HMMER3 (Eddy, 2009) to compare the annotation to protein derived HMM. Finally it, detects PKS or NRPS domains using a preexisting compound library. Newer versions of antiSMASH introduce a new powerful tool: the ClusterFinder algorithm (Cimermancic *et al.,* 2014), which uses evolutionary distances and widely divergent sequences to find unknown classes of BGCs without any experimentally characterized member.

Bagel3 is a mining tool for identification of bacteriocin and ribosomally synthesized and post translationally modified peptides (RiPPPs). It does so by using a special Open Reading Frame (ORF) calling that does not bypass potentially small ORFs encoding for bacteriocins. It also uses context genes to decide whether a bacteriocin might be present or not. Another handy tool is Natural Product Domain Seeker or NaPDoS which uses a phylogenetic based classification to detect ketosynthase (KS) and condensation (C) domains through the use of BLASTX against the domains database. NaPDoS can also use a BLASTP using a six frame translated version of the input data.

The next logical step would be to validate putative BGCs identified by bioinformatics. This can be done by expressing these BGCs in a host strain, but to do so it is necessary to develop new strategies for heterologous expression, gene cluster amplification, transposon mutagenesis or overexpression of positive regulators (Ziemert *et al.,* 2016). Nevertheless, at this point, a set of new obstacles arise: assuming there already is a bioengineered strain that can produce a newly

identified NP, how can this production be escalated to industrial levels? Is there a way to increase the yield? How can the process be as cheap as possible? What carbon or nitrogen sources would be the most ideal to use? Clearly, even after the identification and expression of BGCs there are a lot of optimization issues to face.

One possible alternative for optimization is to exploit information of NP metabolic pathways. By doing so, it could be possible to find either a more cost-effective, or a more thermodynamically favorable pathway, to achieve the desired product or even block some routes that are consuming the reactants needed.

Reported computational tools for pathway design rely on graph search or optimization techniques proposals based on linear programming (LP) or mixed integer linear programming (MILP) (Thiele *et al.,* 2010). The purpose is to extract a minimal stoichiometry balanced sub-network that converts a source metabolite into a desired product with maximum yield. These networks have recently achieved genome scale size (Hamilton *et al.,* 2014). The downside in these procedures is that they usually fail to find the optimal conversion stoichiometry and therefore they do not reach maximum yield. Such issues prompt the development of a pathway designing tool that first optimizes the overall stoichiometry (i.e., OptStoic) (Chowdhury *et al.,* 2015) by finding co-reactant and co-products coefficients. It then identifies the reactions that must be involved, from a database (i.e., minRxn) (Chowdhury *et al.,* 2015), and finally ties the reactants and products with the desired stoichiometric ratios (i.e., minFlux) (Chowdhury *et al.,* 2015), all of this while maintaining thermodynamic feasibility.

This project aims to follow a reported bioinformatics pipeline (Alzate *et al.,* 2015) which includes genome assembly, annotation and prediction to identify BGCs encoded in the genomes of Actinobacteria isolated from Colombian samples. The microbial isolates studied here were

obtained from soil and saline samples, taken at two different sites: Parque Nacional Natural Los Nevados (PNN) and the Zipaquirá salt mines. The isolates were originally characterized microscopically and macroscopically by growth on various media and for their capacity to produce metabolites against either bacteria or cancerous cell lines (Cantillo *et al.,* 2015) (**Table1**). Bacterial identification was done using traditional PCR, sequencing and analysis using BLAST and RDP (Cole et al., 2014). In this work, we analyzed the complete genome of six isolates and ran optimizations algorithms to seek more efficient pathways to produce NPs encoded by the identified BGCs.

II. MATERIALS AND METHODS

Genome sequence analysis

Draft genomes were generated at the DOE Joint Genome Institute (JGI) using Illumina technology. An Illumina 300 bp insert shotgun library was constructed and sequenced using the Illumina MiSeq platform. Raw sequence data was filtered using BBDuk and BBMAP (Bushnell, 2014). An initial assembly was also performed at the JGI using Velvet and Allpaths along with an *in silico* generated reads library. Raw reads were downloaded and also used on different assemblers.

For this, sequence reads were quality controlled using FastQC (V 0.11.5) and then trimmed with the default set of options of Trimmomatic (V 0.36). FastQC analyses were repeated to verify improvement on read quality.

Genome assemblies

The most adequate k-mer length for each set of reads was calculated using Kmergenie (V 1.7016). The first assembly was done using Velvet (V 1.2.10) with default option for paired -end reads and k-mer lengths of 29 bp for *Nesterenkonia sandarakina, Isoptericola halotolerans*, *Isoptericola sp* (CG1183) and 31 bp for *Streptomyces sp.* (CG 926), *Streptomyces avidinii* and *Streptomyces*

microflavus . SPAdes (V 3.11) assembler was also tested using k-mer sizes near those calculated by Kmergenie **(Table 2).** Quality of FASTA files coming out of the assembly was assessed using BUSCO (V 3), while assembly metrics were calculated using the Quast (V 4.5) web server. Additionally, 16S rRNA sequences of the six strains were used to create a phylogenetic tree using MrBayes (Huelsenbeck *et al.,* 2001), a free software program which performs Bayesian inference of phylogeny. The analysis was run using a GTR substitution matrix over 100,000 generations with a sample tree taken every 100 generations.

Annotation and BGC mining

Genome annotation was carried out through RAST (V 0.11.5) web server, using Classic RAST, Fix Frameshifts and Debug options. BGC prediction was performed using three different approaches: antiSMASH (V 3.0.5), Bagel3 and NaPDoS. antiSMASH was run on annotated genomes using three additional features provided by the web server: ClusterFinder algorithm, whole-genome PFAM analysis and EC number prediction.

Bagel3 was run using default options, and finally NaPDoS was run four times for each genome, using 1) the FASTA file as an input and seeking for KS domains, 2) using the same input and seeking for C domains. The remaining two runs were with the FASTA aminoacid files (obtained in RAST annotation) and seeking for KS and C domains.

Pathway design

The bacillobactin-like (only reported on *B. subtilis*) BGC of the *Streptomyces sp.* (CG 926) genome was used to define the following objective function:

$$C_6H_{12}O_6 + H^+ + NH_3 + C_{10}H_{10}O_6 \rightarrow C_{39}H_{42}N_6O_{18} + CO_2 + H_2O$$

Chorismate ($C_{10}H_{10}O_6$) is used in the equation as a direct precursor for bacilibactin ($C_{39}H_{42}N_6O_{18}$) biosynthesis.

