Deciphering genomes: Comparative genomic analysis of legume associated *Micromonospora*

Memoria presentada por: **Raúl Riesco Jarrín**

para optar al Grado de Doctor por la Universidad de Salamanca.

Directora: Martha E. Trujillo Toledo

Universidad de Salamanca

Departamento de Microbiología y Genética

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS VNiVERSiDAD D SALAMANCA 1218 ~ 2018

AUTORIZACIÓN DEL TUTOR

MARTHA E. TRUJILLO TOLEDO, CATEDRÁTICA DEL ÁREA DE MICROBIOLOGÍA DEL DEPARTAMENTO DE MICROBIOLOGÍA Y GENÉTICA DE LA UNIVERSIDAD DE SALAMANCA

CERTIFICA:

Que la memoria de la Tesis Doctoral titulada "**Deciphering genomes: Comparative genomic analysis of legume associated *Micromonospora***" presentada por Raúl Riesco Jarrín para optar al grado de Doctor por la Universidad de Salamanca, ha sido realizada bajo su dirección en el Departamento de Microbiología y Genética, y se **autoriza** su deposito y presentación.

Y para que así conste, extiendo el presente certificado.

En Salamanca a 2 de julio de 2020

TRUJILLO
TOLEDO
MARTHA ESTELA
- 70891687Y

Firmado digitalmente por TRUJILLO TOLEDO MARTHA ESTELA - 70891687Y
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-70891687Y, givenName=MARTHA ESTELA, sn=TRUJILLO TOLEDO, cn=TRUJILLO TOLEDO MARTHA ESTELA - 70891687Y
Fecha: 2020.07.02 11:57:26 +02'00'

Martha E. Trujillo Toledo

AGRADECIMIENTOS

Me resulta difícil encontrar palabras suficientes para agradecer a tantas personas. Es importante remarcar que a pesar de que es mi nombre el que figura en la portada, este trabajo no me pertenece. Pertenece a mi tutora, la Dra. Martha Trujillo, que con un gran esfuerzo personal y una interminable paciencia ha conseguido llevar este proyecto a buen puerto. Pertenece a mi familia, a mis padres y mi hermana, que me han apoyado incluso en mis días más bajos, dándome un amor incondicional en cada paso. Pertenece a mis compañeros de laboratorio, los que están y los que se fueron, porque me han mostrado en que consiste trabajar en equipo. Pertenece a mis amigos, que me han dado apoyo moral y felicidad sin esperar nada a cambio, aun estando en la lejanía.

No necesito ninguna cita celebre de adorno, y no la encontraréis más adelante. Si hay algo que he aprendido y grabado a fuego en mi mente es que no hay trabajo que te pertenezca individualmente, y mucho menos en ciencia. Son todos fruto del trabajo en equipo y como tal, nos pertenecen a todos. Es una lección que llevaré siempre conmigo, allá donde vaya, y que posee una belleza difícil de expresar con palabras.

Sois la mente, el cuerpo y el alma que sustenta esta tesis.

A todos, gracias.

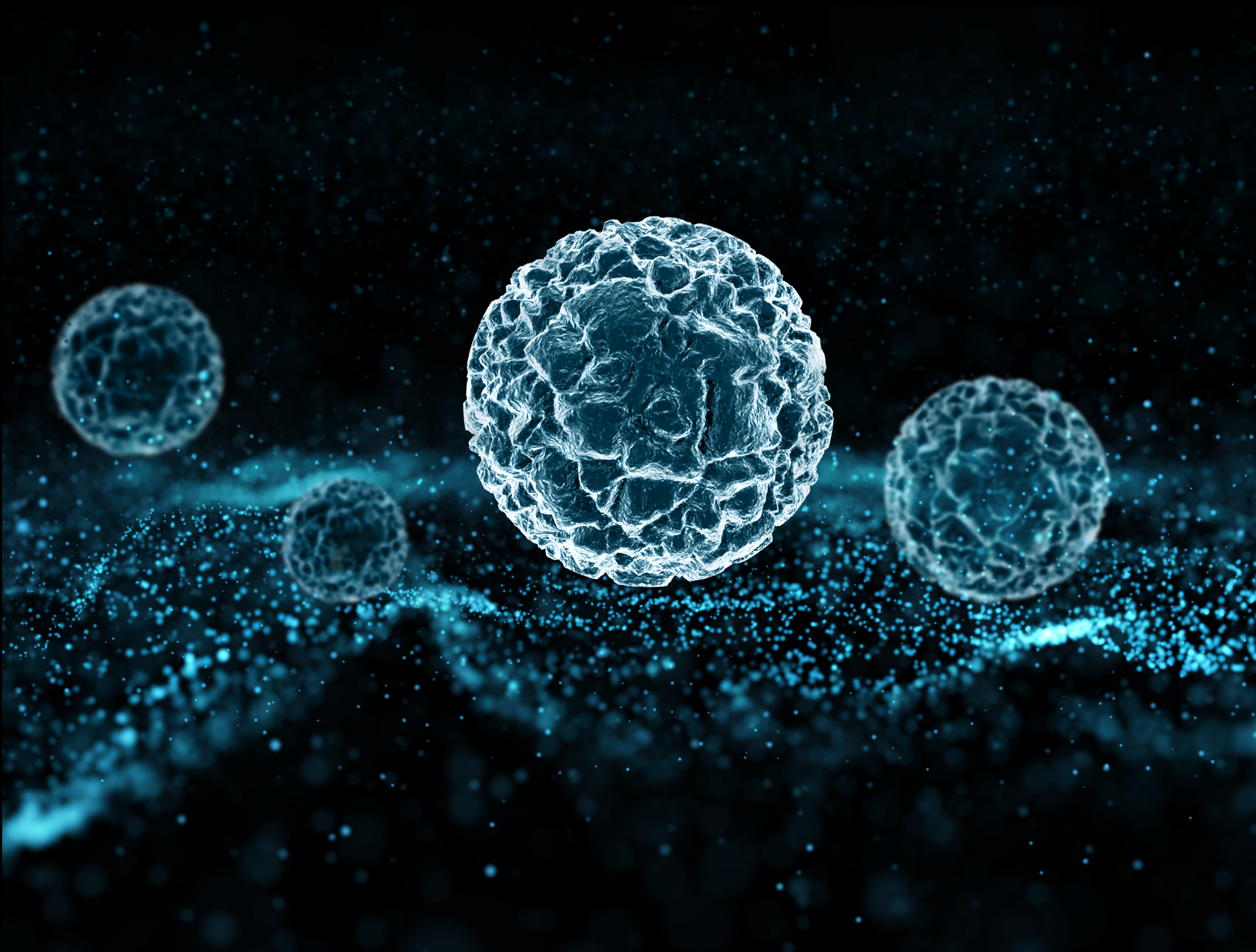
CONTENTS

From this point onwards, the reader can return to the table of contents by clicking on the house symbol situated at the bottom of each page.

LIST OF FIGURES

LIST OF TABLES

THESIS OUTLINE



THESIS OUTLINE

Micromonospora is now considered as a plant growth promoting bacterium (PGPB), but until 2015, the genomes of only two endophytic strains had been sequenced. In this thesis, a comparative genomic approach was used to elucidate some of the biological processes that drive the relationship between *Micromonospora* and the plant at a genomic level. To achieve this, the following specific objectives were proposed:

1. Creation of new bioinformatic tools that complement or enhance widely used programs in microbial genomic characterization and creation of a pipeline to characterize plant-related bacterial features in the genus *Micromonospora.*

2. Characterization of the main genomic features that drives the relationship between *Micromonospora* and the plant, using a comparative genomic approach.

3. Genome based characterization of strains isolated from different parts of the plant, closely related to the main species found in the nodules, *Micromonospora saelicesensis* and *Micromonospora noduli.*

This thesis is therefore divided in three parts, one for each objective. In the first part some of the most popular programs and applications in microbial genomic characterization were briefly analyzed and their strong and weak features were singled out. Several bioinformatic solutions were proposed, to help improve the end-user management of some of these frequently used programs. Finally, an R based pipeline was proposed to analyze and manage genomic data, to decipher significant differential features that drive the relationship between *Micromonospora* and its host plants. All these bioinformatic solutions were implemented as R based scripts, with the additional benefit that they can be easily introduced in future pipelines with only a minimal amount of commands and variables needed.

The second chapter was focused on the second objective, the characterization of the main genomic features that drive the plant-bacteria relationship in *Micromonospora*. The genomes of seventeen *Micromonospora* strains isolated from different legumes (*Cicer* sp., *Medicago* sp., *Lupinus* sp., *Ononis* sp., *Pisum* sp. and *Trifolium* sp.), and plant tissues (nodule and leaves) were sequenced. With the addition of these newly sequenced genomes, we constructed a database of 74 genomes, with an almost equal number of soil-related and endophytic-related *Micromonospora* representatives. Using a novel comparative genomic approach, based on the database generated in 2018 (Levy et al., 2018) and the proteome of known host plants, we determined several genomic features that could potentially be related to the *Micromonospora*-plant interaction.

The third chapter focused on the final objective. As the most frequently isolated *Micromonospora* in plant nodules (Carro, 2009; Carro et al., 2012; Cerda, 2008; de la Vega, 2010; Trujillo et al., 2010, 2015), the species *M. saelicesensis* and its close phylogenetic neighbor

M. noduli probably play an important role in the relationship with their legume host. However, these two species share many features, and the question arose whether they should be merged into a single species. Using different genomic approaches, in chapter III we have studied the taxonomic relationship between the species *Micromonospora saelicesensis* and *Micromonospora noduli* using genome-based data.

REFERENCES

Carro, L. (2009). Avances en la Sistemática del Género *Micromonospora*: Estudio de Cepas aisladas de la Rizosfera y Nódulos de *Pisum sativum*.

Carro, L., Spröer, C., Alonso, P., and Trujillo, M. E. (2012). Diversity of *Micromonospora* strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. *Syst. Appl. Microbiol.* 35, 73–80. doi:10.1016/j.syapm.2011.11.003.

Cerda, M. E. (2008). Aislamiento de *Micromonospora* de Nódulos de Leguminosas Tropicales y Análisis de Su Interés Como Promotor del Crecimiento Vegetal.

de la Vega, P. A. (2010). Distribución, caracterización e importancia ecologica de *Micromonospora* en nódulos fijadores de nitrogeno de *Lupinus.*

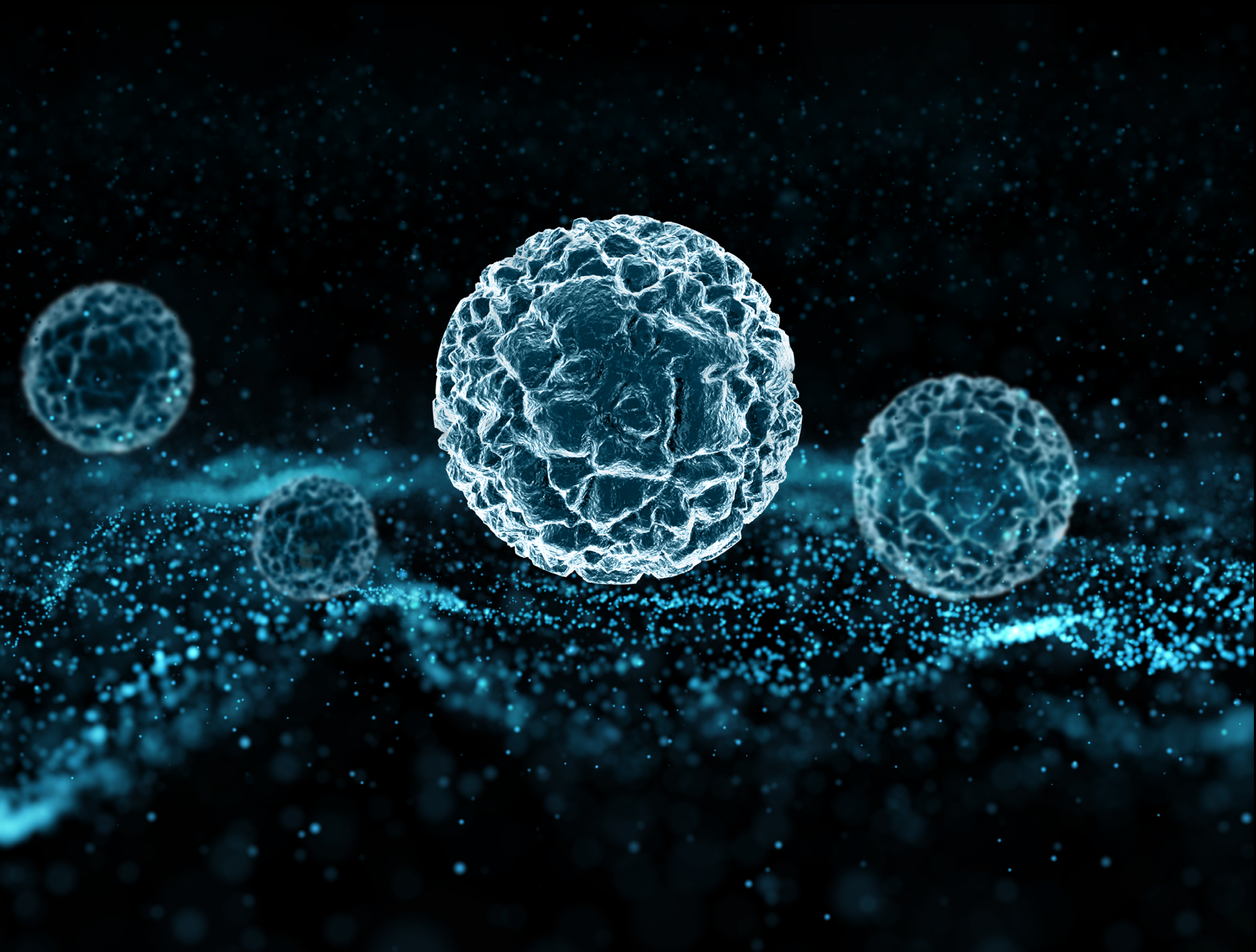
Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018). Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50, 138–150. doi:10.1038/s41588-017-0012-9.

Trujillo, M. E., Alonso-Vega, P., Rodríguez, R., Carro, L., Cerda, E., Alonso, P., et al. (2010). The genus *Micromonospora* is widespread in legume root nodules: The example of *Lupinus angustifolius. ISME J.* 4, 1265–1281. doi:10.1038/ismej.2010.55.

Trujillo, M. E., Riesco, R., Benito, P., and Carro, L. (2015). Endophytic actinobacteria and the interaction of *Micromonospora* and nitrogen fixing plants. *Front. Microbiol.* 6, 1–15. doi:10.3389/fmicb.2015.01341.

CHAPTER I

Script development and implementation



1.1 INTRODUCTION

1.1.1-Genomic information and its place in systematics

Genomics has revolutionized Life Sciences. The great potential of genomic information to generate new data cannot be underestimated, and as we speak is revolutionizing biomedical research, bacterial ecology and systematics.

A quick search in the US National Center for Biotechnology Information (NCBI) reveals a total of 285,747 bacterial genomic assemblies and 3,514 archaeal assemblies, of which 10,711 and 384 respectively are labeled as type material. Of course, this is a raw count, and does not take in count the quality of the genome, nor include all the diversity obtained in metagenomic studies. An enormous effort has been made to improve public databases with new genomic information. Global projects like the Genomic Encyclopedia of Bacteria and Archaea (GEBA) (Wu et al., 2009), with more than 15.000 bacterial and archaeal genomes already sequenced (see jgi.doe.gov for more information), and the recently proposed Global Catalogue of Microorganism 10K Type Strain Sequencing Project (Wu and Ma, 2019) are providing good quality reference data to the public repositories. It is to be expected that these databases will grow exponentially in the following years with the impulse of the previously mentioned global projects and the constant publication of genomic data in scientific works.

Advancement and drastic drop of the costs of DNA sequencing technology, the publication of new and reliable methods and tools to characterize bacterial diversity and ecology, and the development of new reliable and simpler bioinformatic pipelines, have contributed to the incorporation of genomics as part of many scientific studies. Most of the novel species described nowadays are described with genome assemblies associated with, at least, the type strain, and many journals request this information as part of their requirements for publication.

1.1.2- Genomic information in the field of bacterial systematics

Current taxonomy relies on the use of the polyphasic approach (Tindall et al., 2010), considering multiple aspects of the organism in study, including genotypic, phenotypic and chemotaxonomic features. However, the use of phenotypic traits and chemotaxonomy in prokaryotic systematics are of limited value, due to intra-species variability and lack of reproducibility (Amaral et al., 2014; Baek et al., 2018; Sutcliffe et al., 2013). As a result, genotypic characterization is growing in importance.

1.1.2.1- 16S rRNA gene phylogenies

16S rRNA gene sequence similarity and phylogeny is the most used genetic tool in bacterial systematics. However, as 16S is a very conserved gene, it presents problems in resolving phylogenetic relationships in closely related organisms (Hahnke et al., 2016; Katayama et al., 2007; Na et al., 2018).

Other housekeeping genes have been proposed to improve the resolution in closely related species. For example, the use of *gyr*B is highly recommended in the genus *Micromonospora*, given its higher resolution (Garcia et al., 2010; Kasai et al., 2000). However, these genes also present some drawbacks. Housekeeping genes, even having conserved sequences, can be subjected to Horizontal Gene Transfer events (HGT) that can hinder the interpretation of the inferred phylogenetic trees (Creevey et al., 2011). Moreover, topologies of single gene phylogenies can be very different, even if very conserved core genes are selected, making the identification of close neighbors a difficult task.

1.1.2.2- Multi-locus sequence approach

To overcome the limitation of single gene phylogenies, multi-locus sequence analysis (MLSA) approach was developed (Gevers et al., 2005). The MLSA consists in the concatenation of several housekeeping genes and the subsequent phylogenetic analysis of the concatenated sequence. The selection of genes is a crucial part in the process, as the genes ideally should be single copied in the genome as well as homologous and ubiquitous in all prokaryotes (for future comparations), or at least ubiquitous in large groups of bacteria and archaea (Gevers et al., 2005; Glaeser and Kämpfer, 2015). The final topology of the MLSA phylogenetic analysis is usually very stable, increasing its stability with the increasing number of housekeeping genes included in the analysis, being even proposed for some time as an alternative to DNA-DNA hybridization (DDH) (Cole et al., 2010; Glaeser and Kämpfer, 2015; Martens et al., 2008). The number of genes in the analysis can vary depending on the scope of the analysis, but it usually ranges from five genes to more than one hundred (Carro et al., 2012; Na et al., 2018). Before genomes were available, this approach required the partial amplification and high quality sequencing of the selected genes (Glaeser and Kämpfer, 2015), and depended greatly on the genes deposited in the databases. Furthermore, no consensus on the genes to be chosen was made between different taxa, therefore a comparative study above the genus level could be expensive and laborious, as new core-genes had to be sequenced for all the strains of interest. The advent of the genomic era has solved most of these problems. All genes of interest can be easily and automatically screened in the genome, allowing the reconstruction of trees with hundreds of genes.

Genomes can facilitate phylogenetic reconstruction using the core-genome as the base pool of genes (Liu et al., 2016b). The idea of selecting common genes for all bacteria has already been explored several times, Na and collaborators (2018) even have developed a phylogenetic tool called up-to-date bacterial core gene (UBCG), based on a database of ninety two bacterial genes, ubiquitous in the domain *Bacteria.*

1.1.2.3- Overall relatedness indexes

Genomes can also be used to establish overall general relatedness indices (OGRI), a term proposed by Chun and Rainey (2014) that englobe all indices inferred from genomes that establish a general relatedness between two genomes, in a similar fashion to the DNA-DNA hybridization. The most and frequently used OGRIs are the Average Nucleotide Identity (ANI) (Goris et al., 2007) and the digital DNA-DNA hybridization, based on the Genome BLAST Distance Phylogeny method (GBDP) (Henz et al., 2004; Meier-Kolthoff et al., 2013).

1.1.2.3.1- ANI

ANI is a similarity-based index, the query genome is divided in 1020 base pairs long sequence fragments, and then searched directly against the reference genome, usually using BLASTn program (Altschul et al., 1990; Chun and Rainey, 2014). All matches with identity above 30% are considered to calculate a mean identity value that will constitute the ANI index. As the calculation of the matches is not symmetrical (the best hit for the query-reference can be different to the reference-query), ANI values can be slightly different depending on the direction of the analysis. To avoid this, the ANI is usually calculated in both directions and a mean value is used as the ANI final value.

Prokaryotic species delimitation for the ANI index is stablished in the range of 95-96%, that is equivalent to the 70% traditional DNA-DNA hybridization value (Goris et al., 2007; Richter and Rosselló-Móra, 2009).

As the ANI is the most used OGRI, many variations of the original method have been proposed (Arahal, 2014). Since traditional BLASTn is computationally cost intensive and slow, the main variation consists in the algorithm used to search for the sequence. MUMmer software (Kurtz et al., 2004), and USEARCH (Edgar, 2010) have been proposed as alternatives that demonstrated not to compromise the accuracy in favor of the calculation speed (Richter and Rosselló-Móra, 2009; Yoon et al., 2017b). To differentiate these two approaches, ANI based on BLAST is called ANIb, ANI based on MUMmer ANIm and ANI based on USEARCH is called ANIu.

To avoid the non-symmetrical pairwise calculation problem, another approach called OrhoANI was devised (Lee et al., 2016). This approach uses orthologs, sequences with symmetrical homology (the reference and the query are the best hit in both ways), to make the pairs and calculate the index. As the resulting homology values are always symmetrical, a unique value is given as result of the ANI calculation, despite of the direction of the analysis. As expected by the use of very close sequences, OrthoANI is more restrictive and generate slightly higher values, but it is more stable than the regular ANI that have proven to give slightly different values, even when using the same platform (Beaz-Hidalgo et al., 2015).

There are many stand alone and web-based programs to calculate ANI, just to cite some examples: web-based Kostas lab ANIb calculator (Rodriguez and Konstantinidis, 2014), stand-alone and java based Jspecies with its online web based service JspeciesWS that calculates ANIm (Richter et al., 2016; Richter and Rosselló-Móra, 2009), and finally web server EzBioCloud that allows to calculate both ANIu and OrthoANI and also has a stand-alone tool called Orthologous Average Nucleotide Identity Tool (OAT) (ezbiocloud.net/tools/orthoani) that calculates both ANIb and OthoANI, implementing a matrix presentation of the results (Lee et al., 2016; Yoon et al., 2017a, 2017b).

1.1.2.3.2- Digital DNA-DNA hybridization (dDDH)

dDDH based on GBDP is a distance-based approach. Genome sequences are directly aligned without any trimming, using BLAST+ (Camacho et al., 2009) as the preferred alignment tool to obtain high scoring segment pairs or HSP. These HSP are then used to calculate a distance using a formula optimized to the overlapping of the HSP (Meier-Kolthoff et al., 2013). On the web server Genome to Genome Distance Calculator (GGDC) (ggdc.dsmz.de), three algorithms are offered as a result: "greedy", "greedy with trimming" and "coverage", but normally the second one is recommended. This method is offered as a standalone version for the old version GGDC v1.0 (Henz et al., 2004), but only as web server service for the version 2.0, offering more sophisticated and improved statistical models with interval estimations and other important features (like the calculation of pseudo-bootstrapping replicates) (Meier-Kolthoff et al., 2013).

As GBDP method is a distance-based approach, phylogeny based on distance matrices can be inferred. However, to give statistical support a pseudo-bootstrap must be inferred by generating multiple distance matrices with modifications in the sampling of HSP, a sampling with substitutions and sampling with deletion process (bootstrapping and jackknifing) to produce at least one hundred additional matrices to make the phylogenetic inference (Meier-Kolthoff et al., 2014). The GBDP method to make phylogenetic trees has been used to make genome based classifications of the phyla *Bacteroidetes,* and *Actinobacteria*, the family *Geodermatophilaceae* and the genus *Micromonospora* (Carro et al., 2018; Hahnke et al., 2016; Montero-Calasanz et al., 2017; Nouioui et al., 2018). The method is very promising however, up to this date, the program implementing this method has not been released to the public. Only a web based implementation called Type strain Genome Server (TYGS) based on the GBDP method have been recently released, but do not give total control of the strains included in the phylogenetic analysis nor the possibility to include it in a bioinformatic pipeline (Meier-Kolthoff and Göker, 2019).

1.1.3- Improvements in bioinformatic workflows to facilitate the use of bioinformatic tools in systematics

New alternatives to the already known bioinformatic tools are appearing every day. There are many algorithms, all with pros and cons. However, most of the times the worst enemy of a bioinformatic tool is not the algorithm, but the usability by the end-user. If the tool is difficult to use, involves many steps or is not freely available to customize, the final user can choose another program, even if the final result is less desirable.

As an example, given the drawbacks mentioned previously for the GBDP method (lack of a method to make distance or dDDH matrices and no availability of the tool to integrate it in a bioinformatic pipeline), the end-user may be prone to select other tools, like ANI for the OGRI and UBCG multilocus approach for the phylogenomic tree inferring. The algorithm is probably better, as it captures genome variability by aligning whole genomes instead of highly similar fragments, but the final user can choose other methods if he wants to include many genomes or wants to customize the tool.

Other disadvantage that can discourage the end-user is the use of console-based approaches. Most of the time, this cannot be avoided, because the approach may depend on Linux based tools. However, simplifying the tool to the minimum coding must be an essential priority, as the final user usually works with “click and go” tools, and normally tends to avoid the coding language as much as possible. An example of this problem can be found in the UBCG pipeline (Na et al., 2018): UBCG depends entirely on Linux based programs, and it requires at least a command line for every genome in the analysis. To the end-user of the program, it means a lot of time spent in writing cumbersome command lines. In the end, the final user can choose another method, even if it is less flexible, like TYGS server (Meier-Kolthoff and Göker, 2019).

New bioinformatics pipelines can aid to solve these flaws, even adding a better visual representation of the results, making the tool more appealing to the user.

1.1.4- Genomic applications on the field of microbe-plant interaction ecology

Molecular biology advances, not only in the field of genomics, but in all the commonly called “*omics*” (genomics, transcriptomics, proteomics, and metabolomics) have revolutionized ecology. Commonly used microbial ecological tools until now have relied on amplicon sequencing of a selection of core genes (16S or *gyr*B), allowing the characterization of the microbial community, but not the description of the features that drives the relationships with the environment. Advances in technology have made possible the reconstruction of the structure of a microbial community using the entire genomic information available, allowing the accurate screening not only of the microbial diversity, but also the pool of genes that drives the interactions between the microbial community and the environment (Levy et al., 2018a).

The study of model organisms at the genomic level and the use of comparative genomics approaches have contributed to expand the knowledge of factors that drive bacteria-plant interactions, both positive and negative. Studies of these factors have been appearing in the literature for the last decade (Liu et al., 2016a; Loper et al., 2012; Trujillo et al., 2014). These studies have focused on known bacterial genes involved in plant relationships, inferring new plant related functions at the genus and even species level. However, the molecular mechanisms that drive the plant-microbe interactions are still not fully understood. There are many gaps to fill, but the use of genomic data seems like a good alternative to complete the picture.

Genomics studies can be used to elucidate some of these unknown functions: Comparative genomics can be used for the screening of genes that potentially contribute in the bacterial adaptation to plants, amplifying the plant related bacterial gene databases even with genes with unknown functionality. Based on this idea, new, dedicated databases are now available. A good example is the Genomic Features Of Bacterial Adaptation to Plants database (GFOPAP), that uses a dataset of 3837 genomes to create a database of potential bacterial mechanisms that could be involved in plant-bacterial interactions (Levy et al., 2018b). These databases represent potential biological processes that could transform the way we see the interaction between the plant and its associated microbiome, however they must be validated in subsequent studies in the laboratory.

The drawback (and the advantage) of these functional databases is that they tend to be very general, covering multiple phyla across the domain bacteria. A bioinformatic pipeline to make genus-based functional databases, incorporating not only bacterial information but also that of the host plant, can be very useful in our understanding of the ecological function of a bacterium in the plant environment.

1.2- OBJECTIVES

1. Development of new bioinformatic pipelines that facilitate the use of popular tools in systematics.

   - Up to date Bacterial Core Gene (UBCG)

     - Automatize and improve code preparation, depending only on easily editable and importable feature tables.
     - Improve phylogenetic tree visualization.

   - Genome to Genome Distance Calculator (GGDC)

     - Improve data management and creation of an environment suitable for bioinformatic pipelines.
     - Create distance and dDDH tables that can be exported and visualized externally.
     - Create a simple phylogenomic tree reconstruction, based on distances calculated with GGDC.

2. Development of a new bioinformatic pipeline for the creation of a plant-related bacterial gene database, and the subsequent statistical analysis based on a comparative genomic approach.

   - Screening of the selected functions and genes on the genome.
   - Statistical analysis of all the genomes focusing on differences based on their habitat or any other ecological feature of interest.

1.3- SCRIPT DEVELOPMENT

1.3.1- Script environment selection

R environment offers a very flexible tool for data management and analysis. Among other characteristics R language is:

- A flexible data and storage management tool.
- A perfect tool for indexed variables management, in special tables and matrices.
- A tool for statistical analysis and visual representation of the data.
- An effective open source program language, simple, flexible and open to changes and customization, as it is constructed to run using "packages", more specific tools that are created by the community.

Taking advantage of these characteristics, many of the data of this work have been treated or inferred using personalized R scripts. These scripts offer a solution to several needs, from the simple organization of the databases used in the study, to almost final graphic representation of the results.

Three main scripts have been created for this thesis, "UBCG_iTOL_maker", "GGDC Output Management Assistant (GOMA)", and "*Micromonospora* Plant Associated Gene tool (MicroPLAGE)".

1.3.2- UBCG_iTOL_maker

Up-to-date Bacterial Core Gene (UBCG) is a tool to construct phylogenetic trees using genome sequences based on a multilocus sequence approach (Na et al., 2018). The workflow of the UBCG tool is composed of two steps. In the first one, the program does a quick protein-coding gene prediction using Prodigal (PROkaryotic DYnamic programming Gene-finding Algorithm), a gene prediction software for prokaryotes (Hyatt et al., 2010) and then screens each genome for 92 bacterial core genes using HMMER software v.3.2.1 (hmmer.org), against a predetermined Hidden Markov Model (HMM) profile database. When the screening is complete, it generates an intermediate file with all the gene data in a program readable file, in *bcg* format. In the second step, every *bcg* file in the working folder is retrieved to make an alignment using MAFFT program (Multiple sequence Alignment based on Fast Fourier Transform) (Katoh, 2002) of each of the 92 core genes and the concatenated sequence of all the genes. All these alignments are used to infer a phylogenetic tree of all the genes and the concatenated sequences using FastTree (Price et al., 2010) or RAxML (Randomized Axelerated Maximum Likelihood) (Stamatakis, 2014) , according to the user choice. Finally, it calculates the Gene Support Index, a value of support in the tree independent to the bootstrap that indicates how many individual genes support each branch in the tree (Na et al., 2018).

In the end, UBCG program uses two main files: the fasta nucleotide file of the genome, and the self-created *.bcg* file. The method of storing these files completely depends on the user, and it usually becomes quite chaotic when several trees are made and new fasta files and *.bcg* files are added to the database. Also, the user must make the terminal order for each genome, providing details like the accession number, the complete species taxonomic nomenclature, the strain identification and the path for both the fasta file and the folder in which to store the necessary *.bcg* files for each tree. All these requirements make UBCG program difficult to use in the long term, especially for the novel user that does not know how to operate in a Linux environment, leading to human error in the command prompts and duplicated work.

UBCG_iTOL_maker has been created to organize and facilitate the UBCG workflow, managing the internal fasta and *bcg* databases and generating all the necessary commands to run the UBCG program in the Linux terminal, significantly reducing the time spent in genomic phylogenetic tree reconstruction. Additionally, it integrates some tools for final tree visualization and presentation of the data, as it integrates R script table2itol.R (Göker, M., freely available in https://github.com/mgoeker/table2itol), for iTOL annotation (Letunic and Bork, 2016).

1.3.2.1- Dependencies

UBCG program is created to work in a UNIX environment, and has been tested only in Linux and Mac X 10 or higher Operating Systems (OS), not working under MS Windows OS (see user´s manual on help.ezbiocloud.net/ubcg-users-manual/). For that reason, UBCG_iTOL_maker has been optimized to work under a Linux environment, with the idea that the bioinformatic workflow can be carried out in one place and the databases can be handled *in situ.* The script has also been tested in a Windows environment, successfully working, but the resulting command file must be run in Linux since UBCG only works in a UNIX environment, and all the databases must be copied from Windows to Linux every time the user wants to run the script and run the analysis.

The script has been designed to work with only two packages, but it also inherits all the dependencies of the table2itols that runs in the final step of the workflow. In the end, the script has six dependencies:

1. Dependencies of the script:

    - ***readxl:*** This package is designed to read *.xls* and *.xlsx* Excel files and import them to the R environment (Wickham and Bryan, 2018). Feature database is designed in Excel format (*.xlsx)* to make it readable in multiple operating systems and making it easily copied and updated in the operating system the user prefers. For UBCG_iTOL_maker (and any other R script) a simpler format like comma separated values (*.csv*) or tab separated values (*.tsv*) are easier to read and import, but sometimes these files are difficult to read for the regular user (I.e.: In Spain the *csv* format is usually codified with a ";" separator instead of the regular ",", because this symbol is used as decimal separator). To avoid these problems the Excel format is favored as it can be used in almost all the regular spreadsheet reader programs, therefore allowing the user to modify all the input tables in their preferred OS.

- ***tidyr:*** As its name suggests, this package is designed for data tidying (Wickham and Henry, 2018), and it is used to handle irregularities in the feature database, created in the data integration to the R environment, like missing cells. These irregularities are often created as the feature table, an Excel *.xlsx* is often imported from one OS to another, as the user incorporates new data on their personal computer. Ideally, this process should be easy, but sometimes the incompatibilities of the OS make subtle differences in the file that incorporates written empty cells.

2. Dependencies of the script table2itol.R (github.com/mgoeker/table2itol)

- ***plotrix:*** This package aids in labeling, axis and color scaling functions (Lemon, 2013). It is used to generate branch annotations in the final phylogenetic tree.
- ***yaml:*** This package was created to convert YAML format, a standardized format used in data serialization, to R readable format (Stephens et al., 2018). It is used to define color vectors manually.
- ***readxl:*** As explained before, this package reads Microsoft Excel files.
- ***readODS:*** This package is similar to *readxl*, as it also converts spreadsheets into R readable data tables, but this programs is used to read OpenOffice ***.ods*** files (Schutten et al., 2018).
- ***optparse:*** The package is a command line parser, inspired in Python´s "optparse" library (Davis, 2019). This package is used to configure a command like orders in R for the management of certain variables of the script, that otherwise would be entered manually. In blas2itol.R this package is used for running the non-interactive mode and show the help message used to configure your command line.

The package has been constructed under R v3.5.2 (R Development Core Team and R Core Team, 2011), and developed and ran in RStudio (RStudio Team, 2016), outdated versions of R are not guaranteed to work, as the dependences could be unusable.

1.3.2.2- Preparation of the environment and databases

UBCG_iTOL_maker, like many other R scripts focused on database managing depends greatly on the organization of the environment. The script is pretended to work under the default construction of the UBCG program, that comprise the main folder, in which the UBCG main program and two additional folders are placed. These two folders are the "fasta" folder, where all the genome nucleotide fasta will be placed and the "bcg" folder, in which the *bcg* will be created.

The script uses 3 databases that will be placed in the default folders "fasta" and "bcg":

1. **Genome *fasta* database**: This database intends to contain all the *fasta* nucleotide files of the genomes that has been or will be used. It must be formatted with these characteristics:

    - All the genome records have to be fasta nucleotide files of the whole genome, or a *fasta* file with all the contigs, as provided in most of the frequently used public databases (NCBI, JGI, RAST…). To download these assemblies, the following methods are provided (valid in March 2019):

        - NCBI: Go to ncbi.nlm.nih.gov, select Genome in the search toolbar and look for the desired genome. Then go to "Download sequences in FASTA format for **genome**" and click on **genome**. Ensure that it only has one assembly (if it does not, it should appear something like "Sequence data: genome assemblies: X

          (**See Genome Assembly and Annotation report**)". Click on **See Genome Assembly and Annotation report** and select the desired strain.

          If you do not see the link to **genome**, go to the RefSeq number, click on it, scroll down to the end of the page, and click the numbers after WGS (contig accession numbers). Go to Download tab and select the FASTA link under Contigs category.

        - JGI: Go to genome.jgi.doe.gov and look for the desired genome in the search toolbar. Click the IMG number (2585427558) of the genome. In the IMG/MER page, click **Add Genome to cart** → click to select genome →Upload & Export & Save tab → Download selected genomes via JGI Portal (click **Download Genomes**). Select *.fna* file inside the downloaded folder.

        - RAST: Go to rast.nmpdr.org, click on **Go to the Jobs Overview**, select the desired genome clicking in [**view details**]. Then go to "Available downloads for this job", select DNA Contigs and click Download.

- All assembly records must be in *.fasta* format, not *.fna, .fas* or *.fa* for the construction of the bcg creation script. If they are not, a direct change of the extension can be done manually, as all above mentioned formats are equivalent to the canonical *fasta* format.
- All *fasta* names must be as simple as possible and must be constructed without spaces or symbols (with the unique exception of "_"). The name must be unique, as it will be used as key reference in the database scanning.
- All *fasta* files must be contained in a folder, named after the genus of the strain. First letter of the genus name must be in capital letter.
- All genus folders must be contained in the "fasta" folder of the UBCG program.