To find alternative pathways, three different algorithms were used: OptStoic, MinRxn and MinFlux. These algorithms came with a pre-built reaction, metabolites and stoichiometric matrix database they need to perform the analysis. Before running OptStoic, it was necessary to include in the databases a considerable amount of metabolites and reactions reported for *Streptomyces* and *B. subtilis,* in order for the software to function correctly. The required metabolites were downloaded from the KEGG (Kanehisa *et al.,* 2000) database and reactions from Metacyc (Caspi *et al.,* 2014).

To analyze the minimum set of reactions and the flux for the alternate pathway, a new stoichiometric matrix was constructed using an Excel macro.

The objective function underwent changes to see different results from the algorithms tested and to find the one that would yield bacillobactin using the simplest reactants (**Table 3**).

III. RESULTS

Microbial isolates obtained from soil and saline samples were originally chosen based on their capacity to produce metabolites against either bacteria or eukaryotic cell lines (Cantillo *et al.,* 2015) (**Table 1**). Isolates were further characterized by the analysis of the 16S rRNA gene. These strains were identified as: *Isoptericola halotolerans*, *Isoptericola sp* (CG1183), *Nesterenkonia sandarakina*, *Streptomyces sp.* (CG 926)*, Streptomyces avidinii* and *Streptomyces microflavus*. Given that one *Isoptericola* and one *Streptomyces* strain could not be identified at the species level, they will be referred to as *Isoptericola sp.* (CG1183) and *Streptomyces sp.* (CG 926). A posterior FastQC analysis of the sequenced data obtained from the JGI showed that reads had a

low quality towards the end for all the six strains studied which is somewhat common for Illumina sequencing. This can be seen for strain Streptomyces *sp.* (CG 926) which had a quality Phred socre of 19 (**Figure 1A**).

After running Trimommatic, which removes low quality data, the mean quality towards the end of the read improved to at least 25 Phred score (**Figure 1B**). Next, a Kmergenie run was performed on every genome to estimate the most appropriate k-mer length for the assembly. For *I. halotolerans* and *Isoptericola sp.* (CG1183) the best length was 29 bp, 107 bp for *N. sandarakina* , 91 bp for *Streptomyces sp.* (CG 926), and 117bp for *S. avidinii* and *S. microflavus.*

Velvet assemblies were then carried out using default options for paired-end reads. However, the tool used a k-mer length of 31 bp for all *Streptomyces* assemblies, which was very different from the number reported by Kmergenie. This led to a fragmented and incomplete assembly of those strains, with over 500 contigs and less than half of the single copy orthologues found using BUSCO (**Table 4**), indicating a low quality of assembly.

SPAdes generated a better assembly than Velvet, as shown by the QUAST metrics. However, these same metrics indicated that the best assemblies were the ones obtained from the JGI (**Table 4**). Based on these results, the annotation, prediction and metabolic analysis were carried using the assemblies provided by the JGI, together with the raw sequence data which showed the best metrics (**Table 4**).

MrBayes results (**Figure 2**) showed that the six strains can be classified in two different groups. One contains the two *Isoptericola* isolates plus *N. sandarakina* while the other one contains the three *Streptomyces* isolates. On the first group both *I. halotolerans* and *Isoptericola sp.* (CG1183) are closely related, and since the tree doesn't show new branches towards them, it is possible that they might even be the same species. On the second group, the same closeness is also present

but now involving *S. avidinii* and *S. flavus,* while *Streptomyces sp.* (CG 926) is on another branch but still more similar to the two *Streptomyces* strains than to *N. sandarakina* to the *Isoptericola* strains.

Annotation was carried out for each genome with the RAST web server. These annotations reported a mean of 5,182 genes per strain for all *Streptomyces* strains, which had almost 7,000 genes while *Isoptericola* and *Nesterenkonia* isolates had around 3,000 genes (**Table 5**). The RAST annotation for *Streptomyces sp* (CG 926) can be seen in **Figure 3**.

BGC predictions on the annotated genomes were then carried out using various tools.

antiSMASH found 342 BGCs in all six genomes: 38 on *I. halotolerans,* 36 on *Isoptericola sp.* (CG1183)*,* 35 on *N. sandarakina,* 93 on *Streptomyces sp.* (CG 926), 72 on *S. avidinii and* 68 on *S. microflavus.* 187 (56.67%) of these 342 BGCs were putative clusters identified by the ClusterFinder algorithm incorporated in antiSMASH, many of these clusters can be considered as novel BGCs. 32 (9.35%) of these BGCs were for putative saccharides and 23 (6.72%) were for fatty acids. This information means that more than half of the total amount of BGC could only be detected using the antiSMASH implementation of the ClusterFinder algorithm.

The rest of the 342 clusters identified in these six genomes with antiSMASH were classified as follows: 23 (6.72%) were NRPS clusters, 19 (5.55%) were terpene related clusters, 18 (5.26%) corresponded to PKS clusters, 9 (2.63%) were siderophore-related clusters, 5 (1.46%) were for ectoine and butyrolactone synthesis, 2 (0.58%) for lassopeptide, lantipeptide and betalactamic clusters and finally 1 (0.29%) for melanin synthesis.

The use of other algorithms revealed different gene clusters for synthesis of NPs. Bagel3 found 16 bacteriocin clusters in the six annotated genomes. From these, one cluster was found on the *I.*

293 *halotolerans* genome, one in *Isoptericola sp* (CG1183), six were found in *Streptomyces sp* (CG 926),
294 four in *S. avidinii,* and four in *S. microflavus.*

295 From this total of 16 clusters, seven (43.75%) are Class III high molecular weight bacteriocins, six
296 (37.5%) are lantipeptides, two (12.5%) are lassopeptides and one (6.35%) is a tiopeptide (**Figure 4**).
297 These results indicate that these genomes also have the potential to produce bacteriocins, even
298 though only a few were found per genome.

299 Finally, the use of NaPDoS predicted 149 NPs with KS or C domains on the annotated genomes: 76
300 matched the C domain and 73 were identified as KS domain. Of the 149 domains, 2 were found *I.*
301 *halotolerans*, two in *Isoptericola sp.* (CG1183), four in *N. sandarakina*, 70 in *Streptomyces sp.* (CG
302 926), 41 in *S. avidinii* and finally 30 were found in *S. microflavus.*

303 The NaPDoS tool also performs a classification that takes into account the metabolic pathway
304 involved in product formation. By doing so, the 149 domains were identified as follows: 67
305 (44.96%) were for NRPS products, 11 (7.38%) for fatty acid synthesis products, nine (6.04%) were
306 hybrid products and two (1.34%) were polyunsaturated fatty acids products (**Figure 5**).

307 In total, the three different tools used to predict BGC identified a total of 502 putative BGCs with
308 capacity to produce a broad range of NPs. Of the tools used, antiSMASH found the most clusters
309 (**Table 6**)**.** It also became clear that the genomes varied greatly in the number of BGCs identified
310 and that the *Streptomyces* strains had the largest biosynthetic potential (**Table 6).** Of these strains,
311 Streptomyces *sp* (CG 926) contained the most predicted BGCs.