2. **Genome feature database**: It is an Excel (*.xlsx*) file, containing all relevant characteristics. All genome in the database must have an entry. The database will be placed in the "fasta" folder, with the Genome *fasta* database. At this point, this database contains the following fields, but only the specifically named as mandatory are really needed, the rest will be used to make the iTOL feature table:
   - **Fasta:** *fasta* file name without the extension (i.e.:M_auratinigra_DSM_44815): This field is mandatory, as it will be used as reference in the analysis.
   - **Phylum:** (i.e.: *Actinobacteria)*
   - **Class:** (i.e.: *Actinobacteria*)
   - **Order:** (i.e.: *Actinobacteria)*
   - **Family:** (i.e.: *Micromonosporaceae*)
   - **Genus:** (i.e.: *Micromonospora)*: This field is mandatory to locate the *fasta* file.
   - **Species:** (i.e.: *Micromonospora auratinigra*): This field is mandatory for *bcg* file creation.
   - **Strain:** (i.e.: DSM 44815): This field is mandatory for *bcg* file creation.
   - **type_strain:** "yes" if it is a type strain, "no" otherwise. (i.e.: yes): This field is mandatory for *bcg* file creation.
   - **genes:** number of genes (i.e.: 6183)
   - **genome_length:** total length of the genome, in pair of bases (i.e.: 6758600)
   - **GC:** G+C content percentage without % symbol (i.e.: 70.25)
   - **Habitat:** origin of the strain (i.e.: soil)
   - **Ref:** accession number (i.e.: NZ_LT594323, jgi_2585427558, rast_66654556). This field is mandatory for *bcg* file creation.
   - **Contig:** number of contigs (i.e.: 2)

3. **bcg database:** The database is handled entirely by UBCG_iTOL_maker. It looks for the default "bcg" folder of the UBCG program, and it locates a subfolder named "bcg_general_db", where the general database of *bcg* is placed. In the case that one or none of the folders exist, the script will create them. All the newly *.bcg* files created by UBCG will be placed automatically in the bcg_general_db subfolder. This subfolder will be scanned by UBCG_iTOL_maker to extract all the already done *.bcg* files for the next analysis of the UBCG program, preventing duplicated work and saving computational time.

For the script to work, the genome *fasta* and the feature databases must be well formatted and in the correct folder. UBCG_iTOL_maker R script file can be placed in any folder but must be accompanied with the table2itol.R file (available in github.com/mgoeker/table2itol). Also, it must be in the same location as the search table, a simple Excel file with one column named "fasta" in where all fasta names used for the analysis are included (without the extension, they can be copy-pasted from the feature database).

1.3.2.3- Proposed bioinformatic workflow

The proposed bioinformatic workflow is divided in 3 steps:

1. **UBCG_iTOL_maker.** The script must be run separately in R console command or RStudio. Before running the script, the user must stablish the variables between quotes:

   - Name of the analysis: will be used as name of the tree (i.e.: "tree_Microm"). It must be unique, if there is a previous work with the same name, the script will display an error message and the program will end.
   - The UBCG main program path (i.e.: "/home/user/desktop/UBCG_v3/UBCG").
   - Table2itol and UBCG_iTOL_maker containing folder path (i.e.: "/home/user/desktop/UBCG_v3/UBCG/UBCG_itol_maker").
   - Feature database name: As the "fasta" folder has been set as the predetermined folder for the database, only the name (with the extension) must be provided (i.e.: "genome_feat_DB.xlsx").
   - Search table name: The name of the search table, that must be in the same folder as the UBCG_iTOL_maker ("search_fasta_table.xlsx").

When all the variables are set, the user can run the script. The script will scan the environment and create all the necessary folders for the analysis. If the folder for the phylogenetic tree already exists, due to a repetition of the work, the script will generate an error message and will terminate. On a second step, it will load all the variables from the feature database, using the search table as a reference. These variables will be used for four processes:

- Scan for the *.bcg* files in the general database, and it exists, it will copy it to the analysis folder. If the *.bcg* file does not exists, it will create a command prompt in a text file for the generation of the file in the UBCG program and its posterior copy to the analysis folder.
- Generation of a command prompt to do the phylogenetic tree in the UBCG program.
- Construction of a readable table for the table2itol script. After that, the script will launch table2itol using this table as model to make the final iTOL feature files.
- Construction of a genomic accession table: with the complete strain identification and the accession number in the public database, in *.csv* format.

2. **UBCG**: Using the script with all the command prompts generated previously by UBCG_iTOL_maker, UBCG program can be launched in the GNU to generate all the necessary *.bcg* files in the “general_bcg_db” folder and copy them to the bcg tree folder where the analysis will be performed. After that, UBCG will make the alignments and then will infer the phylogenetic trees for each gene and the concatenated sequences. This process will generate alignment and raw tree files for each gene and the concatenated sequence.
3. **iTOL visualization:** Raw tree files and all the desired features can be loaded in the iTOL visualization tool (Letunic and Bork, 2016), freely available in itol.embl.de. From this platform the final details can be arranged, and the final phylogenetic tree can be exported to different formats.

A summary of the proposed bioinformatic workflow can be found in Figure 1.

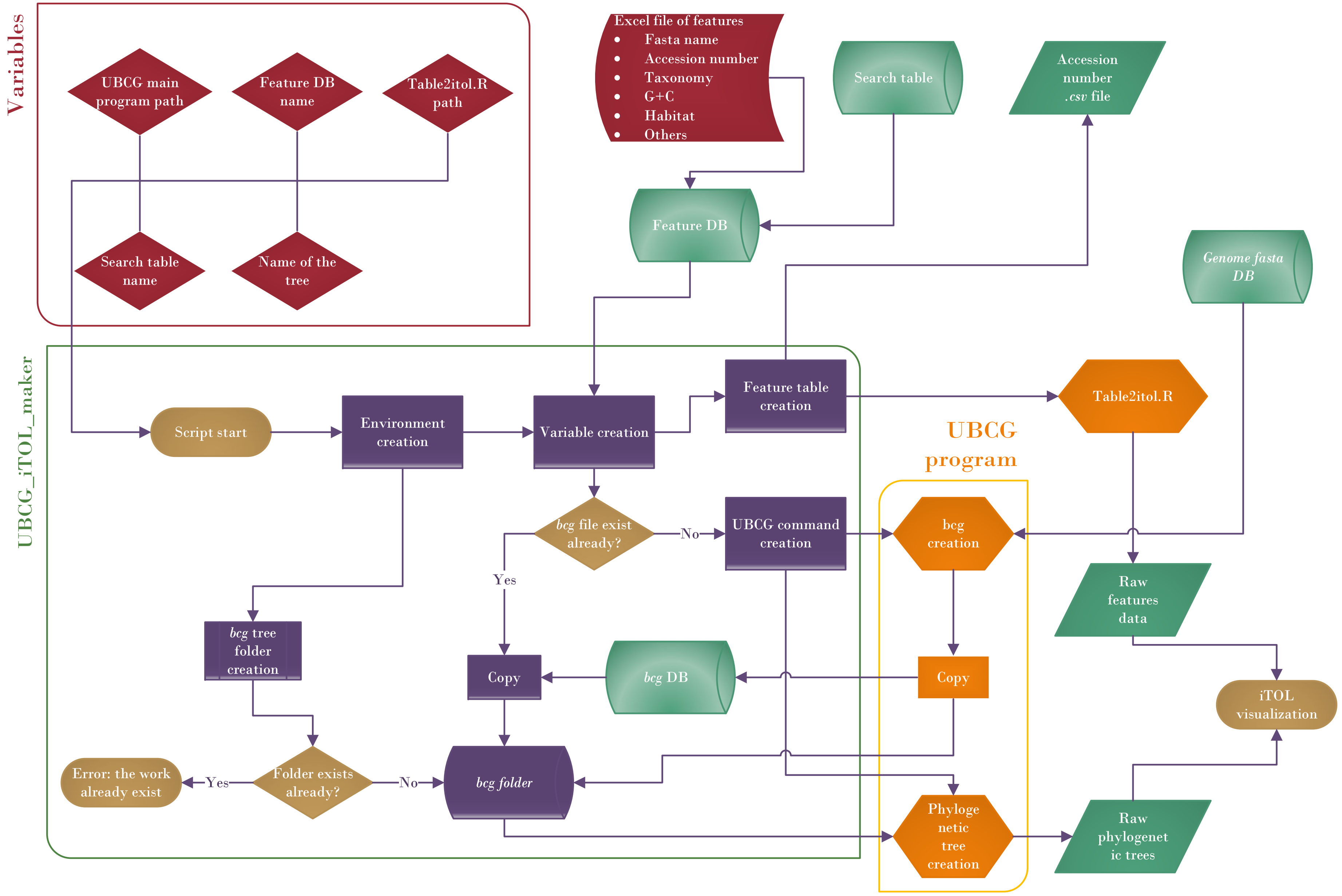


Figure 1: Proposed bioinformatic workflow of UBCG_iTOL_maker. Diamonds represents choices, cylinders represent databases, squares represent process of the pipeline, rhomboids represent output data. In purple parts of the UBCG_iTOL_maker, in orange parts of external program, including UBCG main program and table2itol, in red variables and in green databases and output data.

1.3.3- GGDC Output Management Assistant (GOMA)

Genome-to-Genome Distance Calculator (GGDC) is an online based tool that makes a distance calculation between a pair of genomes and estimates a digital hybridization percentage, mimicking the conventional DNA-DNA hybridization (Meier-Kolthoff et al., 2013). The workflow of the UBCG method is composed of three main steps, that are done automatically in a dedicated server: In the first step the pair of genomes is reciprocally blasted against each other, looking for "High-scoring segment pairs" (HSP), namely zones (or matches) of high similarity between the two genomes. In the second phase, these matches are converted to a single distance value, applying one of three formulas:

1. Length of the HSP/genome length.
2. Sum of all identities in HSP/overall HSP length
3. Sum of all identities in HSP/genome length

The recommended formula for almost all cases is formula 2, as it does not consider the genome length, being able to make an accurate analysis on incomplete assemblies. In the last step, GGDC uses a generalized linear model (GLM) inferred from empirical data (Meier-Kolthoff et al., 2013) to calculate the digital DNA-DNA hybridization (dDDH).

From an end-user point of view, GGDC server takes one genome sequence and compares it against up to 75 genomes, making pairs between the query sequence and each one of the references. As final product, the server sends the user an automated email with an attached comma separated value spreadsheet (*.csv*) with all the data, including calculated distances for each of the formulas, inferred dDDH and G+C difference. This final display is quite useful if the user only wants the comparison between one genome and the rest, but quite cumbersome when the work involves doing a dDDH or a distance matrix, that will require multiple files in a not so easy to read form for many countries (like Spain), with automatic numerical names. When the work surpasses the 75 genomes, the complexity rises as all the interactions must be done in phases. The analysis can take several days (putting the server in a reasonable 50-60% of workload) yielding results of hundreds of files (Figure 2).

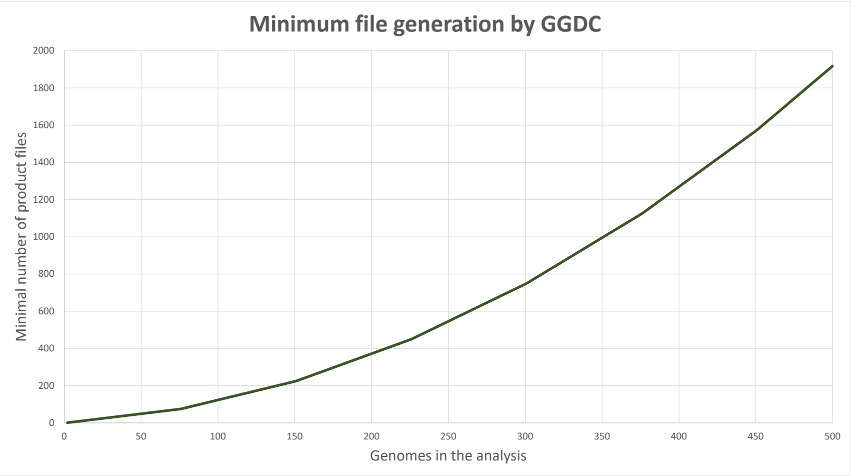


Figure 2: Minimum number of files generated by GGDC needed to make a matrix, assuming a maximum number of 75 references against one query.

GGDC Output Management Assistant (GOMA) offers a simple solution to treat the GGDC data, facilitating the creation of matrices from these automatically generated files, providing a more comprehensible labeling, graphical representation and means to make a phylogenetic reconstruction from the distance data using FastME 2.0 online service (Lefort et al., 2015).

1.3.3.1- Dependencies

GOMA uses as input the output of the GGDC online tool, canonical comma separated value files that can be used under any OS. As GGDC is created specifically to work web based, no restrictions in the environment are imposed by the original program. GOMA has been created under MS Windows 10 OS, with one package that involves the utilization of a java environment (*xlsx* package), presenting some issues in a Linux based environment and implying that the program can only be fully used in a Linux environment with root administration privileges. An UNIX adapted GOMA version has been consequently been created, with the final export in Microsoft Excel format (*.xlsx*) stripped from the script, only exporting dDDH and distance matrices in canonical *.csv* format.

The script has been designed to work with three packages:

1. ***data.table:*** This package is designed to operate with large sets of tabular data, allowing reorganization of column data with simple syntax (Dowle and Srinivasan, 2018). It is used in GOMA to read the original data and create the final matrices using the identifier of the query (the *fasta* name of the query genome) and the references (the *fasta* name of the references genome).

2. ***ComplexHeatmap***: This package is an efficient way to visualize association between data from different sources and reveal potential connections and relationships in a visual way, providing a flexible way to arrange heatmaps and annotations (Gu et al., 2016). It is used in GOMA script to create the final heatmaps of distances and dDDH, providing also a clustering function that can be used to create a reference dendrogram of the genomes in the analysis.

3. ***xlsx:*** This package provides a way to read, write and format Microsoft Excel files (*.xls* and *.xlsx*) in an R environment (Dragulescu and Arendt, 2018). It is used in the script to export the final dDDH and distance matrix to an Excel format. This package is not strictly necessaire for the core functions of the scripts, as it is only used to generate an easily format to work in a Microsoft environment. This package sometimes gives privileges problems in UNIX based environments, therefore it has not been included in the Linux version of the script.

The package has been constructed under R v3.5.2 (R Development Core Team and R Core Team, 2011), and developed to run in RStudio (RStudio Team, 2016), outdated versions of R are not guaranteed to work, as the dependencies could be unusable.

1.3.3.2- Preparation of the environment and databases

GOMA does not use a strict environment to work, as it mostly creates its own working environment, only being mandatory that all GGDC output files are contained in the same folder. However, there are some prerequisites in the data uploaded to the GGDC server that must be taken in consideration for the script to work. The script captures the names of the *fasta* uploaded to the server from the output file and uses them as key references to construct the matrices. For that reason, the *fasta* names must be unique and must contain the minimal number of symbols, as they can be mistaken as operators in R environment and lead to errors in the matrix construction. If there is a necessity to use a symbol, for example to separate two names, the underscore ("_") can be used with confidence. No name changing is needed for the output files, they will be recognized in the folder for the name pattern and introduced in the pipeline automatically.

1.3.3.3- Proposed bioinformatic workflow

The proposed workflow is divided in three main steps:

1. **GGDC:** The user must upload the *fasta* files to the server, taking in consideration the requirements presented previously. This step can be quite cumbersome, as all the data must be uploaded manually, and all the outputs must be downloaded from the user mail. As of today, there is no alternative because the code for the GGDC v2.1 program has not been made public (only depreciated v1.0 is available), nor an API (Application Programming Interface) has been developed to upload the data to the server. In addition, the minimal number of uploads of datasets is equal to minimal number of product files that can be seen in Figure 2; therefore, if the user wants to include many genomes in the analysis a lot of time will be required.

2. **GOMA:** The script must be run in a R console or RStudio. Before running the script, the user must stablish a unique variable:

   - Working directory path: It is the path of the folder that contains all the GGDC generated files.

   When the variable is set the script can be launched. First, the script will look for all the GGDC files available in the folder, looking for the pattern "ggdc_" that has all the GGDC files. The script will scan the file for the name of the query genome and rename the GGDC output file with it. If there is more than one file for the same query, the script will add a numerical tag to distinguish them (I.e.: M_sae.csv, M_sae_1.csv, M_sae_2.csv, etc.). After that, the script will create a list of all genomes included in the analysis and use it to load each of the *.csv* files and construct both distance and dDDH matrix with the data corresponding to formula

2. The script is designed to accept new data, therefore new GGDC data can be provided and the script reloaded to construct the matrix using old and new data even if the names has been automatically changed in previous analysis.

After the matrices are completed, a heatmap will be created for each one of the matrices. GOMA will create a folder called "output" were all the product files will be allocated. Heatmaps will be created as *.pdf* files inside the "output" folder and matrices will be exported as *.xlsx* Microsoft Excel files and *.csv* comma separated values.

Finally, and only if all the interactions have been made and the distance matrix is complete, the script will generate a *phylip* formatted distance matrix called "FastME_matrix.txt" that can be used to infer phylogenomic tree reconstruction. If any interaction is missing, the script will generate an error message in *.txt* format that indicates which coordinates of the matrix are empty.

3. **FastME tree reconstruction:** Phylip formatted distance matrix can be uploaded to FastME v2.0 online tool (Lefort et al., 2015), freely available in atgc-montpellier.fr/fastme/, and used to infer a genomic phylogeny. However, no statistical support will be generated as only one distance matrix is uploaded. The resulting tree will be generated in newick standard format (*.nwk*) that can be uploaded to most of the commonly used visualization programs (like MEGA or iTOL (Kumar et al., 2016; Letunic and Bork, 2016)).

A summary of the proposed bioinformatic workflow can be found in Figure 3.