312 One of the identified clusters in *Streptomyces sp.* (CG 926), which had 47 genes, showed on
313 average 38% similarity to the bacillobactin BGC, a catechol based siderophore, first described in *B.*
314 *subtilis (*Crosa *et al.,* 2002*).* According to the antiSMASH annotation 15 out of 47 genes in this

bacillobactin cluster were responsible for biosynthesis. These genes had an average of 72% identity to the biosynthesis bacillobactin genes in *B.subtilis* according to a BLASTP analysis. Based on this analysis, and the fact that the genes have been previously reported, this bacillobactin cluster was used for subsequent studies using computational tools for metabolic pathways modelling.

To carry out the modelling of alternative bacillobactin biosynthesis pathways, the *S. coelicolor* (A3)2 and *B. subtilis* reactions databases were downloaded from Metacyc, a highly curated metabolic pathways database. These included 1,655 and 1,159 reactions from *S.coelicolor* A3 (2) and *B.subtilis*, respectively. However, 319 and 156 reactions for these strains, respectively, were not considered as they did not have a specific Enzyme Commision number identifier needed to do a posterior cross reference with the database of OptStoic, a tool that identifies optimum overall stoichiometry (Chowdhury *et al.,* 2015). The total number of reactions was therefore reduced to 1,336 and 1,003 for each organism for a total of 2,339 reactions. Using these 2,339 reactions, a cross reference was made with OptStoic's database and 532 new reactions and 11418 new metabolites were added.

OptStoic algorithm's output is a .csv file with information about the product, objective value, Delta Free Gibbs Energy and the stoichiometric coefficient of each metabolite allowed to participate in a reaction. With an established database of reactions and metabolites, 21 simulations were performed (**Table 3),** and only 5 were thermodynamically feasible (**Table 7**). Notably, reactions that did not use chorismate as an intermediate were always unfeasible.

The algorithms MinRxn and MinFlux identify the minimal amount of intervening reactions and fluxes to meet overall stoichiometry obtained by OptStoic (Chowdhury *et al.,* 2015). They need a stoichiometric matrix, which associates each reaction with the stoichiometric coefficient of both

reactants and products, in order to run properly. Since new reactions were added to the reactions database, a new matrix was needed as well. To build it, 14 out of de 532 reactions had to be deleted. This was done because the same compound appeared on both sides of the equation as it was acting like a chemical catalyst, which cannot be accepted by the Metadat2014 macro. Finally the matrix and its vector, where all the stoichiometric coefficients are collected to a vector, were generated from 518 reactions. The results from modelling experiments showed all 5 simulations obtained by OptStoic to be unfeasible using current metabolites and reactions database. This means that even if pathways different to the one reported for bacillobactin biosynthesis are thermodynamically feasible, they cannot be achieved using metabolites and reactions reported for *S. coelicolor* A3 (2) and *B. subtilis*.

IV. DISCUSSION

Actinobacteria have already been reported as a very productive bacterial phylum in terms of NP production. In this work, results showed that several BGCs could be identified in six *Actinobacteria* genomes examined, and that of these, the *Streptomyces* isolates were well above the known average for each genome (Ikeda *et al.,* 2014; Ohnishi *et al.,* 2008). Indeed the *Streptomyces sp.* (CG 926) genome had over 160 potential BGCs identified with 3 different tools, when the mean of clusters found in similar species has been reported to be between 40 and 60 (Ikeda *et al.,* 2014). This again, indicates these isolates are promising on the quest to find novel NPs with clinically relevant biological activities (Barka *et al.,* 2016).

With the addition of new tools for genome mining, it seems clear that the actual number of BGCs a genome can have is larger than the numbers reported based on experimental data. Taking into account that most of the JGI assemblies were composed of over 40 contigs (**Table 4**), it is very likely that future improvements on genome sequencing and assembly will be able to yield even more robust data for BGC identification, making genome mining a far more complex but more

362 informative procedure (Lee *et al.,* 2016). A high number of contigs means the genome is more
363 fragmented, and some parts of it could even be missing, making it difficult to correctly identify
364 BGCs. The goal behind the use of several assembly strategies was to obtain draft genomes as little
365 fragmented as possible; however, JGI assemblies were by far the best ones. This indicates that
366 Velvet and SPAdes are not good enough when used on default parameters and that some
367 modifications on such parameters would be necessary to achieve a better assembly. For instance,
368 it might have been necessary to try more k-mer lengths around the number predicted by K-
369 mergenie. Besides, coverage cutoff value used by Velvet could also have been modified. Finally,
370 JGI assemblies get rid of all the contigs shorter than 1 Kbp while Velvet and SPAdes only discarded
371 contigs shorter than 500 bp. Therefore while the former assemblies contain fewer contigs with
372 better N50 and L50 metrics, they are losing more information about the genomes.

373 The so called third generation sequencing technologies, such as PacBio and Nanopore are already
374 producing longer reads than Illumina. However, at the time the isolates were sequenced, the error
375 rate on base calling was still too high for these to be suitable options for *de novo* genome
376 assemblies (Lee *et al.,* 2016). Since current error rates are considerably reduced, longer reads not
377 only mean easier assemblies but a better genome annotation and therefore more accurate BGC
378 predictions. That is why today; a combination between PacBio and Illumina sequencing is
379 preferred for similar projects (Ziemert *et al.,* 2016).

380 All strains analyzed in this study contained multiple BGCs, determined through the use of diverse
381 tools. Even though, *Isoptericola* strains were not as rich in BGCs as the *Streptomyces* strains, they
382 did show a fair amount of BGCs on their genomes, most notably bacteriocins identified by Bagel3.
383 Considering that these bacteria were isolated from a saline environment, their BGCs could encode
384 for metabolites very different from the rest, because they would be intended to face a different

set of competitors and to gain fitness advantage in a more extreme environment (Atanasova *et al.,* 2013), contrary to the *Streptomyces* isolated from soils.

Considering the phylogenetic relation between the strains (**Figure 2**), the very similar number of BGCs identified on both *Isoptericola* strains was to be expected. Taking into consideration that both of the strains also share the same number of genes, it is very plausible that they are the same species. The phylogenetic inference shows that the six strains can be divided into two groups, one containing the aforementioned *Isoptericola* strains plus *N. sandarakina* and a second group with only the *Streptomyces* genus. This distinction can also be made when referring to the number of BGCs on each genome, as the number of clusters within the *Streptomyces* group is far larger than the number within the *Isoptericola* and *N. sandarakina* group thus hinting at the possibility that a phylogenetic study of Actinobacteria could be indicator of an expected number of clusters one could find on specific genomes as they could be well preserved on closely related species (Medema *et al.,* 2014).