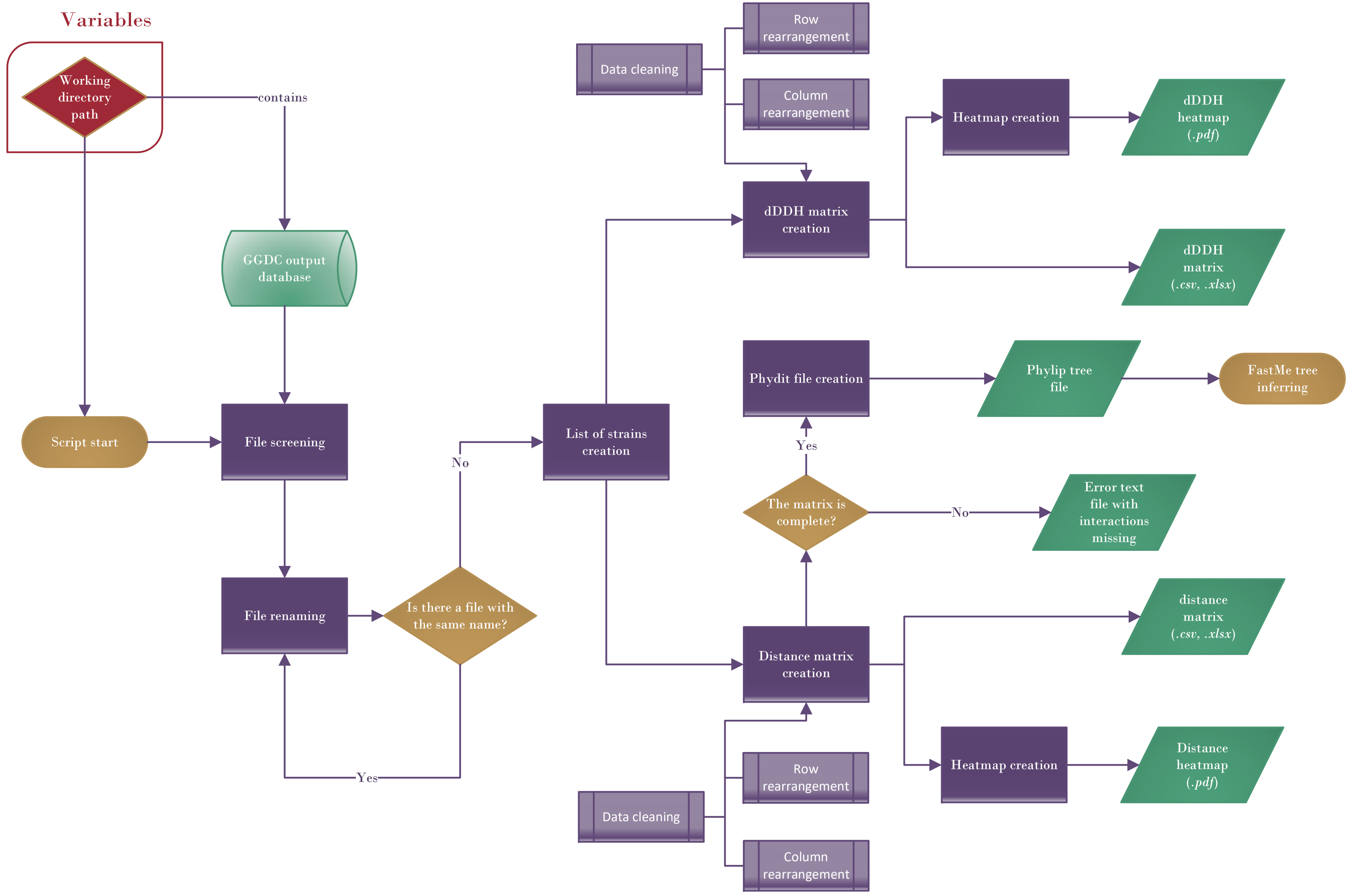


Figure 3: Proposed bioinformatic workflow of GOMA. Diamonds represents choices, cylinders represent databases, squares represent process of the pipeline, rhomboids represent output data. In purple parts of the GOMA, in red variables and in green databases and output data.

1.3.4- *Micromonospora* Plant Associated Gene tool (MicroPLAGE)

In 2018, Levy and collaborators published a huge database of potentially plant-related genes, inferred from several statistical abundance analyses at a genomic level (Levy et al., 2018b). The starting database comprised 3837 genomes, distributed across many phyla in the *Bacteria* domain, including the phylum *Actinobacteria*. Five statistical approaches were used in the analysis, and all the products, that include COG, Pfam, TIGRFAM and KEGG annotations, and hmm profiles of the plant-related genes, were included in a huge and complex database, comprised of Excel, Fasta, HMM and phylogenetic trees data, and most of them were divided according to its phylogenetic distribution.

As it can be imagined, extracting data from this database and inferring results for it can be a challenge by itself. For this work, the database of Levy and colleagues was manually curated, only accepting entries that were covered by two or more of the five statistical analyses, as recommended (Levy et al., 2018b). Considering that the analysis would be focused in the genus *Micromonospora*, only the entries contained in the *Actinobacteria*1 database were used. After performing *de novo* annotations of the genomes (see Materials and Methods in chapter II), all the raw data must be analyzed. Working with a medium to large size databases can be troublesome, as each annotation of interest must be manually screened in each genome and incorporated in the comparative analysis. As reference, in this work a medium size database of seventy-four genomes was used, comprising approximately half million genes. *Micromonospora* Plant Associated Genes tool (MicroPLAGE) was made as a tool to channel all this information, concatenating a query for all the annotations of interest and a set of statistical analyses to infer functional relationships among the genomes in the database. The script finds the potentially plant related bacterial genes, and cluster all strains according to its plant-microbe interaction related functions, in the assumption that there must be functional differences between strains of *Micromonospora* isolated from different niches. The main product of the script is a list with the statistically over or infra-represented functions that characterized each function-based cluster, allowing a characterization of the main functions that potentially drives the plant-microbe relationship in the genus *Micromonospora.*

1.3.4.1- Dependencies

MicroPLAGE is a complex script, mostly dependent on three types of packages based on its functionality: table formatting and data interchange, statistical analysis of table formatted data and graphical representation.

1. **Table formatting**

Table formatting packages *tydir* and *data.table*, already explained before, were used to format input data to find all plant-related annotations in the genome features. The packages will index and cross all the information, to finally generate a new database with all potentially plant-microbe interaction related genes. These two packages constitute the core of the script, since their final output is the central table containing all potentially plant related bacterial genes in each genome, functionally characterized or not, that will be confronted afterwards with the functional annotation tables for the statistical analysis.

2. **Statistical analysis**

   - ***FactoMineR:*** It is a package designed for multivariate Exploratory Data Analysis (Lê et al., 2008). Its main features are the principal component analysis, correspondence analysis, multiple correspondence analysis, clustering and structuration of data. It is used to perform the Principal Component Analysis (PCA) made in the COGs of the input genomes and the COGs of the potentially plant related bacterial genes. It is also responsible of the clustering of the KEGG functional data to obtain all the functional related groups and the statistically significant functions that drives this distribution.

   - ***FactoInvestigate:*** the package is designed to interpret in an autonomous form all the data generated in the PCA analysis, selecting the best graphs to represent and explain the results. It also generates an text file with a report for all the main features of the analysis (Thuleau and Husson, 2018). The script uses this package in order to generate report files of the PCA produced by *FactoMineR in the analysis.*

   - ***cluster:*** This package was created to perform cluster analysis in large sets of data. It finds groups inside complex data tables (Maechler et al., 2018). It is used in the analysis to form clusters in the PCA for its posterior graphical representation, interacting with the package *circlize* to make the final *.pdf* output of the PCA analysis. It was also used to validate the number of clusters generated in the KEGG functional analysis performed by *FactoMiner.*

   - ***ComplexHeatmap:*** This package, previously explained in the GOMA script is used in MicroPLAGE to generate multiple heatmaps, organizing and clustering the data to perform an ideal visual representation. It is used in the representation of the core genome, as well as the functional KEGG representation.

3. **Visual representation**

   - ***factoextra:*** It is used to extract and visualize data from the Multivariate Data Analysis made by *FactoMineR* (Kassambara and Mundt, 2017). In the script, it is used to internally visualize all the PCA data and clustering analysis.

   - ***circlize:*** It is a package for product circular layouts representation of huge amounts of information (Gu et al., 2014). It is used in MicroPLAGE to produce the final *pdf* outputs of the PCA analysis, in conjunction with the package *cluster.*

- ***ggplot2:*** *ggplot2* is a useful tool to generate graphics (Wickham, 2016), being one of the most popular tools in data visualization among R users. It is mainly used in the script to make the bar plots for the COGs of each strain.

- ***ggfortify:*** It works as a complement of *ggplot2*, providing a unified style for complex analysis like time series, PCA and clustering (Tang et al., 2016). It is used along with *ggplot2* for several visual representation, including PCA visualization generated by *cluster* and *circlize.*

- ***RColorBrewer:*** It is used to add color palettes in R visualization tools (Neuwirth, 2014). In the script, this package is used to add color to some graphics, like heatmaps and bar plots.

1.3.4.2- Preparation of the environment and databases

MicroPLAGE is not designed to create, manage or curate its own input database therefore the preparation of the initial data and the working environment is a critic step for the program to work. All the input data must be contained in a folder named input inside the main working directory. The script needs many input data, divided in:

1. **Reference list,** of all genomes included in the analysis in a *csv2* format (a special comma separated values file that changes the comma separator for a semicolon separator (“,” → “;”) and the point decimal indicator for a comma (“.” → “,”)). This file contains all the metadata of the genomes in the study, like the strain identifier (that must be unique), the accession code, the number of coding sequences (CDS), the habitat and the genome length. This file must be called “CDS.csv”. The reference list must have at least the following features:

    - **strain**: This is the reference name and is not necessarily the official strain designation. As it will be used as reference, the field “strain” must be unique for each genome in order to link this table with all the annotation data. No symbols that can be confused as operators can be used in the reference designation, and all spaces must be replaced with underscores.

    - **CDS**: number of coding sequences in the genome. They are used to calculate relative values, like the percentage of COGs in the genome.

    - **Ref:** Accession number of the genome. It is not used in the analysis, but it is the official reference to find the assembly in public databases. Note: It was not used as key reference because some of the genomes were not published at the time the analysis was carried out.

    - **RealStrain_id:** The strain designation.

    - **Original_procedence:** original habitat, directly taken from the description of the strain.

- **Procedence_final:** The habitat designation must be normalized and simplified, due to the complex nature of the data.
- **genome_length:** Genome length, in Mbp.

2. **Strain annotation:** All strains must be annotated with Pfam, TIGRFAM, the HMM database of Orthologs contained in Levy et al. 2018 and the Plant-resembling bacterial genes found in the blast analysis against the plant proteome database. All these data must be contained in a folder with the reference name of the genome, included in "strain" field of the reference list. In this work, all annotations are structured in *csv2* spreadsheet format.

- **Pfam annotation:** It must be contained in a *csv2* file, inside the strain folder, with the name "*dom_strain.csv*" (replacing *strain* with the reference name). It needs to contain at least a column named "domain" with the Pfam annotation and a column name "gene" with the gene identifier.
- **TIGRFAM annotation:** In a similar way to the previous annotation table, it must be contained in the strain folder, with the name "TIGRfam_*strain*.csv" with a *csv2* format. It needs at least a column named "TIGR" with the TIGRFAM annotation and a column named "gene" with the gene identifier.
- **HMM database of orthologs:** All hmm profiles contained in the database *Actinobacteria*1 of the Genomic Features of Bacterial Adaptation to Plants database (freely available in labs.bio.unc.edu/Dangl/Resources/gfobap_website/ (Levy et al., 2018b)) have been screened against the genomes, generating an additional annotation. The annotation file must be named "table_res_*strain.*csv" and be in a *csv2* format inside the strain folder. It must have at least a column named "HMM" with the hmm profile name and a gene column with the gen identifier.
- **Plant-resembling bacterial genes.** All genes have been confronted against a known *Micromonospora* plant host proteome database. The result must be contained in a *csv2* format file named "BLAST_plant_proteome_*strain*.csv", within the strain folder, with at least one column named "gene", with all the genes found in the analysis.
- **Transcriptome comparison.** MicroPLAGE is able to perform a comparison with the results of the transcriptome, if needed. If this function is required, a folder within the strain folder named "transcriptoma" and a simple file with the gene identifier (named "gene") and the result (UP/DOWN), must be included. The file must be named "*strain*_res_Patri_trad.csv".

3. **EggNOG annotation**. All the genomes have been screened in the EggNOG mapper tool (Huerta-Cepas et al., 2017), and the results converted into a *csv2* file. All these files have been named "*strain*_NOG.csv" and situated in a folder named EGGNOG within the input folder. This file will be used to extract all COGs and KEGG information.

4. **Core genome.** Roary (Page et al., 2015) has been used to calculate the core and pan-genome of the set of genomes, producing a final table, named "gene_presence_absence.csv". As this analysis will be mentioned in Materials and Methods in Chapter II, and will be discussed later, no more information about the procedure and the results will be given in this section. The file "gene_presence_absence.csv" generated by Roary will be used to extract all genes contained in the core genome for each strain. This file must be contained in a folder called "core_genome" within the input folder.

   As Roary is very selective with the input format, sometimes a new annotation with prokka (Seemann, 2014) is needed, giving as result other gene identifiers. If a translation is needed, *csv2* files with translation for each strain must be contained in the path "./input/core_genome/core_translation/results" and named "*strain*_trad_core.csv".

At the beginning of the analysis, MicroPLAGE will look for the files needed, and if they are not present the script will generate an error message and terminate.

1.3.4.3- Proposed bioinformatic workflow

The proposed analysis workflow involves three main steps: genome analysis and data preparation, potential plant-related gene table generation and statistical analysis and graphic generation, being responsible MicroPLAGE of the last two parts.

1. **Genome analysis and data preparation.** This analysis will be commented in Materials and Methods (Chapter II). This step involves:

   - Use of HMMER (hmmer.org) to screen the genomes against Pfam, TIGRFAM and the HMM database of orthologs in the database *Actinobacteria*1 of the Genomic Features of Bacterial Adaptation to Plants database (Levy et al., 2018b).
   - Use of BLASTp, available within BLAST+ (Camacho et al., 2009) to compare our genomes against the plant proteome database, generating a set of proteins with high resemblance to the plant host proteome.
   - Use of EggNOG mapper (Huerta-Cepas et al., 2017) to analyze each genome against the EggNOG database (Huerta-Cepas et al., 2016).
   - Use of Roary (Page et al., 2015) to calculate the core and pan-genome, generating a table of absence-presence of the genes of each genome in the pan-genome.

2. **Potential plant-related bacterial gene table generation.** This is the core of MicroPLAGE. In the first place, EggNOG files are screened for KEGG and COGs annotations, and all the files for these annotations are created within each strain folder inside the input folder. It also looks for the presence-absence table, creating a file within each strain folder, containing all genes in the core genome.

   The script screens all the annotation files, looking for the plant-microbe related features described in Levy et al. 2018 *Actinobacteria*1 database, focusing only in features confirmed with two or more statistical approaches. After that, all genes that has been found as highly similar to the plant proteome are added to the analysis. Finally, any core gene found in the previous queries is labeled as not-differential and eliminated from the analysis.

   The final product is a list of the strains with all the potential plant-related genes, that contains both functionally well characterized genes and uncharacterized genes. Only the genes with functional characterization will be used in the functional statistical analysis, but the inclusion of poorly characterized genes is very important, as they could represent hidden potential, and start new lines of work in the future. All genes are exported as a text file, with the gene and KEGG functional annotation, ready to be mapped in KEGG mapper tool (Kanehisa et al., 2012).

3. **Statistical analysis and graphic generation:** All genes with functional COG and KEGG annotation will be analyzed to stablish if some function is statistically differential in plant-related strains. First, the script will make an initial analysis with the COGs of the potentially plant related bacterial genes. All these COGs will be plotted in bar-plots for each strain and the overall data will be included in a PCA analysis. The analysis of this PCA will be the first main output of the program.

   On second place, all the KEGG annotations will be analyzed. First, all the data for all strains will be plotted in a heatmap, to visually see all the differences. After that, considering that unique functions could affect the position of some strains in smaller clusters and will not represent an overall differential plant-related bacterial feature of the genus, all unique genes are separated and analyzed separately (in this thesis, this step only affected one strain with a very differential feature, comprising 75 genes of the strain *M. pattaloongensis* DSM 45245^{T}). All the curated data is then clustered, to see the distribution of the strains and the features that drive this distribution (all the features that significantly deviates from the mean). All the features that characterized each cluster will be written in a *csv2* formatted table of differential elements. All these elements can be used to create a map of differential elements, to visually study them in the KEGG mapper tool.

In addition to the previously mentioned analysis, MicroPLAGE works with the input data to make some initial reference analysis. It will make a PCA analysis of the COGs of all strains with the raw data, to see how the strains distribute themselves before the analysis. It will also make a heatmap with the Roary data, to visualize the pangenome. Additionally, it will compare the plan-related gene table obtained in step 2 to the transcriptomic data, if needed.

Most of the analysis is optional and can be programmed as variables at the start of the script.

A summary of the proposed bioinformatic workflow can be found in Figure 4.