It was interesting to note that more than half of the predicted BGCs were only found using the ClusterFinder algorithm. As this algorithm is based on genealogy of not so closely related species and it finds sequences on the genome similar to already reported BGC sequences, some of these results could be false positives. An experimental approach would therefore be needed to confirm if the identified BGCs actually encode for these related compounds. In this case, techniques like MS/MS could come in handy to probe the metabolome of these isolates grown under different culture conditions. In fact, there have been efforts to unify mass spectrometry results with BGC *in silico* predictions (Sidebottom *et al.,* 2013). If the compounds are detected, then they could be added to public BGC databases like MIBiG (Medema *et al.,* 2015), or Norine (Caboche *et al.,* 2008),

407 which would help future research projects because the antiSMASH database is periodically
408 updated with MIBiG results.

409 From a technical standpoint, knowing the shortest pathway to a product of interest is
410 advantageous, as this potentially minimizes the efforts aimed at production. This can be achieved
411 by carrying out modelling of the biosynthetic pathways for a particular product and searching for
412 alternative routes that could benefit biosynthesis. However, it has been argued that optimized
413 metabolic systems naturally evolve towards the state of maximum entropy production (Unrean *et*
414 *al.,* 2011), which means that evolution itself tries to reduce genetic and metabolic efforts to
415 produce a given a product to save the most amount of energy as possible. In addition, currently
416 available tools for metabolic analysis are based on information of core metabolism (Hamilton *et*
417 *al.,* 2014), which is expressed during the logarithmic growth phase of a microorganism. In this
418 study we attempted to model biosynthesis of bacillobactin, for which there is only one reported
419 pathway (Crosa *et al.,* 2002). Despite this it seemed plausible that the desired product could be
420 obtained via alternate routes.

421 Given that BGCs encode for compounds generally considered as secondary metabolites that
422 originate mainly after exponential growth, as cells enter stationary phase, classic algorithms and
423 softwares may not be the most appropriate approach for analysis of BGC metabolism. However,
424 some tools, like OptStoic can be useful for the metabolic predictions of BGCs. OptStoic is based on
425 some premises and constraints that minimize the amount of flux and reactions on a pathway, or
426 maximize the intake of a carbon source (Henry *et al.,* 2010; Chowdhury *et al.,* 2015). As such, they
427 can be used as basis for work with secondary metabolites. In order to study pathway designs for
428 bacillobactin synthesis, it was essential to start with a large metabolite and reaction database to
429 provide as many options as possible. This can be analogous to a graph with edges and nodes; the

less edges the graph has, the more difficult it will be to arrive at a certain node. However, by adding more information (edges) the nodes would be more connected to the network.

The results obtained after running OptStoic seem to show that bacillobactin production is confined to a simple metabolic pathway and cannot be obtained through intermediates other than chorismate, or via shorter steps. Chorismate is a known precursor of many different molecules like amino acids (tyrosine and phenylalanine), indol, salicylic acid and vitamin K, among others. In this case it seems that chorismate is also required for bacillobactin synthesis. From a metabolic standpoint, current information about reported enzymatic reactions and metabolites from Metacyc and KEGG databases was not enough to find alternate pathways for bacillobactin production using either OptStoic or MinRxn/MinFlux. These tools resulted in thermodynamically unfeasible solutions, meaning that based on the available information no other path can be found for the synthesis of bacillobactin, other than the one reported.

Evolutionary studies have found that genes needed for essential metabolism such as carbohydrate degradation and nucleic acid production are more evolutionary conserved than non-essential genes because purifying selection acting on essential genes is expected to be far more stringent. This means that genetic diversity tends to be very low for these genes, increasing their preservation rate (Jordan *et al.,* 2002). Basically, in essential pathways there is more than one gene that can supply the demand of a particular metabolite, ensuring availability even if the primary source of production is blocked. This level of preservation is lost for non-essential metabolites. It is not advantageous for a microorganism to have multiple sets of genes that encode for a compound that is not required for growth and reproduction, even if such metabolism can increase the cell's fitness and its capability to adapt to a changing environment (Luo *et al.,* 2015). Because of this, metabolites like bacillobactin are usually translated from a single set of

genes, having no other backup to produce them. In such cases, alternative pathway finder algorithms like OptStoic would fail to find additional potential biosynthetic routes. The question remains whether this is true for all the secondary metabolites or if in some cases it is still possible to find simpler, more efficient routes for some other metabolites like erythromycin.

Ultimately this means that the optimizations problem must be tackled using a different approach. For example, the bacillobactin BGC could be overexpressed in a heterologous host strain, causing the intended bacillobactin producing phenotype. Alternatively, some of the pathways involving chorismate consumption could be turned off using gene knockout allowing for the chorismate to be more biologically available for a cell to use. Reported tools like Optknock that suggests gene deletion strategies leading to the overproduction of a given metabolite using a metabolic model, could become very handy in this process (Burgard *et al.,* 2003). Another approach would be to use metabolic engineering to construct a minimal bacterial cell where only the essential metabolism is active (Gil *et al.,* 2004); then, only non-essential pathway added would be the ones involving bacillobactin biosynthesis or another secondary metabolite of interest. This strategy might ensure the cell is using as much resources as possible to synthetize the desire product.

V. CONCLUSIONS

This research showed that bacteria belonging to the *Actinobacteria* phylum are very promising for NP discovery, based on the amount of BGCs identified. Diversity within BGCs was also staggering and comprised molecules like NRPS, PKS, and RiPPS, among others. Nevertheless, the big majority of described BGCs corresponded to not experimentally confirmed molecules which mean that laboratory procedures like MS/MS or HPLC are necessary to assess the existence of novel compounds identified using this approach. Finally, if the final goal is to produce one of these compounds at an industrial scale, a confirmation based on heterologous expression would also be required.

In terms of pathway design, the information of current databases on metabolic processes is not enough to identify alternative pathways leading to synthesis of a secondary metabolite. Future discoveries of new microorganisms, metabolites and enzymatic reactions can allow for the design of more efficient ways to obtain the desired product.

V. REFERENCES

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., … van der Donk, W. A. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural Product Reports*, *30*(1), 108–160. https://doi.org/10.1039/c2np20085f

Alzate, J. D., Restrepo, S., Reyes, A. (2015). Bioinformatics workflow and assessment of software to seek secondary metabolites in Bacteria. *Universidad de los Andes,* masters' thesis

Atanasova, N. S., Pietilä, M. K., & Oksanen, H. M. (2013). Diverse antimicrobial interactions of halophilic archaea and bacteria extend over geographical distances and cross the domain barrier. *MicrobiologyOpen*, *2*(5), 811–825. https://doi.org/10.1002/mbo3.115

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., … Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, *9*, 75. https://doi.org/10.1186/1471-2164-9-75

Baltz, R. H. (2016). Genetic manipulation of secondary metabolite biosynthesis for improved production in Streptomyces and other actinomycetes. *Journal of Industrial Microbiology & Biotechnology*, *43*(2–3), 343–370. https://doi.org/10.1007/s10295-015-1682-x

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Baranasic, D., Gacesa, R., Starcevic, A., Zucko, J., Blažič, M., Horvat, M., … Petković, H. (2013). Draft Genome Sequence of Streptomyces rapamycinicus Strain NRRL 5491, the Producer of the Immunosuppressant Rapamycin. *Genome Announcements*, *1*(4). https://doi.org/10.1128/genomeA.00581-13

Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H.-P., … van Wezel, G. P. (2016). Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *80*(1), 1–43. https://doi.org/10.1128/MMBR.00019-15

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., … Lander, E. S. (2002). ARACHNE: a whole-genome shotgun assembler. *Genome Research*, *12*(1), 177–189. https://doi.org/10.1101/gr.208902

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., … Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, *417*(6885), 141–147. https://doi.org/10.1038/417141a

Bérdy, J. (2012). Thoughts and facts about antibiotics: Where we are now and where we are heading. *The Journal of Antibiotics*, *65*(8), 385. https://doi.org/10.1038/ja.2012.27

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, *84*(6), 647–657. https://doi.org/10.1002/bit.10803

Bushnell B. BBMap short read aligner. Accessed 16 March 2017. https://sourceforge.net/projects/bbmap

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., & Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. *Nucleic Acids Research*, *36*(Database issue), D326–D331. https://doi.org/10.1093/nar/gkm792

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Cantillo, A. P., Zambrano, M.M., Vives, S. J. (2015). Identificación de microorganismos halófilos y halotolerantes con actividad antimicrobiana y citotóxica. *Universidad de los Andes,* masters thesis

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., … Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, *42*(Database issue), D459-471. https://doi.org/10.1093/nar/gkt1103

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31–37. https://doi.org/10.1093/bioinformatics/btt310

Chowdhury, A., & Maranas, C. D. (2015). Designing overall stoichiometric conversions and intervening metabolic reactions. *Scientific Reports*, *5*, 16009. https://doi.org/10.1038/srep16009

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., … Fischbach, M. A. (2014). Insights into Secondary Metabolism from a Global Analysis of

Prokaryotic Biosynthetic Gene Clusters. *Cell*, *158*(2), 412–421. https://doi.org/10.1016/j.cell.2014.06.034

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633–D642. https://doi.org/10.1093/nar/gkt1244

Crosa, J. H., & Walsh, C. T. (2002). Genetics and Assembly Line Enzymology of Siderophore Biosynthesis in Bacteria. *Microbiology and Molecular Biology Reviews*, *66*(2), 223–249. https://doi.org/10.1128/MMBR.66.2.223-249.2002

Cundliffe, E. (2008). Control of tylosin biosynthesis in Streptomyces fradiae. *Journal of Microbiology and Biotechnology*, *18*(9), 1485–1491.

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., & Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics (Oxford, England)*, *24*(8), 1035–1040. https://doi.org/10.1093/bioinformatics/btn074

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23*(1), 205–211.

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Giddings, L.-A., & Newman, D. J. (2013). Microbial natural products: molecular blueprints for antitumor drugs. *Journal of Industrial Microbiology & Biotechnology*, *40*(11), 1181–1210. https://doi.org/10.1007/s10295-013-1331-1

Gil, R., Silva, F. J., Peretó, J., & Moya, A. (2004). Determination of the Core of a Minimal Bacterial Gene Set. *Microbiology and Molecular Biology Reviews*, *68*(3), 518–537. https://doi.org/10.1128/MMBR.68.3.518-537.2004

Gomez-Escribano, J. P., Alt, S., & Bibb, M. J. (2016). Next Generation Sequencing of Actinobacteria for the Discovery of Novel Natural Products. *Marine Drugs*, *14*(4). https://doi.org/10.3390/md14040078

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., … White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Hamilton, J. J., & Reed, J. L. (2014). Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, *16*(1), 49–59. https://doi.org/10.1111/1462-2920.12312

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, *28*(9), 977–982. https://doi.org/10.1038/nbt.1672

Huang, X., Wang, J., Aluru, S., Yang, S.-P., & Hillier, L. (2003). PCAP: A Whole-Genome Assembly Program. *Genome Research*, *13*(9), 2164–2170. https://doi.org/10.1101/gr.1390403

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, *17*(8), 754–755.

Ikeda, H., Kazuo, S., & Omura, S. (2014). Genome mining of the Streptomyces avermitilis genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. *Journal of Industrial Microbiology & Biotechnology*, *41*(2), 233–250. https://doi.org/10.1007/s10295-013-1327-x

Jensen, P. R. (2016). Natural Products and the Gene Cluster Revolution. *Trends in Microbiology*, *24*(12), 968–977. https://doi.org/10.1016/j.tim.2016.07.006

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, *12*(6), 962–968. https://doi.org/10.1101/gr.87702

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Katz, L., & Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology*, *43*(2–3), 155–176. https://doi.org/10.1007/s10295-015-1723-5

Kinch, M. S., Haynesworth, A., Kinch, S. L., & Hoyer, D. (2014). An overview of FDA-approved new molecular entities: 1827-2013. *Drug Discovery Today*, *19*(8), 1033–1039. https://doi.org/10.1016/j.drudis.2014.03.018

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., … Schatz, M. (2016). Third-generation sequencing and the future of genomics. *bioRxiv*, 048603. https://doi.org/10.1101/048603

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Luo, H., Gao, F., & Lin, Y. (2015). Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Scientific Reports*, *5*. https://doi.org/10.1038/srep13210

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Medema, M. H., Cimermancic, P., Sali, A., Takano, E., & Fischbach, M. A. (2014). A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Computational Biology*, *10*(12), e1004016. https://doi.org/10.1371/journal.pcbi.1004016

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., … Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, *39*(Web Server issue), W339–W346. https://doi.org/10.1093/nar/gkr466

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., … Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*, *11*(9), 625–631. https://doi.org/10.1038/nchembio.1890

Nett, M., Ikeda, H., & Moore, B. S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural Product Reports*, *26*(11), 1362–1384. https://doi.org/10.1039/b817069j

Newman, D. J., Cragg, G. M., & Snader, K. M. (2000). The influence of natural products upon drug discovery. *Natural Product Reports*, *17*(3), 215–234.

Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., … Horinouchi, S. (2008). Genome Sequence of the Streptomycin-Producing Microorganism Streptomyces griseus IFO 13350. *Journal of Bacteriology*, *190*(11), 4050–4060. https://doi.org/10.1128/JB.00204-08

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, *26*(2), 544–548.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smeds L, Kunstner A (2011) CONDETRI - A Content Dependent Read Trimmer for Illumina Data. PLoS ONE 6(10): e26314. doi:10.1371/ journal.pone.0026314

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(Web Server issue), W309-312. https://doi.org/10.1093/nar/gkh379

Starcevic, A., Zucko, J., Simunkovic, J., Long, P. F., Cullum, J., & Hranueli, D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Research*, *36*(21), 6882–6892. https://doi.org/10.1093/nar/gkn685

Strieker, M., Tanović, A., & Marahiel, M. A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Current Opinion in Structural Biology*, *20*(2), 234–240. https://doi.org/10.1016/j.sbi.2010.01.009

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Unrean, P., & Srienc, F. (2011). Metabolic networks evolve towards states of maximum entropy production. *Metabolic Engineering*, *13*(6), 666–673. https://doi.org/10.1016/j.ymben.2011.08.003

van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J., & Kuipers, O. P. (2013). BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Research*, *41*(Web Server issue), W448–W453. https://doi.org/10.1093/nar/gkt391

Weissman, K. J. (2015). The structural biology of biosynthetic megaenzymes. *Nature Chemical Biology*, *11*(9), 660. https://doi.org/10.1038/nchembio.1883

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Ziemert, N., Alanjary, M., & Weber, T. (2016). The evolution of genome mining in microbes - a review. *Natural Product Reports*, *33*(8), 988–1005. https://doi.org/10.1039/c6np00025h

Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E., & Jensen, P. R. (2012). The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLOS ONE*, *7*(3), e34064. https://doi.org/10.1371/journal.pone.0034064

FIGURES

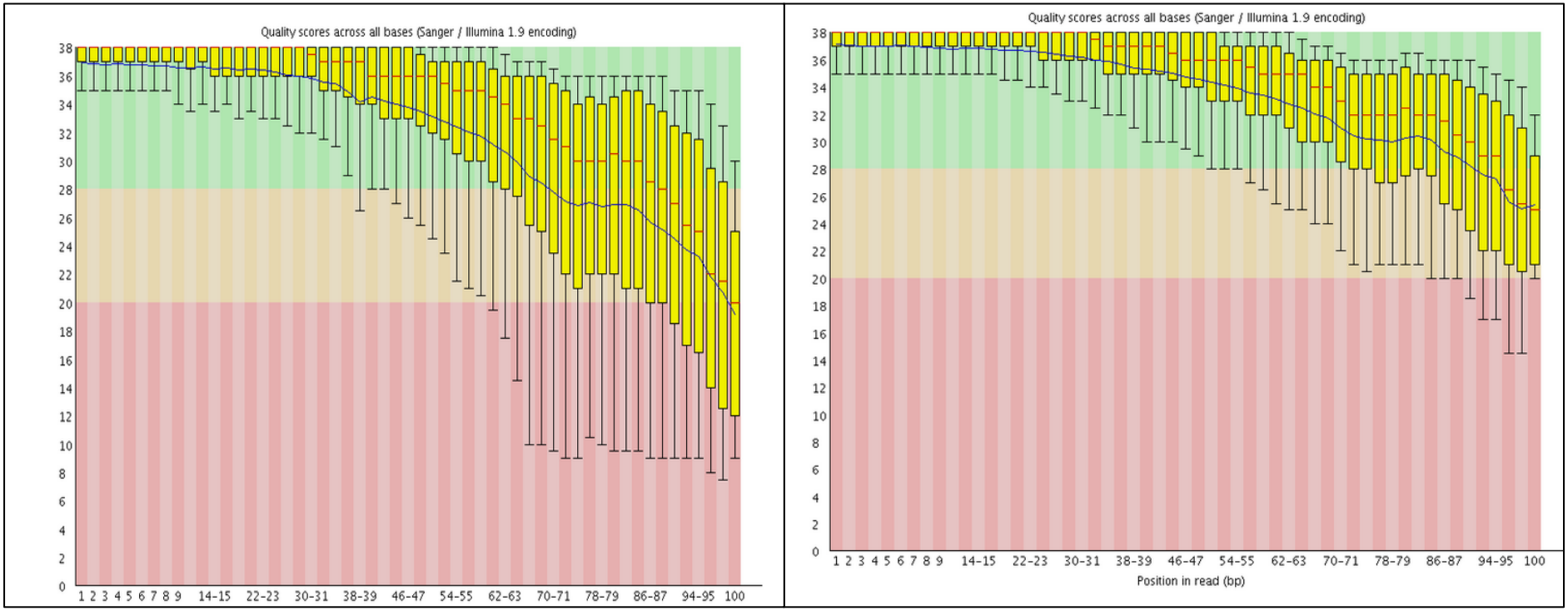

Figure 1. Quality analysis *of Streptomyces sp* (CG 926) forward sequence reads. The Y-axis on the graph shows the quality scores and the x-axis represents the position in the read (bp). The yellow box indicates the interquartile range (25%-75%), the central red line is the median value and the blue line indicates the mean quality. The upper and lower whiskers represent the 10% and 90% data points. (A) Quality analysis of the raw reads with a mean quality of 19. (B) Quality analysis of the reads after Trimmomatic, mean quality improves to 25.

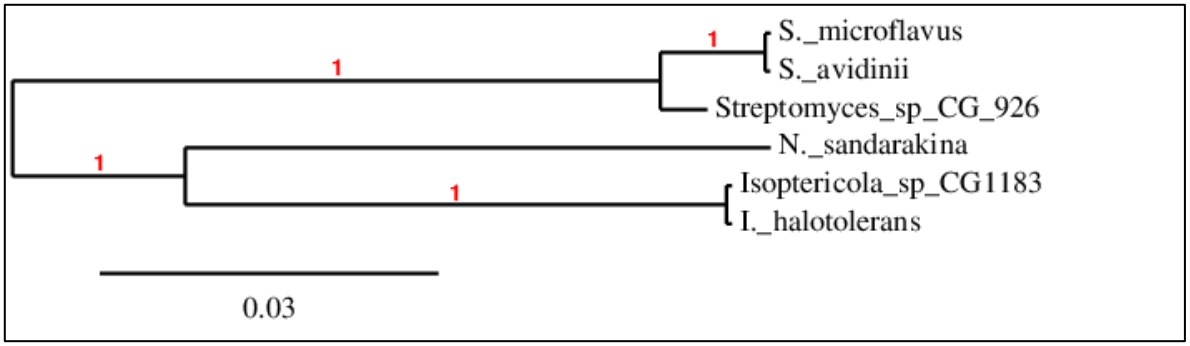

Figure 2. Phylogenetic tree of the six strains from MrBayes. The length of the bar at the bottom provides a scale for the amount of genetic change measure in nucleotides substitutions per site. The red numbers on each branch are the node support values, obtained by bayesian posterior probabilities and indicating a strong evidence that the sequences on the right of the node cluster together.