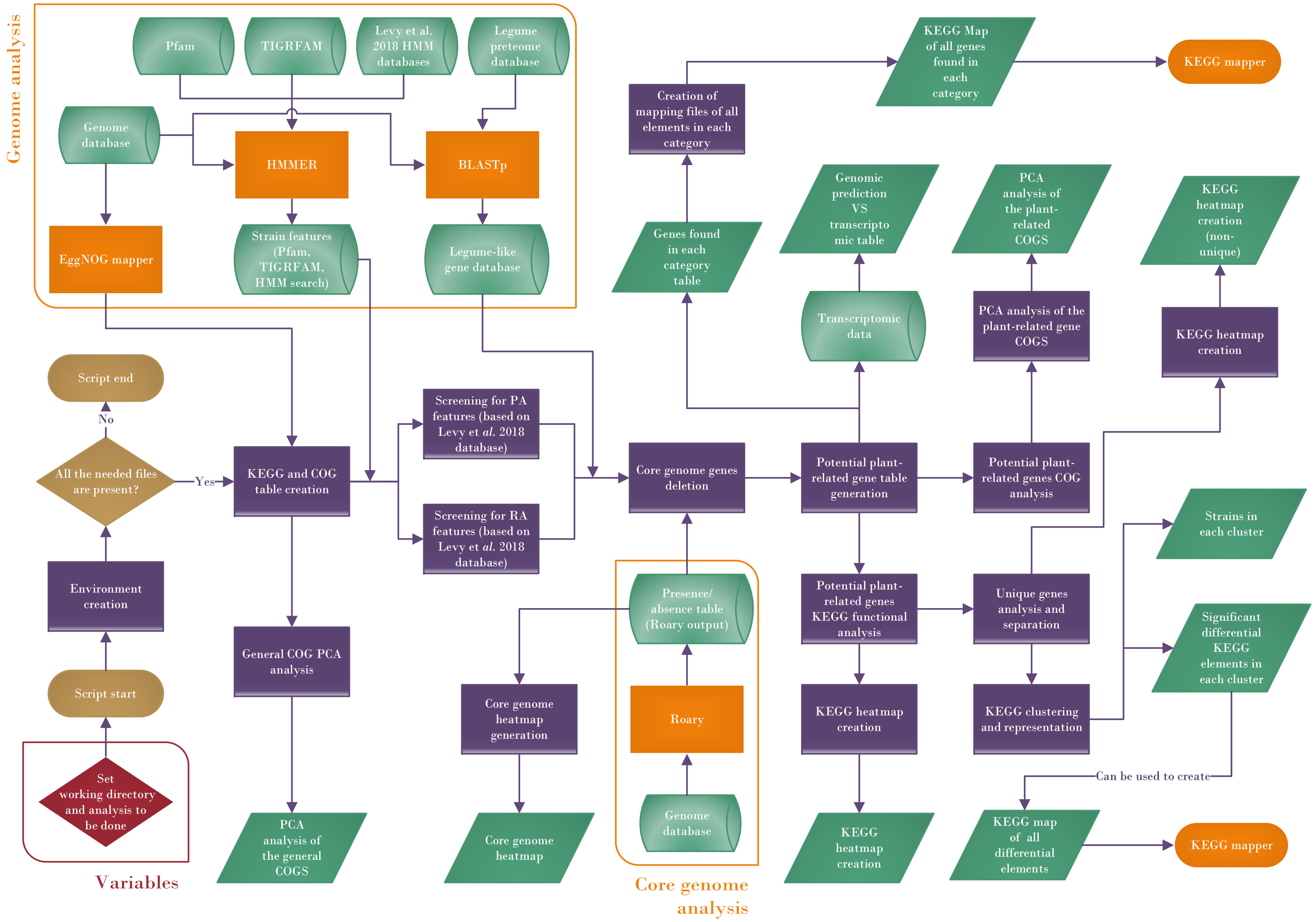

Genome analysis
Pfam
TIGRFAM
Levy et al. 2018 HMM databases
Legume preteome database
Genome database
HMMER
BLASTp
EggNOG mapper
Strain features (Pfam, TIGRFAM, HMM search)
Legume-like gene database
Script end
No
All the needed files are present?
Yes
KEGG and COG table creation
Screening for PA features (based on Levy et *al.* 2018 database)
Screening for RA features (based on Levy et *al.* 2018 database)
Environment creation
General COG PCA analysis
Script start
Set working directory and analysis to be done
Variables
PCA analysis of the general COGS
Core genome heatmap generation
Core genome heatmap
Core genome genes deletion
Presence/ absence table (Roary output)
Roary
Genome database
Core genome analysis
Creation of mapping files of all elements in each category
Genes found in each category table
KEGG Map of all genes found in each category
KEGG mapper
Genomic prediction VS transcriptomic table
Transcriptomic data
Potential plant-related gene table generation
Potential plant-related genes KEGG functional analysis
KEGG heatmap creation
KEGG heatmap creation
PCA analysis of the plant-related COGS
PCA analysis of the plant-related gene COGS
Potential plant-related genes COG analysis
Unique genes analysis and separation
KEGG clustering and representation
KEGG map of all differential elements
KEGG mapper
KEGG heatmap creation (non-unique)
KEGG heatmap creation
Strains in each cluster
Significant differential KEGG elements in each cluster
Can be used to create

Figure 4: Proposed bioinformatic workflow of MicroPLAGE. Diamonds represents choices, cylinders represent databases, squares represent process of the pipeline, rhomboids represent output data. In purple parts of the MicroPLAGE, in orange parts of external program, including HMMER, EggNOG, BLASTp and Roary, in red variables and in green databases and output data.

1.4- CONCLUSIONS

Three bioinformatic pipelines have been proposed in this thesis. These pipelines depend on three R scripts, that intend to be flexible enough to incorporate in any bioinformatic pipeline, with independence of the operative system used.

"*Micromonospora* Plant Associated Gene tool (MicroPLAGE)" intends to be an innovative method to screen for bacterial genomic features related to the bacterial-plant interaction at a genus level. It works under the assumption that there must be a genomic adaptation to the plant environment in the genus of study. *Micromonospora*, as a genus present in diverse environments, including plant-related environments, soils (from desertic soils, to forest soils) and marine environments, is the perfect candidate to test this script. In the second part of this thesis the proposed pipeline for MicroPLAGE will be the central analysis tool in the comparative genomic analysis.

"UBCG_iTOL_maker" and "GGDC Output Management Assistant (GOMA)" have been created to organize the input data and present the results of their corresponding base program (UBCG and GGDC) in an intuitive and more visual form. Of course, the scripts depend on the main base program, and sometimes they cannot avoid the main drawbacks of the base programs. For example, GOMA script starts from the output of the GGDC online service, and therefore require that the user upload all the genomes to make all the interactions. As GGDC server does not have any way to automatize the data upload at present, this drawback could not be solved. However, the capacity of the scripts to be incorporated in any bioinformatic pipeline entails that all these problems could be solved in the future if the main program is released. UBCG_iTOL_maker and GOMA will be used in the third part of this thesis, to highlight the presentation of the results.

1.5- REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.

Amaral, G. R. S., Dias, G. M., Wellington-Oguri, M., Chimetto, L., Campeão, M. E., Thompson, F. L., et al. (2014). Genotype to phenotype: Identification of diagnostic *Vibrio* phenotypes using whole genome sequences. *Int. J. Syst. Evol. Microbiol*. 64, 357–365. doi:10.1099/ijs.0.057927-0.

Arahal, D. R. (2014). "Whole-Genome Analyses," in *New Approaches to Prokaryotic Systematics*, eds. M. Goodfellow, I. Sutcliffe, and J. B. T.-M. in M. Chun (Academic Press), 103–122. doi:10.1016/bs.mim.2014.07.002.

Baek, I., Kim, M., Lee, I., Na, S.-I., Goodfellow, M., and Chun, J. (2018). Phylogeny Trumps Chemotaxonomy: A Case Study Involving *Turicella otitidis. Front. Microbiol*. 9, 834. doi:10.3389/fmicb.2018.00834.

Beaz-Hidalgo, R., Hossain, M. J., Liles, M. R., and Figueras, M. J. (2015). Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for *Aeromonas* genomes in the genbank database. *PLoS One* 10, 1–13. doi:10.1371/journal.pone.0115813.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Carro, L., Nouioui, I., Sangal, V., Meier-Kolthoff, J. P., Trujillo, M. E., Montero-Calasanz, M. del C., et al. (2018). Genome-based classification of *micromonosporae* with a focus on their biotechnological and ecological potential. Sci. Rep. 8, 525. doi:10.1038/s41598-017-17392-0.

Carro, L., Spröer, C., Alonso, P., and Trujillo, M. E. (2012). Diversity of *Micromonospora* strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. *Syst. Appl. Microbiol.* 35, 73–80. doi:10.1016/j.syapm.2011.11.003.

Chun, J., and Rainey, F. A. (2014). Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea. Int. J. Syst. Evol. Microbiol.* 64, 316–324. doi:10.1099/ijs.0.054171-0.

Cole, J. R., Konstantinidis, K., Farris, R. J., and Tiedje, J. (2010). *Microbial diversity and phylogeny: Extending from rRNAs to genomes.*

Creevey, C. J., Doerks, T., Fitzpatrick, D. A., Raes, J., and Bork, P. (2011). Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PLoS One 6. doi:10.1371/journal.pone.0022099.

Davis, T. L. (2019). *optparse:* Command Line Option Parser. Available at: https://cran.r-project.org/package=optparse.

Dowle, M., and Srinivasan, A. (2018). *data.table:* Extension of `data.frame`. Available at: https://cran.r-project.org/package=data.table.

Dragulescu, A. A., and Arendt, C. (2018). *xlsx:* Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. Available at: https://cran.r-project.org/package=xlsx.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.

Garcia, L. C., Martinez-Molina, E., and Trujillo, M. E. (2010). *Micromonospora pisi* sp. nov., isolated from root nodules of *Pisum sativum. Int. J. Syst. Evol. Microbiol.* 60, 331–337. doi:10.1099/ijs.0.012708-0.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733–739. doi:10.1038/nrmicro1236.

Glaeser, S. P., and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst. Appl. Microbiol.* 38, 237–245. doi:10.1016/j.syapm.2015.03.007.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities*. Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.*

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812.

Hahnke, R. L., Meier-Kolthoff, J. P., García-López, M., Mukherjee, S., Huntemann, M., Ivanova, N. N., et al. (2016). Genome-based taxonomic classification of *Bacteroidetes. Front. Microbiol.* 7. doi:10.3389/fmicb.2016.02003.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., Schuster, S. C., Huson, D. H., et al. (2004). Whole-genome prokaryotic phylogeny. *Bioinformatics* 21, 2329–2335. doi:10.1093/bioinformatics/bth324.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi:10.1093/molbev/msx148.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi:10.1093/nar/gkv1248.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 40, D109–D114. doi:10.1093/nar/gkr988.

Kasai, H., Tamura, T., and Harayama, S. (2000). Intrageneric relationships among *Micromonospora* species deduced from *gyr*B-based phylogeny and DNA relatedness. *Int. J. Syst. Evol. Microbiol.* 50, 127–134. doi:10.1099/00207713-50-1-127.

Kassambara, A., and Mundt, F. (2017). *factoextra*: Extract and Visualize the Results of Multivariate Data Analyses. Available at: https://cran.r-project.org/package=factoextra.

Katayama, T., Tanaka, M., Moriizumi, J., Nakamura, T., Brouchkov, A., Douglas, T. A., et al. (2007). Phylogenetic analysis of bacteria preserved in a permafrost ice wedge for 25,000 years. *Appl. Environ. Microbiol.* 73, 2360–2363. doi:10.1128/AEM.01715-06.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes*. Genome Biol.* 5, R12. doi:10.1186/gb-2004-5-2-r12.

Lê, S., Josse, J., and Husson, F. (2008). {*FactoMineR*}: A Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi:10.18637/jss.v025.i01.

Lee, I., Kim, Y. O., Park, S. C., and Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi:10.1099/ijsem.0.000760.

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. *Mol. Biol. Evol.* 32, 2798–2800. doi:10.1093/molbev/msv150.

Lemon, J. (2013). *Plotrix:* a package in the red light district of R. *R-News* 6, 8–12.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi:10.1093/nar/gkw290.

Levy, A., Conway, J. M., Dangl, J. L., and Woyke, T. (2018a). Elucidating Bacterial Gene Functions in the Plant Microbiome. *Cell Host Microbe* 24, 475–485. doi:10.1016/j.chom.2018.09.005.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018b). Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50, 138–150. doi:10.1038/s41588-017-0012-9.

Liu, W., Wang, Q., Hou, J., Tu, C., Luo, Y., and Christie, P. (2016a). Whole genome analysis of halotolerant and alkalotolerant plant growth-promoting rhizobacterium Klebsiella sp. D5A. Sci. Rep. 6, 26710. doi:10.1038/srep26710.

Liu, Y.-Y., Chiou, C.-S., and Chen, C.-C. (2016b). PGAdb-builder: A web service tool for creating pan-genome allele database for molecular fine typing. Sci. Rep. 6, 36213. doi:10.1038/srep36213.

Loper, J. E., Hassan, K. A., Mavrodi, D. V., Davis, E. W., Lim, C. K., Shaffer, B. T., et al. (2012). Comparative genomics of plant-associated *Pseudomonas* spp.: Insights into diversity and inheritance of traits involved in multitrophic interactions. *PLoS Genet.* 8, e1002784. doi:10.1371/journal.pgen.1002784.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). *cluster*: Cluster Analysis Basics and Extensions.

Martens, M., Dawyndt, P., Coopman, R., Gillis, M., De Vos, P., and Willems, A. (2008). Advantages of multilocus sequence analysis for taxonomic studies: A case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *Int. J. Syst. Evol. Microbiol.* 58, 200–214. doi:10.1099/ijs.0.65392-0.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2014). Highly parallelized inference of large genome-based phylogenies. *Concurr. Comput. Pract. Exp.* 26, 1715–1729. doi:10.1002/cpe.3112.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., Göker, M., and Access, O. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., and Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun.* 10, 2182. doi:https://doi.org/10.1038/s41467-019-10210-3.

Montero-Calasanz, M. del C., Meier-Kolthoff, J. P., Zhang, D. F., Yaramis, A., Rohde, M., Woyke, T., et al. (2017). Genome-scale data call for a taxonomic rearrangement of *Geodermatophilaceae*. *Front. Microbiol.* 8. doi:10.3389/fmicb.2017.02501.

Na, S.-I., Kim, Y. O., Yoon, S.-H., Ha, S., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi:10.1007/s12275-018-8014-6.

Neuwirth, E. (2014)*. RColorBrewer*: ColorBrewer Palettes. Available at: https://cran.r-project.org/package=RColorBrewer.

Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., et al. (2018). Genome-Based Taxonomic Classification of the Phylum *Actinobacteria. Front. Microbiol*. 9, 2007. doi:10.3389/fmicb.2018.02007.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi:10.1093/bioinformatics/btv421.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5, e9490. doi:10.1371/journal.pone.0009490.

R Development Core Team, R., and R Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria doi:10.1007/978-3-540-74686-7.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2016). JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. doi:10.1093/bioinformatics/btv681.

Rodriguez, L. M., and Konstantinidis, K. T. (2014). Bypassing Cultivation To Identify Bacterial Species. *Microbe Mag.* 9, 111–118. doi:10.1128/microbe.9.111.1.

RStudio Team (2016). RStudio: Integrated Development Environment for R. Available at: http://www.rstudio.com/.

Schutten, G.-J., Chan, C., Leeper, T. J., and other contributors (2018). *readODS*: Read and Write ODS Files. Available at: https://cran.r-project.org/package=readODS.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Stephens, J., Simonov, K., Xie, Y., Dong, Z., Wickham, H., Horner, J., et al. (2018). *yaml*: Methods to Convert R Data to YAML and Back. Available at: https://cran.r-project.org/package=yaml.

Sutcliffe, I. C., Trujillo, M. E., Whitman, W. B., and Goodfellow, M. (2013). A call to action for the International Committee on Systematics of Prokaryotes. *Trends Microbiol.* 21, 51–52. doi:10.1016/j.tim.2012.11.004.

Tang, Y., Horikoshi, M., and Li, W. (2016). *ggfortify:* Unified Interface to Visualize Statistical Results of Popular R Packages. R J. 8.2 8, 478–489. doi:10.1016/j.cattod.2005.03.010.

Thuleau, S., and Husson, F. (2018). *FactoInvestigate:* Automatic Description of Factorial Analysis. Available at: https://cran.r-project.org/package=FactoInvestigate.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W., and Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol*. 60, 249–266. doi:10.1099/ijs.0.016949-0.

Trujillo, M. E., Bacigalupe, R., Pujic, P., Igarashi, Y., Benito, P., Riesco, R., et al. (2014). Genome Features of the Endophytic Actinobacterium *Micromonospora lupini* Strain Lupac 08: On the Process of Adaptation to an Endophytic Life Style? PLoS One 9, e108522. doi:10.1371/journal.pone.0108522.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York Available at: http://ggplot2.org.

Wickham, H., and Bryan, J. (2018). *readxl:* Read Excel Files. Available at: https://cran.r-project.org/package=readxl.

Wickham, H., and Henry, L. (2018). *tidyr:* Easily Tidy Data with "spread()" and "gather()" Functions. Available at: https://cran.r-project.org/package=tidyr.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea. Nature* 462, 1056–1060. doi:10.1038/nature08656.

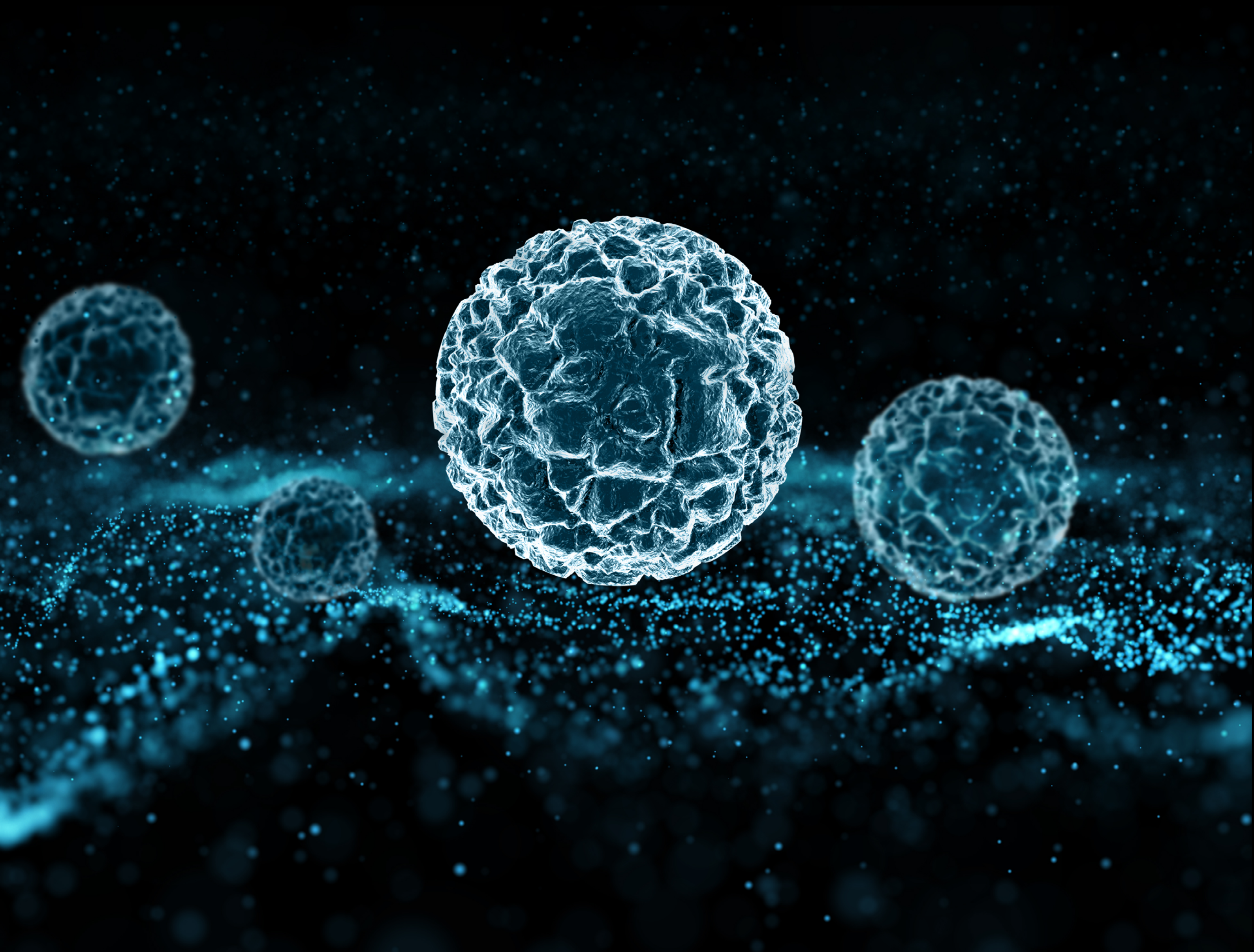
Wu, L., and Ma, J. (2019). The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/ijsem.0.003276.

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017a). Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* 67, 1613–1617. doi:10.1099/ijsem.0.001755.

Yoon, S. H., Ha, S. min, Lim, J., Kwon, S., and Chun, J. (2017b). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol*. 110, 1281–1286. doi:10.1007/s10482-017-0844-4.

CHAPTER II

On the hunt for genomic features of *Micromonospora* adaptation to plants



2.1- INTRODUCTION

Micromonospora is a Gram-positive bacterium with a wide geographical distribution. Although it is mainly isolated from soils, members of this genus have been recovered from marine and freshwater environments and even from animal tissues (de Menezes et al., 2012; Genilloud, 2015; Veyisoglu et al., 2016; Zhang et al., 2012). In the last decade, *Micromonospora* has been recovered from diverse plant tissues, specially from nitrogen fixing nodules of actinorhizal and leguminous plants (Carro et al., 2012b, 2013; Trujillo et al., 2010; Valdés et al., 2005). Although this bacterium has been isolated from nodules using the same protocol as rhizobia, it had been largely overlooked due to its slow growth when compared to rhizobial strains (Carro, 2009; Cerda, 2008; de la Vega, 2010).

The presence of *Micromonospora* in nodular tissues was confirmed for the first time using fluorescence *in-situ* hybridization (FISH) and transmission electron microscopy (TEM), suggesting a close interaction between the bacterium and the plant (Trujillo et al., 2010). Subsequent monitorization of the colonization process in three different plants, *Lupinus, Medicago* and *Trifolium* using a GFP-tagged *Micromonospora* strain, Lupac 08, in combination with the corresponding nitrogen fixer (rhizobia), and coupled with immunogold labeling (Benito et al., 2017) confirmed the capacity of *Micromonospora* to colonize the plant cells and suggested that a non-specific relationship takes place between the bacterium and the plant (Benito et al., 2017). Strain Lupac 08 was localized in all nodular tissues, confirming its capacity to enter and colonize the three hosts.

In order to unveil potential traits in the *Micromonospora*-plant relationship, genomic information is essential to help explain some of the complex mechanisms involved in this interaction. The genome sequence of *M. lupini* Lupac 08, isolated from a lupin nodule (Trujillo et al., 2007), was determined to identify genomic features potentially involved in this plant-microbe interaction (Alonso-Vega et al., 2012; Trujillo et al., 2014). The annotated genome disclosed various traits potentially involved in the capacity of this bacterium to alternate a lifestyle as a saprophyte in the soil and as an endophyte inside the root nodules (Trujillo et al., 2014).

These strategies included several characteristics commonly found in endophytic strains, such as the presence of siderophores, phytohormones and survival systems against plant defenses. This study also highlighted a wide array of plant cell wall degrading enzymes encoded in the genome. However, as no additional genomic information was available, this study could not compare these results with other *Micromonospora* strains, and no common plant-related genomic features could be defined above the strain level.

In recent years, an important number of *Micromonospora* genomes has been sequenced (Carro et al., 2018). This work opened up the possibility to carry out comparative genomic analyses to search for plant-related traits in this genus. Nevertheless, the number of *Mi-*

cromonospora strains isolated from plant tissues with available genomes is still low when compared to the soil environment. To increase the number of sequenced endophytic *Micromonospora* genomes, in this work we have sequenced seventeen new genomes from *Micromonospora* strains isolated from several legumes and different tissues. With the addition of these newly sequenced genomes, we have constructed a database of 74 genomes, with an almost equal number of soil-related and endophytic-related *Micromonospora* genomes. Using a novel comparative genomic approach, supported by the database generated in 2018 (Levy et al., 2018) and the proteome of known host plants, we have determined several genomic features that could potentially be related to the *Micromonospora*-plant interaction.

Genome-wide association studies (GWAS) are a very powerful tool in genomics that tries to match a genetic component with its correspondent phenotype by comparing multiple genomes (Falush and Bowden, 2006; Saber and Shapiro, 2020). In comparison with traditional molecular approaches that select a DNA sequence and test its effect in the phenotype (bottom-up approach), GWAS are a top-down approach that starts with the phenotype and associates differences in phenotype with differential regions of th e bacterial genome (Falush and Bowden, 2006; Sillanpää and Corander, 2002). GWAS has the potential to reveal the genetic features involved with relevant microbial phenotypes such as antibiotic resistance and virulence (Bandoy and Weimer, 2019; Sutton et al., 2019). In 2018, Levy and colleagues developed a computational approach to identify plant-associated genes and root-associated genes based on comparison of phylogenetically related genomes with known origin of isolation. This work used a wide association analysis of 3837 genomes to identify what genomic traits could explain their isolation origin, generating a huge plant-related features database (Levy et al., 2018). In this work, our aim was to develop an alternative pipeline to identify *Micromonospora* strains with the highest association to plants, using a comparative functional analysis of 69046 genes potentially related to the microbial adaptation to plants.

2.2- OBJECTIVES

The present study was designed to study the genomic features that drives the relationship between *Micromonospora* and the plant, using a comparative genomic approach in a database of 74 genomes. To achieve this, several specific aims were established:

1. To generate a database of all potentially plant-related genes contained in the genomic database, using a novel bioinformatic pipeline that comprises a screening in known microbial-plant related features databases, a comparation with known *Micromonospora* plant host proteomes and a comparative genomic analysis of the *Micromonospora* genomic database.

2. To characterize all functional features that differentiate strains that are close to the plant life in the genomic database from the rest, using the Clusters of Orthologs Groups (COG) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) for the comparative functional analysis.

3. To create a novel and reliable bioinformatic pipeline that englobes the two above-mentioned objectives.

2.3-MATERIALS AND METHODS

2.3.1- Isolation of strains

Soil from an agriculture field, collected at the "Fundación Vicente Rodríguez Fabrés" was sampled on October 2015. The complete coordinates for the collection site were N 40°57'26.96", W 5°39'37.28" (Figure 5). The soil was stored in plastic sealed bags and kept at 4 °C until processing.

Commercial *Medicago* sp. seeds were sterilized using a solution of 70% (v/v) ethanol, for 30 seconds, followed by 2.5% (v/v) $HgCl_2$ for five minutes and finishing with six washes in sterile distilled water, for one minute each (Benito et al., 2017). Sterilized seeds were placed in pots using the collected soil and kept in a plant growth chamber for a month, programmed with mixed incandescent and fluorescent lighting for a 16h photoperiod, day-night cycle, a constant temperature of 21-22 °C and 50-60% relative humidity. The pots were watered alternating sterile distilled water and nitrogen-free nutrient solution (Rigaud and Puppo, 1975) as needed.

Mature plants were harvested after one month of growth. To eliminate all the remaining soil, root systems were washed with sterile distilled water. For bacterial isolation, all nodules were surface sterilized with 2.5% (v/v) $HgCl_2$ for two minutes, rinsed five times in sterile distilled water and then crushed in a microtube using a sterile homogenizing pestle. The slurry obtained was plated on yeast mannitol agar (YMA1[1]) (Vincent, 1970) and incubated for 3-4 weeks at 28°C.

After incubation, colonies with a morphological resemblance to *Micromonospora* (Genilloud, 2015) were selected and picked under a stereoscopic microscope. All selected colonies were plated on DSMZ medium 65 (M652[2]) (Shirling and Gottlieb, 1966) to obtain pure cultures which were then conserved at -80 °C in 20% (v/v) glycerol suspensions.

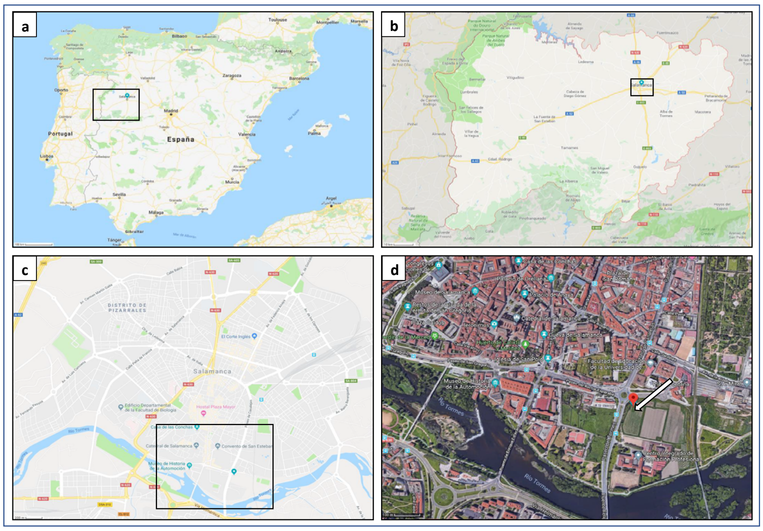


Figure 5: Location of the sampling site. Maps were downloaded from Google® Maps 2019 database.

[1] YMA: Mannitol (10g), yeast extract (3g), K_2HPO_4 (0.2g), $MgSO_4$ (0.2g), NaCl (0.1g), agar (18g), distilled water (1l).

[2] M65: Glucose (4g), yeast extract (4g), malt extract (10g), $CaCO_3$ (2g), agar (18g), distilled water (1l).

2.3.2- 16S rRNA gene identification of isolated strains

Genomic DNA was extracted for molecular identification of the isolated strains to confirm their preliminary identification based on their morphology. A pellet of cells after 7-14 days of growth on M65 agar was harvested in a microtube, resuspended in 300 µl of distilled water and centrifuged at 12000 rpm[3], for 10 minutes. Biomass was stored at -20 °C until further processing.

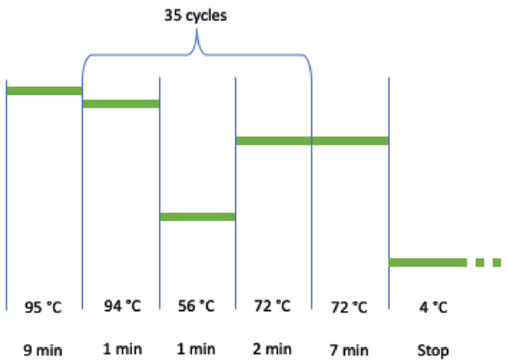
"REDExtract-N-Amp Plant PCR Kit" (Sigma™) was used for DNA extraction according to the following protocol (de la Vega, 2010):

1. Biomass is resuspended in 80µl of "extraction solution".
2. Heat at 95 °C for 10 minutes for cellular lysis.
3. Add 80 µl of "dilution solution".
4. Add 100 µl of phenol-chloroform-isoamyl alcohol solution (25:24:1).
5. Briefly vortex the sample to obtain a uniform milky suspension and centrifuge at 13000 rpm for 10 min.
6. Transfer 100 µl of the supernatant to a new microtube and discard the rest. Add 100 µl of a chloroform-isoamyl alcohol solution (24:1) and mix well.
7. Centrifuge the sample at 13000 rpm for 5 min. Collect 80 µl of the supernatant to a new microtube and discard the rest. Samples were stored at -20 °C.

Amplification of the 16S rRNA gene by polymerase chain reaction (PCR) was carried out in a final volume of 25 µl that contained:

REDExtract mix	12 µl
Extraction-Dilution 1:1 (v/v)	2.5 µl
SF1 primer (Table 1)	0.8 µl
1522R primer (Table 1)	0.8 µl
milliQ H_2O	8 µl
DNA (template)	1 µl

PCR was carried out under the following conditions:



[3] If not specified otherwise, the sample is centrifuged in an Eppendorf centrifuge model 5418, using an 18-place Aerosol-tight fixed rotor for 1.5-2.0 ml tubes (FA-45-18-11) with a rotor radius of 7.7cm. For a quick conversion, one revolution per minute (rpm) equals approximately to 0.8609 g.

PCR amplicons were loaded in a 1% agarose gel and electrophoresed for one hour at 100V, using Thermo Scientific™ GeneRuler 100bp DNA ladder as reference for fragment size. Fragments of approximately 1500 base pairs (bp) were selected and purified using FavorPrep™ GEL/PCR Purification kit following the manufacturer's protocol with some modifications:

1. Transfer up to 300 mg of the excised gel to a microtube and add 500 µl of FADF Buffer.
2. Incubate at 55 °C in a thermoblock until the agarose gel melts, manually mixing the sample every 1-2 minutes.
3. Cool down the mixture and transfer up to 800 µl to a FADF Column. Centrifuge at 14000 rpm for 30 seconds and discard the flow-through. If the volume is higher than 800 µl in the sample, repeat step 3 in the same column until all sample is transferred.
4. Add 750 µl of Wash buffer to the column and wait 5 minutes. Centrifuge 30 seconds at 14000 rpm and discard the flow-through.
5. Centrifuge again at 14000 rpm for 3 minutes to dry the column matrix and discard the flow-through.
6. Place the column in a new microtube. Add 35 µl of distilled milliQ water (or elution buffer provided with the kit) and wait 5 minutes for DNA elution.
7. Centrifuge at 14000 rpm for five minutes and eliminate the column. The final DNA sample in the microtube is stored at -20 °C until further use.

The purified amplicon sample was quantified using Invitrogen life technologies™ Qubit® fluorometer, following the recommended protocol:

1. Prepare 200 µl/sample of working solution (Qubit reagent- Qubit Buffer 1:199 (v/v)). Each time a new Working solution is prepared, Qubit fluorometer must be calibrated using two standard DNA samples.
2. Prepare the assay tubes, using Qubit 0.5 ml tubes:
   - For the standards, 10 µl of the standard is mixed with 190 µl of the working solution.
   - For the samples, 1-20 µl of the DNA sample is used, and mixed with working solution in a final volume of 200 µl. For 16S rRNA gene amplicons, a volume of 3 µl was used.
3. Vortex all assay tubes for 2-3 seconds and incubate the tubes at room temperature for at least 2 minutes for the reaction between the DNA and the Qubit® reagent to take place. The sample can be stored at room temperature for a maximum of 2 hours.
4. Insert the tubes in the Qubit fluorometer and take the readings. For the 16S rRNA gene, "dsDNA" and "High sensitivity" reading parameters were used.

Once quantified, the 16S rRNA gene amplicons were partially sequenced at the DNA Sequencing NUCLEUS service at the University of Salamanca. Primer SR2 (Table 1) (Carro et al., 2012b) was used to obtain a sequence of approximately 450 bp between the positions 1 and 487 of the 16S rRNA gene sequence. The strains were identified using EzBiocloud 16S identification service (Yoon et al., 2017). Strains with high similarity to *M. saelicesensis* were chosen for complete 16S rRNA gene sequencing, using primers SR3, SR4 and 1522R (Table 1) (Carro et al., 2012b). All reads were assembled using DNAStar® SeqMan v5.0 and identified using EzBiocloud 16S identification service (Yoon et al., 2017).

Table 1: Primers used for amplification and sequencing of the 16S rRNA gene. Primer melting temperature (Tm), hairpin melting temperature (Hairpin Tm) and self-dimer melting temperature (Self Dimer Tm) were estimated using Primer3 v2.3.7 under GeneiousTM v2019.0.4 environment.

Primer	**Sequence**	**Primer length**	**Direction**	**%GC**	**Tm**	**Min**	**Max**	**Hairpin Tm**	**Self Dimer Tm**
	(5´-3´)	**(bp)**			**(°C)**	**(bp 5′ -3′)**	**(bp 5′ -3′)**	**(°C)**	**(°C)**
SF1	AGAGTTTGATCMTGGCT-CAG	20	forward	47.4	55.1-56.9	5	24	-	18.0
SR2	GWATTACCGCGGCKGCTG	18	reverse	64.7	58.7-61.9	487	504	53.1	38.8
SR3	CCGTCAATTC-MTTTRAGTTT	20	reverse	33.3	49.9-54.1	877	896	44.9	-
SR4	GGGTTGCGCTCGTTG	15	reverse	66.7	55.5	1067	1081	-	9.7
1522R	AAGGAGGTGWTCCARCC	17	reverse	56.3	52.5-55.8	1497	1513	52.3	-

2.3.3- Strains selected for whole genome sequencing

A set of sixteen *Micromonospora* strains isolated from plant tissues (nodules and leaves) of six legumes (*Medicago, Lupinus, Pisum, Trifolium, Cicer* and *Ononis*) was selected for whole genome sequencing. All selected strains were chosen according to their 16S rRNA gene sequence similarity to the type strains of *Micromonospora saelicesensis* and *Micromonospora noduli*, as these species appear to be the most abundant in legumes (Table 2) (Carro et al., 2012b). The genome of *M. noduli* GUI43^{T} was also selected for whole genome sequencing.

Table 2: List of selected strains for whole genome sequencing. Identity values against the type strains of *M. saelicesensis* and *M. noduli* were determined using BLAST v2.7.1 against GenBank reference sequences (accession numbers AJ783993 and FN658649).

Strain	Host legume	Isolation tissue	*M. saelicesensis* (%)	*M. noduli* (%)
GAR05	*Cicer arietinum*	Nodule	99.9	99.6
GAR06	*C. arietinum*	Nodule	99.9	99.7
LAH08	*Lupinus angustifolius*	Leaf	99.7	99.9
LAH09	*L. angustifolius*	Leaf	99.4	99.3
Lupac 06	*L. angustifolius*	Nodule	99.9	99.6
Lupac 07	*L. angustifolius*	Nodule	99.9	99.7
MED01	*Medicago sp.*	Nodule	99.9	99.7
MED15	*Medicago sp.*	Nodule	99.2	100
NIE79	*Trifolium sp.*	Nodule	99.9	99.6
NIE111	*Trifolium sp.*	Nodule	99.9	99.6
ONO23	*Ononis sp.*	Nodule	99.6	100
ONO86	*Ononis sp.*	Nodule	99.7	99.9
PSH03	*Pisum sativum*	Leaf	99.5	99.6
PSH25	*P. sativum*	Leaf	99.7	99.4
PSN01	*P. sativum*	Nodule	99.9	99.7
PSN13	*P. sativum*	Nodule	99.9	99.7
***M. noduli* GUI43^T**	*P. sativum*	Nodule	99.6	100

2.3.4- Sequencing, Assembly and Annotation

Strains selected for genome sequencing were grown at 28 °C for 7-14 days in M65 broth, harvested and washed with 0.8% (w/v) NaCl solution in a centrifuge tube. The following protocol was used for DNA extraction:

1. One gram of cells is resuspended in 5 ml EC buffer (Tris-Cl 6 mM, Ethylenediaminetetraacetic acid (EDTA) 0.1 M, *N*- Lauroylsarcosine sodium salt 1% w/v, sodium deoxycholate 0.2% w/v). 60 µl of lysozyme (300 mg/ml) and 50 µl of mutanolysin (1000 U/ml) are added to the suspension.