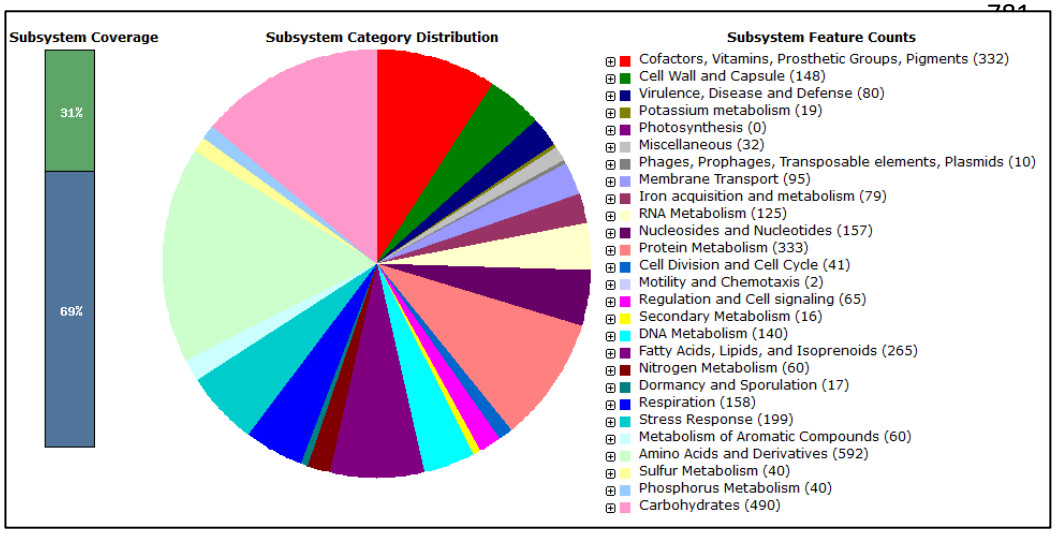

Figure 3. RAST annotation analysis for *Streptomyces* sp. Different colors indicate the subsystem and the number in brackets indicates the amount of genes found for each subsystem.

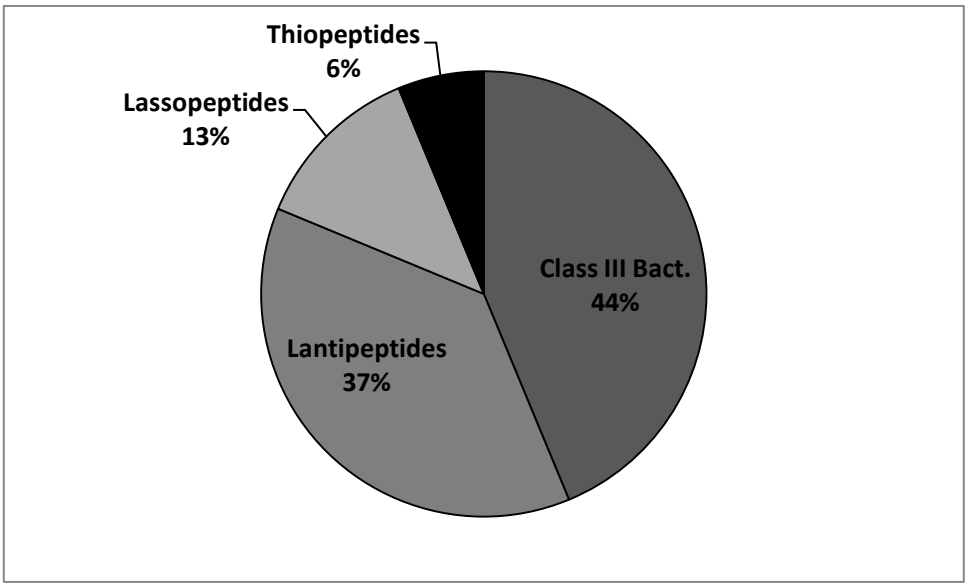

Figure 4. Bacterocins identified by Bagel3 within the 6 genomes.

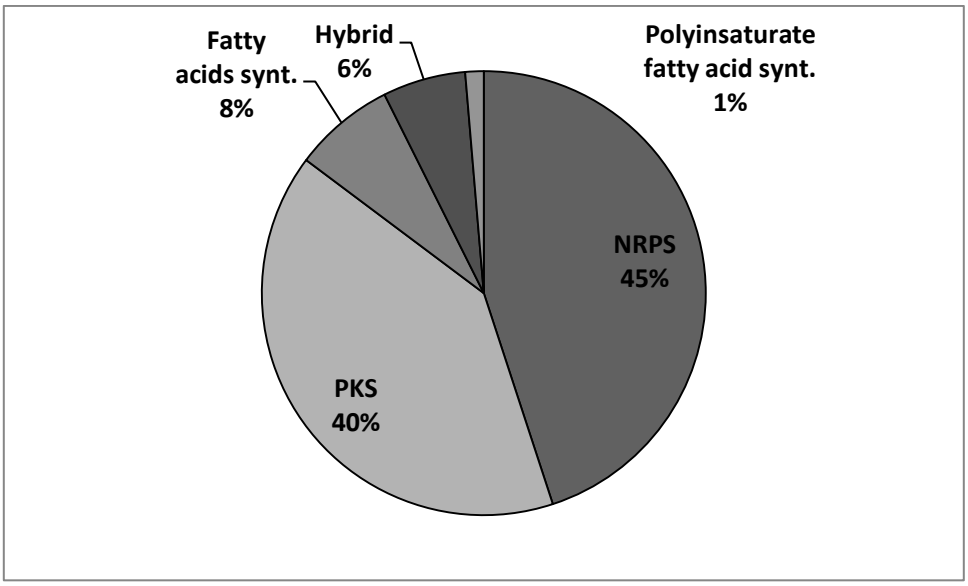

Figure 5. BGC identified with NaPDoS within the six genomes.

828 **TABLES**

829 **Table 1. *Actinobacteria* isolated from PNN and Zipaquirá salt mines samples (Cantillo *et al.,***
830 **2015).**

Isolate	Antimicrobial activity	Cytotoxic activity
Isoptericola halotolerans	NA*	+4T1 y MCF7
Isoptericola sp (CG1183)	NA	+4T1
Nestrenkonia Sandarakina	NA	+4T1 y MCF7
Streptomyces sp (CG 926)	*B. subtilis, M. smegmatis*	NA
Streptomyces avidinii	*E.coli, M. smegmatis, K. pneumonie, A. halotolerans*	NA
Streptomyces microflavus	*E.coli, B. subtilis, M. smegmatis*	NA

831 *** NA= No activity was observed. Strains were assayed for antimicrobial activity against known**
832 **bacteria and for cytotoxic activity against cell lines 4T1 and MCF7 (Cantillo *et al.,* 2015). Only**
833 **those showing activity are reported**

834

835

836

837

838

839

840

841

842

843

844

845

846

847

Table 2. K-mer lengths tested using the assembler SPAdes. In bold, the best k-mer length identified by Kmergenie.