2. Incubate tubes 1 - 1.5h at 37 °C in a water bath.
3. Add 5ml of 2% (w/v) SDS and 200 µl of proteinase K (10 mg/ml) with gentle mixing.
4. Incubate the tube at 55 °C in a water bath for 3 hours with gentle mixing from time to time (approx. 1 per hour).
5. Add one volume of phenol- chloroform -isoamyl alcohol (25:24:1) to the lysate. Gently mix by manually inverting the tube several times.
6. Centrifuge the mixture at 6000 rpm (5000 g approx.) for 15 minutes. Transfer the supernatant into a fresh tube.
7. Add 35 µl of RNAse A (10mg/ml) to the supernatant and incubate for 1 hour at 55 °C in a water bath.
8. Add one volume of chloroform- isoamyl alcohol (24:1). Mix the sample by inverting the tube several times.
9. Centrifuge at 6000rpm (5000 g approx.) for 10 minutes at room temperature.
10. Transfer the aqueous phase (upper) into a clean tube.
11. Add 1/10 volume of 3M sodium acetate pH 7.0, followed by 2.5-3 volumes of cold ethanol (-20 °C).
12. Invert tube to mix the sample. DNA should be visible as white strands. Leave the sample approximately 30 minutes at -20 °C and spool the DNA into a clean microtube.
13. Wash the DNA with 70% ethanol and centrifuge at max speed for 5 minutes.
14. Discard the supernatant and allow the pellet to dry for 10-15 minutes. If the pellet is not dry, incubate at 37 °C in the thermoblock. Resuspend the DNA in 150 µl of TE buffer (Tris-HCl 10 mM, EDTA 1mM) or 150 µl of milliQ H_2O.

DNA samples were sequenced at ChunLab Inc. Libraries were prepared using TruSeq™ library kit, according to the manufacturer instructions. Sequencing was performed using an Illumina® MiSeq™ platform (300-bp paired-end), with a coverage superior to 100X.

Illumina sequencing data was assembled with SPAdes 3.10.1 (Algorithmic Biology Lab, St. Petersburg Academic University of the Russian Academy of Sciences). Protein-coding sequences (CDSs) were predicted by Prodigal 2.6.2 (Hyatt et al., 2010). Genes coding for tRNA were searched using tRNAscan-SE 1.3.1 (Schattner et al., 2005). The rRNA and other non-coding RNAs were searched by a covariance model search with Rfam 12.0 database (Nawrocki et al., 2015). All genomes were functionally annotated using the new EggNOG-mapper (Huerta-Cepas et al., 2017) with HMMER mapping mode against actNOG and bacterial HMM databases using all Orthologs (Huerta-Cepas et al., 2016). To confirm annotation, the predicted CDSs were compared with Swissprot

(The UniProt Consortium, 2017), KEGG (Kanehisa et al., 2014) and SEED (Overbeek et al., 2005) databases using UBLAST program (Edgar, 2010). CRISPR elements were retrieved using the online application CRISPR-finder, available in http://crispr.i2bc.paris-saclay.fr (Grissa et al., 2007) using default parameters.

2.3.5- Data compilation and proteome annotation

Fifty-seven *Micromonospora* and *Salinispora* genomes were retrieved from GenBank and IMG databases (Clark et al., 2016; Markowitz et al., 2012) and added to the seventeen *Micromonospora* genomes obtained. The final database, comprising seventy-four genomes, contained representatives isolated from soil (39%), endophytes (34%), sediment (16%), marine (4%) and diverse environments that included animals, air and fresh water (7%) (Table 3). All genomes were checked for contamination using CheckM in KBase environment (Arkin et al., 2018; Parks et al., 2015).

For data normalization, the 74 proteomes were re-annotated using the following databases:

1. ***Pfam:*** Is a database based on the UniProt Reference Proteome database (The UniProt Consortium, 2017), that contains a large collection of protein families classified according to their functional regions (or domains), each represented by sequence alignments and hidden Markov models (HMMs) (Finn et al., 2016). The version used in this study was the 31.0, released in 2017.

2. ***TIGRFAM:*** Is a curated database of multiple sequence alignments and HMMs for prokaryotic protein sequence classification, designed to increase functional identification of proteins (Haft et al., 2001). The version used in this study was the 15.0, released in 2014.

3. ***EggNOG:*** is a public database that contains orthologous groups of proteins, constructed with a taxonomical grouping approach that generates a more accurate identification of proteins and allows to increase the overall annotation coverage. EggNOG database provides functional annotation of Clusters of Orthologous Groups (COGs), Gene Ontology terms (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Smart/Pfam domains (Huerta-Cepas et al., 2016). The version used in this study was the 4.5.1, released in 2016.

4. ***Genomic Features Of Bacterial Adaptation to Plants (GFOBAP):*** Released in 2018, it contains several plant-related bacterial features, inferred from a comparative genomic analysis of 3837 genomes (Levy et al., 2018). This database comprises the results of the statistical enrichment analysis and the *fasta* files, alignments and HMM profiles of the orthogroups generated in the study of 2018. The database is divided according to taxonomic monophyletic groups, being the *Actinobacteria1* database the one used in this study.

hmmsearch, included within HMMER version 3.1b2 program (hmmer.org) was used to annotate all proteomes against Pfam, TIGRFAM and GFOBAP HMM protein profiles, under a Linux environment. EggNOG-mapper online tool (Huerta-Cepas et al., 2017) was used to annotate all proteomes against the EggNOG bacterial database. Secondary metabolite biosynthetic gene clusters were screened using antiSMASH 4.0 (bacterial version), with *KnownClusterBlast*, *smCoG* analysis, *ActiveSiteFinder* and *SubClusterBlast* extra features (Blin et al., 2017).

Table 3: Genome sequence accession numbers and isolation source of the strains used in this work.

Designation	Accession number	Habitat	Designation	Accession number	Habitat
M. acroterricola **5R2A7^{T}**	QGKR00000000	Soil	***M. noduli*** **Lupac 07**	PYAB00000000	Endophyte
M. aurantiaca **ATCC 27029^{T}**	NC_014391	Soil	***M. noduli*** **MED15**	PYAC00000000	Endophyte
M. aurantiaca **DSM 45487**	FMHX00000000	Soil	***M. noduli*** **ONO23**	PYAD00000000	Endophyte
M. aurantiaca **L5**	NC_014815	Endophyte	***M. noduli*** **ONO86**	PYAE00000000	Endophyte
M. auratinigra **DSM 44815^{T}**	LT594323	Soil	***M. olivasterospora*** **DSM 43868^{T}**	jgi_2585427559	Soil
M. avicenniae **DSM 45758^{T}**	jgi_2681813563	Endophyte	***M. pallida*** **DSM 43817^{T}**	FMHW00000000	Soil
M. carbonacea **DSM 43168^{T}**	FMCT00000000	Soil	***M. palomenae*** **DSM 102131^{T}**	VIXA00000000	Animal
M. chaiyaphumensis DSM 45246^{T}	FMCS00000000	Soil	***M. pattaloongensis*** **DSM 45245^{T}**	jgi_ 2693429860	Sediment
M. chalcea **DSM 43026^{T}**	MAGP00000000	Air	***M. peucetia*** **DSM 43363^{T}**	FMIC00000000	Soil
M. chersina **DSM 44151^{T}**	FMIB00000000	Soil	***M. pisi*** **DSM 45175^{T}**	jgi_2758568729	Endophyte
M. chokoriensis **DSM 45160^{T}**	LT607409	Soil	***M. purpureochromogenes*** **DSM 43821^{T}**	LT607410	Soil
M. citrea **DSM 43903^{T}**	FMHZ00000000	Sediment	***M. rhizosphaerae*** **DSM 45131^{T}**	FMHV00000000	Sediment

M. coriariae **DSM 44875^{T}**	LT607412	Endophyte	***M. rifamycinica*** **DSM 44983^{T}**	LT607752	Sediment
M. costi **CS1-12^{T}**	RBAN00000000	Endophyte	***M. rosaria*** **DSM 803^{T}**	jgi_2728369162	Soil
M. coxensis **DSM 45161^{T}**	LT607753	Soil	***M. saelicesensis*** **Lupac 06**	PYAJ00000000	Endophyte
M. cremea **DSM 45599^{T}**	FSQT00000000	Soil	***M. saelicesensis*** **DSM 44871^{T}**	FMCR00000000	Endophyte
M. eburnea **DSM 44814^{T}**	FMHY00000000	Soil	***M. saelicesensis*** **GAR05**	PXXW00000000	Endophyte
M. echinaurantiaca **DSM 43094^{T}**	LT607750	Soil	***M. saelicesensis*** **GAR06**	PYAH00000000	Endophyte
M. echinofusca **DSM 43913^{T}**	LT607733	Animal	***M. saelicesensis*** **PSN01**	PYAI00000000	Endophyte
M. echinospora **DSM 43816^{T}**	LT607413	Soil	***M. saelicesensis*** **PSN13**	PYAG00000000	Endophyte
M. endolithica **DSM 44398^{T}**	jgi_2585427558	Soil	***M. sagamiensis*** **DSM 43912^{T}**	jgi_2585427560	Soil
M. globispora **S2901^{T}**	QGGF00000000	Sediment	***M. sediminicola*** **DSM 45794^{T}**	FLRH00000000	Sediment
M. haikouensis **DSM 45626^{T}**	FMCW00000000	Sediment	***M. siamensis*** **DSM 45097^{T}**	LT607751	Soil
M. halophytica **DSM 43171^{T}**	FMDN00000000	Saline	***Micromonospora sp.*** **LAH09**	-	Endophyte
M. humi **DSM 45647^{T}**	FMDM00000000	Soil	***Micromonospora sp.*** **MED01**	-	Endophyte
M. inaquosa **LB39^{T}**	QGSZ00000000	Soil	***Micromonospora sp.*** **NIE111**	-	Endophyte
M. inositola **DSM 43819^{T}**	LT607754	Soil	**Micromonospora sp.** **NIE79**	-	Endophyte
M. inyonensis **DSM 46123^{T}**	FMHU00000000	Soil	***Micromonospora sp.*** **PSH03**	-	Endophyte
M. krabiensis **DSM 45344^{T}**	LT598496	Sediment	***Micromonospora sp.*** **PSH25**	-	Endophyte
M. lupini **Lupac 08**	CAIE00000000	Endophyte	***M. tulbaghiae*** **DSM 45142^{T}**	FMCQ00000000	Endophyte

M. marina **DSM 45555^T**	FMCV00000000	Saline	***M. viridifaciens*** **DSM 43909^T**	LT607411	Soil
M. matsumotoense **DSM 44100^T**	FMCU00000000	Soil	***M. wenchangensis*** **CCTCC AA 2012002^T**	MZMV00000000	Sediment
M. mirobrigensis **DSM 44830^T**	FMCX00000000	water	***M. yangpuensis*** **DSM 45577^T**	FMIA00000000	Animal
M. narathiwatensis **DSM 45248^T**	LT594324	Soil	***M. zamorensis*** **DSM 45600^T**	LT607755	Soil
M. nigra **DSM 43818^T**	FMHT00000000	Saline	***S. arenicola*** **CNH-643^T**	jgi_2561511037	Sediment
M. noduli **GUI43^T**	PYAK00000000	Endo-phyte	***S. pacifica*** **CNR-114^T**	AZWO00000000	Sediment
M. noduli **LAH08**	PYAA00000000	Endo-phyte	***S. tropica*** **CNB-440^T**	NC_009380	Sediment

2.3.6- Selection of plant-related *Micromonospora* genes

2.3.6.1- Core genome analysis

A cut-off BLAST value was calculated using a pre-established bacterial core-gene set comprising 92 bacterial core genes described in the UBCG method (Na et al., 2018). These 92 genes were screened and aligned using UBCG 3.0 (Na et al., 2018) for all genomes in the database. Identity matrices were calculated for all the alignments, and the mean, median, maximum and minimum identity percentages were determined for each gene and the overall set.

Roary v 3.11.2 (Page et al., 2015) was used to define the core and pan-genome, using the previously calculated identity cut-off for the clustering of proteins. Protein clusters were used to generate an absence-presence table used for the analysis and a development plot. *ComplexHeatmap* R package (Gu et al., 2016) was used to make a pan-genome heatmap plot.

2.3.6.2- Selection of plant-related and root-related gene annotations

The selection of bacterial genes was based on a pre-defined dataset of plant and root-related genes, described previously (Levy et al., 2018). Considering the phylogenetic position of *Micromonospora* (Nouioui et al., 2018), the dataset was restricted to the first group of the *Actinobacteria* (*Actinobacteria*1 database).

Orthofinder groups, COGs, KEGG Orthologs (KO), Pfam and TIGRFAM within *Actinobacteria*1, “Reproducible Plant Associated Domains” and “Plant-Resembling Plant-Associated and Root-Associated Domains” (PREPARADOs) were included in the annotation search analysis. These annotations were taken in consideration only if two or more of the five statistical support analyses were positive, as recommended (Levy et al., 2018).

2.3.6.3- Plant-resembling bacterial proteins

Proteomes of known *Micromonospora* host plants were screened in UniprotKB database (release 2018_6) (The UniProt Consortium, 2017). Eighteen proteomes, comprising different species of *Cicer, Glycine, Lupinus, Medicago, Oryza, Phaseolus* and *Trifolium* were used to create a BLAST database, comprising 731325 proteins.

Proteomes of the 74 bacterial strains were blasted against the plant proteome database, using BLASTp stand-alone program, included in BLAST+ executables v. 2.7.1 (Camacho et al., 2009), with a threshold of $1e^{-30}$ for the E-value, 70% coverage and 30% identity. All identified proteins found in the analysis were labeled as "plant-resembling bacterial proteins" and included in the overall analysis.

2.3.7- Data management, statistical analysis and visual representation

MicroPLAGE script, described in Chapter I, was used to screen for all the potential plant related gene annotations and to carry out the statistical analyses.

In summary, all annotations for the 74 genomes were screened against GFOBAP curated annotation database (Levy et al., 2018) using *data.table* and *tidyr* packages (Dowle and Srinivasan, 2018; Wickham and Henry, 2018) in R v 3.5.1 (R Development Core Team and R Core Team, 2011). Additionally, plant resembling bacterial proteins were added to the database and all core conserved proteins were deleted.

Non-parametric relationships between habitat and genome length and number of potential plant-related genes and habitat were obtained using IBM® SPSS® Statistics v.25. Bar-plots for COG analyses of each strain were made using *ggplot2* and *ggfortify* packages (Tang et al., 2016; Wickham, 2016). *FactoMineR, factoextra, FactoInvestigate* and cluster packages (Kassambara and Mundt, 2017; Lê et al., 2008; Maechler et al., 2018; Thuleau and Husson, 2018) were used for Principal Component Analysis (PCA) and clustering analysis of the COGs and plant-related functional KEGG characterization (Chapter I). To further analyze the functional differentiation between all the strains included in the database and the groups detected, all the data in the functional KEGG analysis were compared using *factoextra* package principal component analysis and hierarchical clustering tools to highlight differences with statistical significance ($p<0.05$), not only due to absence/presence of the KEGG ortholog, but also in relation to their abundance. All unique strain elements were analyzed separately.

P values generated in KEGG abundance analysis were corrected using *p.adjust* tool, included in *stats* R native package, with Bonferroni adjustment method (Jafari and Ansari-Pour, 2019; R Development Core Team and R Core Team, 2011). Corrected p value (hereafter, q value) resulted in a corrected threshold for statistical significance of 1.31 x 10^{-5} for KEGG elements abundance in each calculated cluster (0.05 q value = 1.31 x 10^{-5} p value). Clusters generated in the KEGG analysis and habitat distribution of the strains in each cluster were tested for statistical correlation using a Pearson chi-square statistical approach, followed by a contingency table post-hoc analysis using multiple regression approach (Beasley and Schumacker, 1995) using IBM® SPSS® Statistics v.25.

ComplexHeatmap package (Gu et al., 2016) was used for the heatmap construction of the KEGG functional analysis and the core genome visual representation. Complete pathway mapping files of plant-associated, root-associated and plant-resembling bacterial proteins for each genome and also for the differential traits between the clusters generated in the PCA analysis were generated using R and visualized using KEGG Mapper online tool (Kanehisa et al., 2012) (Chapter I).

A flowchart explaining how the database was constructed is included in Figure 6.

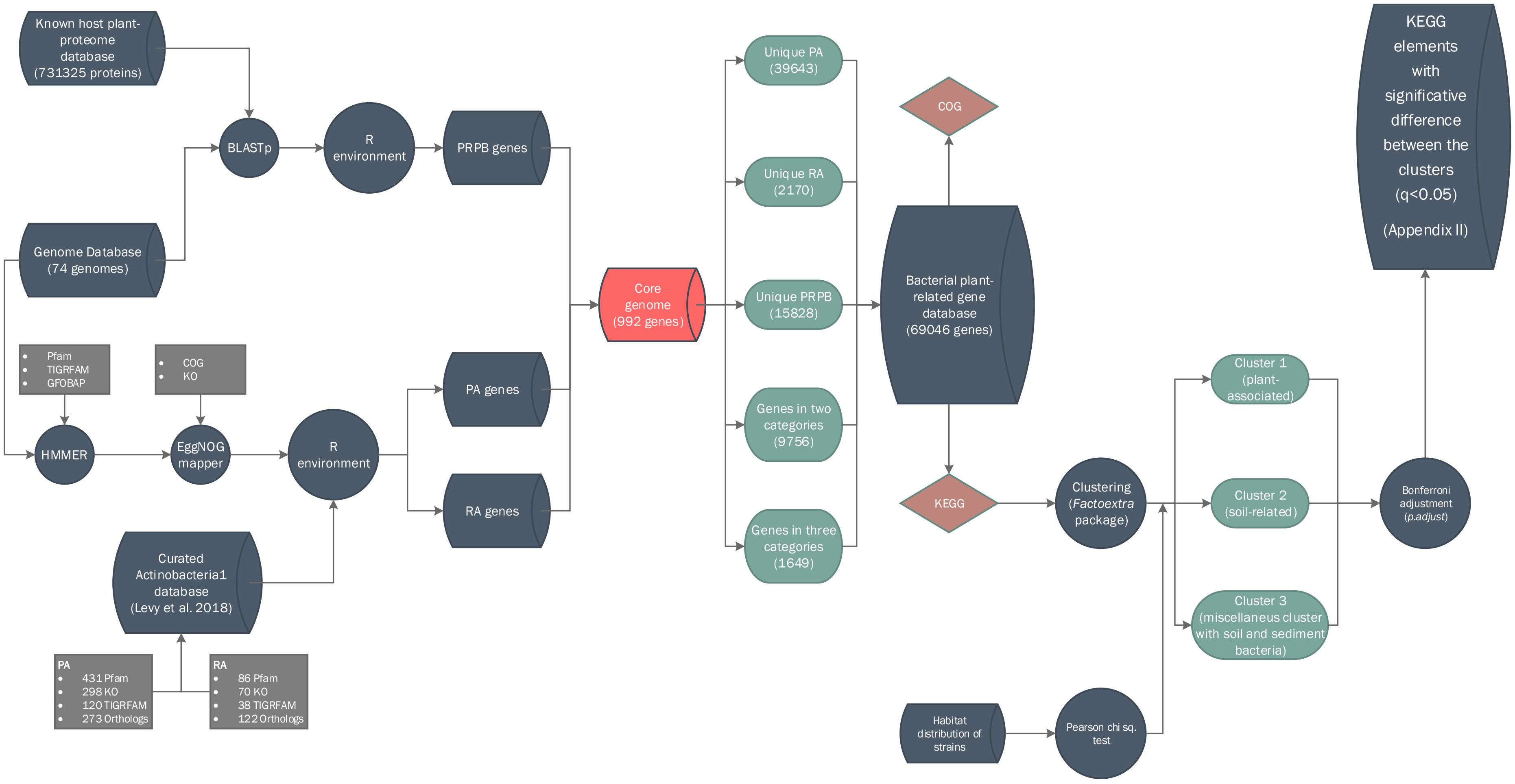

Known host plant-proteome database (731325 proteins)
Genome Database (74 genomes)
BLASTp
R environment
PRPB genes
Pfam
TIGRFAM
GFOBAP
COG
KO
HMMER
EggNOG mapper
R environment
PA genes
RA genes
Curated Actinobacteria1 database (Levy et al. 2018)
PA
431 Pfam
298 KO
120 TIGRFAM
273 Orthologs
RA
86 Pfam
70 KO
38 TIGRFAM
122 Orthologs
Core genome (992 genes)
Unique PA (39643)
Unique RA (2170)
Unique PRPB (15828)
Genes in two categories (9756)
Genes in three categories (1649)
COG
Bacterial plant-related gene database (69046 genes)
KEGG
Clustering (*Factoextra* package)
Habitat distribution of strains
Pearson chi sq. test
Cluster 1 (plant-associated)
Cluster 2 (soil-related)
Cluster 3 (miscellaneus cluster with soil and sediment bacteria)
Bonferroni adjustment (*p.adjust*)
KEGG elements with significative difference between the clusters (q<0.05)
(Appendix II)

Figure 6: Bioinformatic workflow of the construction of the database. Cylinders represent databases, circles represent processes of the pipeline, rhomboids represent annotations, grey squares represent output data. In green partial results.