Isolate	K-mers length (bp)
Isoperitcola halotolerans	**29**, 31, 33
Isoptericola sp (CG1183)	**29,** 31, 33
Nesterenkonia sandarakina	105, 103, **107**
Streptomyces sp (CG 926)	65, 85, **91**
Streptomyces avidinii	39, 59, **117**
Streptomyces microflavus	39, 59, **117**

Table 3. Tested chemical reactions using OptStoic

Reaction number	Carbon Source	Nitrogen Source	Chorismate ($C_{10}H_{10}O_6$)
1	Glucose-6-P	NH_3	Present
2	Glucose-6-P	NH_4	Present
3	Glucose-6-P	Serine	Present
4	Glucose-6-P	Glycine	Present
5	Glucose-6-P	Glycine/Serine	Present
6	Ribulose-5-P	NH_4	Absent
7	Ribulose-5-P	Serine	Absent
8	Ribulose-5-P	Glycine	Absent
9	Ribulose-5-P	Glycine/Serine	Absent
10	Ribose-5-P	NH_4	Absent
11	Ribose-5-P	Serine	Absent
12	Ribose-5-P	Glycine	Absent
13	Ribose-5-P	Glycine/Serine	Absent
14	Eritrose-4-P	NH_4	Absent
15	Eritrose-4-P	Serine	Absent
16	Eritrose-4-P	Glycine	Absent
17	Eritrose-4-P	Glycine/Serine	Absent
18	Glucose-6-P	NH_4	Absent
19	Glucose-6-P	Serine	Absent
20	Glucose-6-P	Glycine	Absent
21	Glucose-6-P	Glycine/Serine	Absent

Table 4. BUSCO and Quast metrics results for each of the assemblies done by Velvet, SPAdes and Velvet/Allpaths

Strain	Assembler	K-mer length	BUSCO	L-50	N-50	Contigs Number	Contigs larger than 1Kbp	Total assembly length (Mbp)*
	Velvet	29	23	23	569	338	0	0
I. halotolerans	Velvet/Allpaths	NA	40	5	280589	22	22	3.85
		29	40	37	27769	235	223	3.83
	SPAdes	31	40	29	31194	176	168	3.83
		33	40	6	214733	44	41	3.84
	Velvet	29	23	5	566	344	0	0
***Isoptericola sp* (CG1183)**	Velvet/Allpaths	NA	40	6	271167	25	25	3.86
		29	40	27	29584	234	222	3.83
	SPAdes	31	40	29	39194	175	167	3.83
		33	40	6	208419	45	42	3.84
	Velvet	29	31	2	537	289	0	0
N. sandarakina	Velvet/Allpaths	NA	40	7	180182	56	56	3.22
	SPAdes	103	40	8	172081	102	69	3.21
		105	40	8	172085	102	69	3.21
		107	40	7	179111	89	63	3.22
	Velvet	31	26	1575	1427	560	289	5
***Streptomyces sp* (CG 926)**	Velvet/Allpaths	NA	40	7	470107	46	46	8.51
		65	40	37	72375	276	225	8.44
	SPAdes	85	40	26	93082	221	172	8.46
		91	40	17	174062	123	102	8.5
	Velvet	31	40	85	28667	557	498	7.5
S. avidinii	Velvet/Allpaths	NA	39	5	630656	44	44	7.65
		39	40	27	90021	245	219	7.59
	SPAdes	59	40	17	157954	245	129	7.61
		117	40	6	404221	67	56	7.64
	Velvet	31	3	369	675	941	77	5.3
S. microflavus	Velvet/Allpaths	NA	38	5	623933	42	42	7.37
		39	40	25	92763	228	206	7.32
	SPAdes	59	40	16	162731	238	120	7.34
		117	40	6	404223	62	50	7.37

NA= Not Applicable. K-mer length used on JGI assemblies was not listed in the assemblies' reports.

*** Total assembly length refers to the accumulative length from contigs larger than 1Kbp**

892 **Table 5. RAST genome annotation results for the six genomes**

Strain	Number of genes
I. halotolerans	3483
Isoptericola sp (CG1183)	3483
N. sandarakina	2894
Streptomyces sp (CG 926)	7701
S. avidinii	6869
S. microflavus	6639

893

894

895

896

897

898

899

900

901

902

903

904

Table 6. BGCs found with each tools in the six genomes. *Streptomyces sp* (CG 926) strain harbors the greatest biosynthetic potential and antiSMASH was the best software for genome mining.

Strain	antiSMASH	NaPDoS	Bagel3	Total (%total)
I. halotolerans	38	2	1	40 (8%)
Isoptericola sp (CG1183)	37	2	1	39 (8%)
N. sandarakina	35	0	0	35 (8%)
Streptomyces sp (CG 926)	93	70	6	169 (8%)
S. avidinii	72	41	4	117 (8%)
S.microflavus	67	30	4	102 (8%)
	342	145	16	502 (8%)

919 **Table 7. Feasible OptStoic stoichiometry results**

Reaction Number	Stoichiometry	ΔG
1	73,067 $C_6H_{12}O_6$ + 150 H^+ + 150 NH_3 + 0,2196 O_2 + 10 $C_{10}H_{10}O_6$ → 12,067 $C_{39}H_{42}N_6O_{18}$ + 67,796 CO_2 + 150 H_2O	-50
2	73,067 $C_6H_{12}O_6$ + 150 H^+ + 150 NH_3 + 0,2196 O_2 + 10 $C_{10}H_{10}O_6$ → 12,067 $C_{39}H_{42}N_6O_{18}$ + 67,796 CO_2 + 150 H_2O	-50
3	145,897 $C_6H_{12}O_6$ + 20 H^+ + 150 $C_3H_7NO_3$ + 10 $C_{10}H_{10}O_6$ → 11,923 $C_{39}H_{42}N_6O_{18}$ + 60,385 CO_2 + 150 H_2O	-276.812
4	120,897 $C_6H_{12}O_6$ + 20 H^+ + 150 $C_2H_5NO_2$ + 10 $C_{10}H_{10}O_6$ → 11,923 $C_{39}H_{42}N_6O_{18}$ + 60,385 CO_2 + 150 H_2O	-270,812
5	150 $C_6H_{12}O_6$ + 20 H^+ + 150 $C_2H_5NO_2$ + 46,327 $C_3H_7NO_3$ + 10 $C_{10}H_{10}O_6$ → 12,517 $C_{39}H_{42}N_6O_{18}$ + 72,857 CO_2 + 150 H_2O	-885,822

920