2.4- RESULTS

2.4.1- Isolation and identification of the strains

Nineteen strains were isolated from *Medicago* internal nodule tissues, with thirteen of them having *Micromonospora*-like morphology (Genilloud, 2015). These strains were selected for partial 16S rRNA gene sequencing (Table 4).

16S rRNA gene amplicons of 386-444 base pairs were identified using EzBiocloud 16S-based identification service (Yoon et al., 2017). Affiliation to the genus *Micromonospora* was confirmed for all sequenced strains. Seven isolates (MED04, MED06, MED08, MED09, MED15, MED16 and MED20) were most similar to the type strain *Micromonospora noduli* GUI43^T; three (MED01, MED02 and MED07) were related to *M. saelicesensis* Lupac09^T, and the remaining three strains were close to either *M. zamorensis* DSM 45600^T (MED13 and MED18) or *M. lupini* Lupac 14N^T (MED05) (Figure 7). All isolates had values of 100% similarity with their closest match except for strain MED15 which yielded a value of 99.6% similarity against *M. lupini* Lupac 14N^T (Table 4).

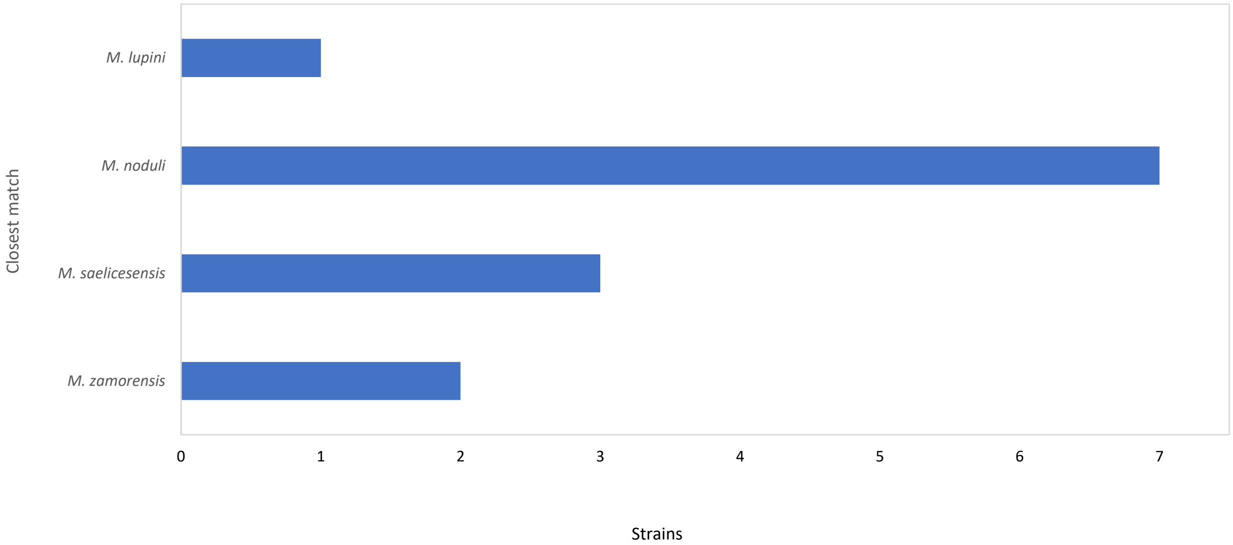


Figure 7: Closest hits according to partial 16S rRNA gene sequence identification in EzBiocloud.

Strains MED01, MED07 and MED15 were selected for complete 16S rRNA gene sequencing, as these were closest to *M. saelicesensis* and *M. noduli*, reported to be the most abundant species in legumes sampled hitherto (Carro et al., 2012b; Riesco et al., 2018). Strains MED01 and MED07 had an identity of 99.9% sequence similarity with *M. saelicesensis* Lupac 09^T, while strain MED15 was close to *M. noduli* GUI43^T, with an identity value of 100% (Table 4)

Table 4: *Micromonospora* strains isolated from *Medicago* sp. and their identity percentage based on partial and complete 16S rRNA gene sequences. Red-colored strains were selected for whole genome sequencing.

Strain	Closest match	Length	Identity
MED01	*M. saelicesensis Lupac* 09^{T}	1433	99.9%
MED02	*M. saelicesensis Lupac* 09^{T}	436	100%
MED04	*M. noduli* GUI43^{T}	435	100%
MED05	*M. lupini* Lupac 14N^{T}	440	99.6%
MED06	*M. noduli* GUI43^{T}	436	100%
MED07	*M.* saelicesensis Lupac 09^{T}	1440	99.9%
MED08	*M. noduli* GUI43^{T}	437	100%
MED09	*M. noduli* GUI43^{T}	415	100%
MED13	*M. zamorensis* DSM 45600^{T}	437	100%
MED15	*M. noduli* GUI43^{T}	1457	100%
MED16	*M. noduli* GUI43^{T}	444	100%
MED18	*M. zamorensis* DSM 45600^{T}	441	100%
MED20	*M. noduli* GUI43^{T}	376	100%

2.4.2- General genomic features of *Micromonospora*

Seventeen high-quality genomes (mean depth of 279x) were obtained from the selected *Micromonospora* strains (Table 2). Their sizes ranged from 6.8 to 7.6 Mb, with strains PSH25 and MED01 having the smallest and largest genomes respectively. G+C mol % values ranged from 70.8 to 71.6 (mean value 71.1 ± 0.2). The number of coding DNA sequences varied from 6182-7060, with a mean value of 6528 ± 196. The number of tRNA coding sequences was more variable, ranging from 45 (strain PSH25) to 84 (strain NIE79), with a mean number of 59 ± 9. rRNA coding sequences also deviated significantly, from 3 in NIE 111 to 8 in strain GAR06. Confirmed clustered regularly interspaced short palindromic repeats sequences (CRISPR) ranged from 0-4. A summary of the genomic characteristics can be found in Table 5.

The final genome database contained the seventeen *Micromonospora* strains sequenced in this work and the fifty-seven genomes retrieved from the public databases. Their genome lengths ranged from 5.2 Mb (*Salinispora tropica* CNB-440^{T}) to 8.7 Mb (*M. pisi* DSM 45175^{T}). Specifically, the *Micromonospora* genomes ranged from 5.4 Mb for *M. pattaloongensis* DSM 45245^{T} to 8.7 Mb for *M. pisi* DSM 45175^{T}, with an average and median values of 7.0 Mb (*Micromonospora aurantiaca* ATCC 27029^{T}). *Salinispora* genomes were significantly smaller, ranging from 5.2 Mb for *S. tropica* CNB-440^{T} to 5.9 Mb for *S. pacifica* CNR114^{T}, with a mean value of 5.6 Mb.

No clear correlation was found between genome size and habitat. Endophytic and soil related strains showed similar genome lengths (7.1 ± 0.4 Mb), while sediment isolates had a mean of 6.6 ± 0.9 Mb, with high dispersion values. Interestingly, *M. tulbaghiae* DSM 45142^T isolated from leaves of *Tulbaghia violacea* (Kirby and Meyers, 2010) and *M. pisi* DSM 45175^T, isolated from nodular *Pisum* tissue (Garcia et al., 2010) were recovered as outliers among the endophytes, with genome sizes of 6.5 and 8.7 Mb respectively (Figure 8).

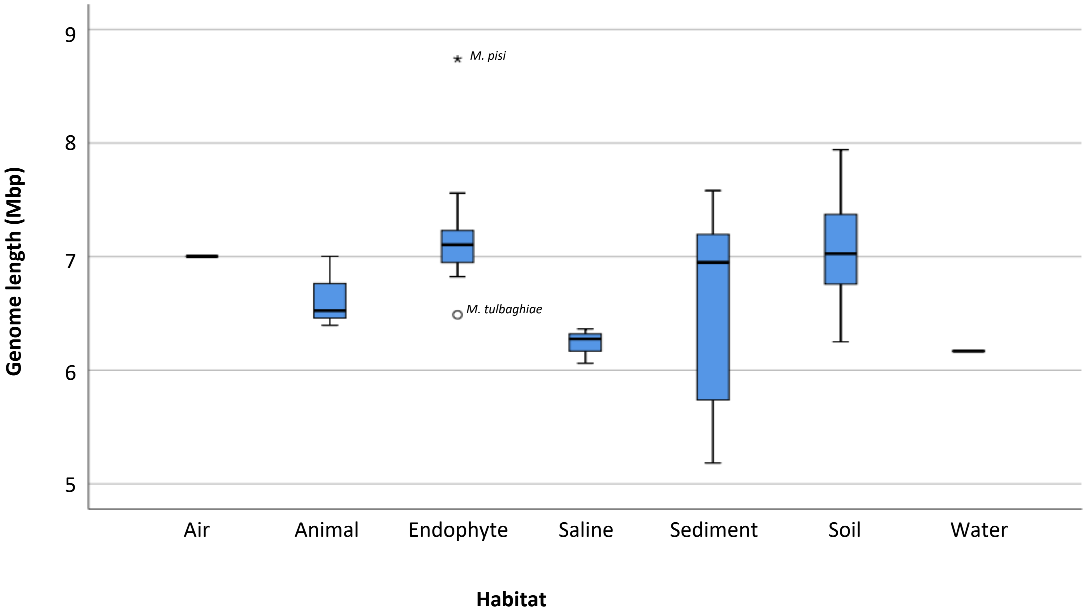


Figure 8: Box plot distribution of the *Micromonospora* strains according to their genome size and habitat. Outliers are specified as circles or asterisks.

Table 5: General genome characteristics of the sequenced genomes. CRISPR number only represent only elements marked as "confirmed CRISPR" in the CRISPRfinder tool (Grissa et al., 2007), all questionable CRISPR have been omitted.

Strain	Genome size (Mb)	G+C ratio (mol%)	CDS	rRNA	tRNA	Contigs	CRISPR	Depth	N50
GAR05	7.1	71.2	6526	6	68	75	0	176	162179
GAR06	7.0	71.2	6410	8	67	60	4	275	180642
M. noduli **GUI43^T**	7.2	70.9	6539	3	57	225	0	784	105196
LAH08	7.3	71.1	6627	4	56	62	3	170	278302
LAH09	6.9	71.6	6182	4	50	133	1	270	102519
Lupac 06	7.1	71.2	6495	4	65	59	1	235	191999
Lupac 07	7.1	71.1	6500	4	51	56	2	216	240674
MED01	7.6	70.8	7060	4	53	63	3	246	232044

MED15	7.2	71.1	6554	4	56	43	1	191	354266
NIE111	6.9	71.1	6298	3	52	71	2	155	232317
NIE79	7.2	71.1	6566	5	84	82	3	155	168635
ONO23	7.2	71.0	6565	5	56	135	2	315	99720
ONO86	7.1	70.9	6591	4	55	407	2	322	29811
PSH03	7.0	71.0	6395	4	54	45	2	342	295666
PSH25	6.8	71.2	6395	5	45	586	2	303	19082
PSN01	6.9	71.1	6458	4	67	456	4	361	26003
PSN13	7.4	71.1	6823	3	64	41	1	227	485329

2.3.3- Functional characterization of the *Micromonospora* database

Analysis of Clusters of Orthologous Groups (COGs) assigned a COG category to 85.0% of the proteins. Main categories were transcription (K, 7.7% ± 0.8), carbohydrate metabolism and transport (G, 5.5% ± 0.6), amino acid metabolism and transport (E, 5.0% ± 0.3), energy production and conversion (C, 4. 3% ± 0.3), and inorganic ion transport and metabolism (P, 4.1% ± 0.3). In addition, 31.7% ± 0.0 of the annotated proteins were assigned to the S category (unknown function). The complete COG distribution is given in Appendix I.

Principal Component Analysis (PCA) of the COG distributions and their relation to the strain habitats were highly influenced by four categories which accounted for 84% of the variance. Namely, these were transcription (K, ~30%), replication and repair (L, ~26%), carbohydrate metabolism and transport (G, ~16%), and secondary metabolism (Q, ~12%). The endophytic strains were recovered as a well-recognized cluster based on high values of K and G, and low Q values. Most of the soil strains were found in the center of the plot, with medium or low values for K, G and C categories, and high values of Q (secondary metabolism). Marine and several soil related strains showed high values for L, and low values of G, T (signal transduction mechanisms) and K, while microorganisms isolated from sediments and other environments were scattered all over the plot, with no clear correlation (Figure 9).

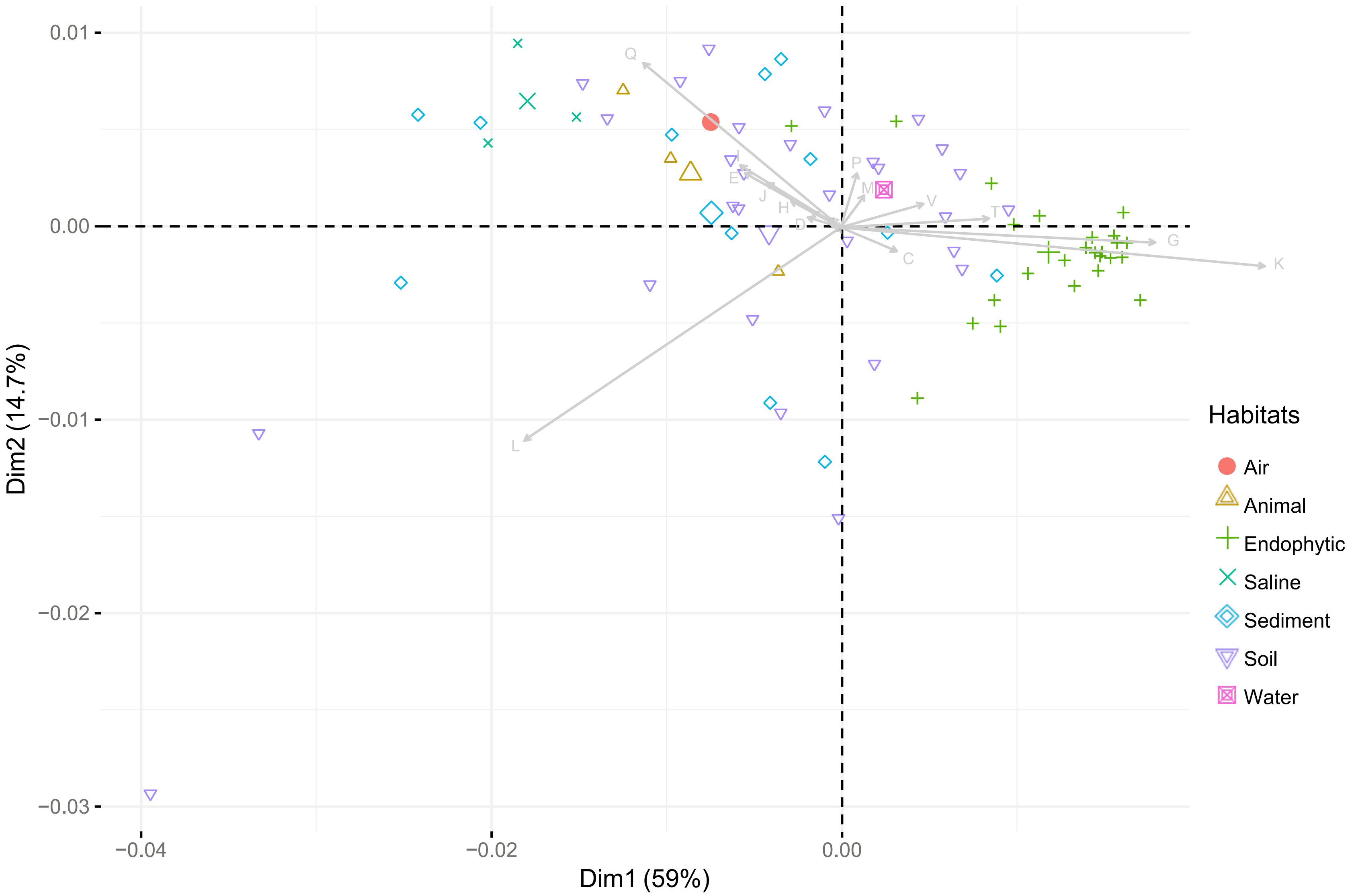


Figure 9: COG distribution of the *Micromonospora* strains and their relation to their habitat.

The number of biosynthetic clusters related to secondary metabolites varied from 6 (*M. globispora* S2901^{T}) to 46 (*M. matsumotoense* DSM 44100^{T}), being the most abundant type terpenes, type I polyketide synthases (t1pks), non-ribosomal peptide synthase (nrps), lanthipeptides, t1pks–nrps, bacteriocines, type II polyketide synthases (t2pks), siderophores and type II polyketide synthases (t3pks) (Figure 10). Secondary metabolite biosynthetic gene cluster analysis revealed no significant correlation between the number of clusters found and the strain habitats, as previously reported (Carro et al., 2018). In general, the endophytic *Micromonospora* strains had less biosynthetic gene clusters (mean of 14), than the soil, sediment and marine isolates (mean of 20, 20 and 23 clusters respectively). The exception was *Micromonospora pisi* DSM 45175^{T}, with 26 biosynthetic clusters identified.

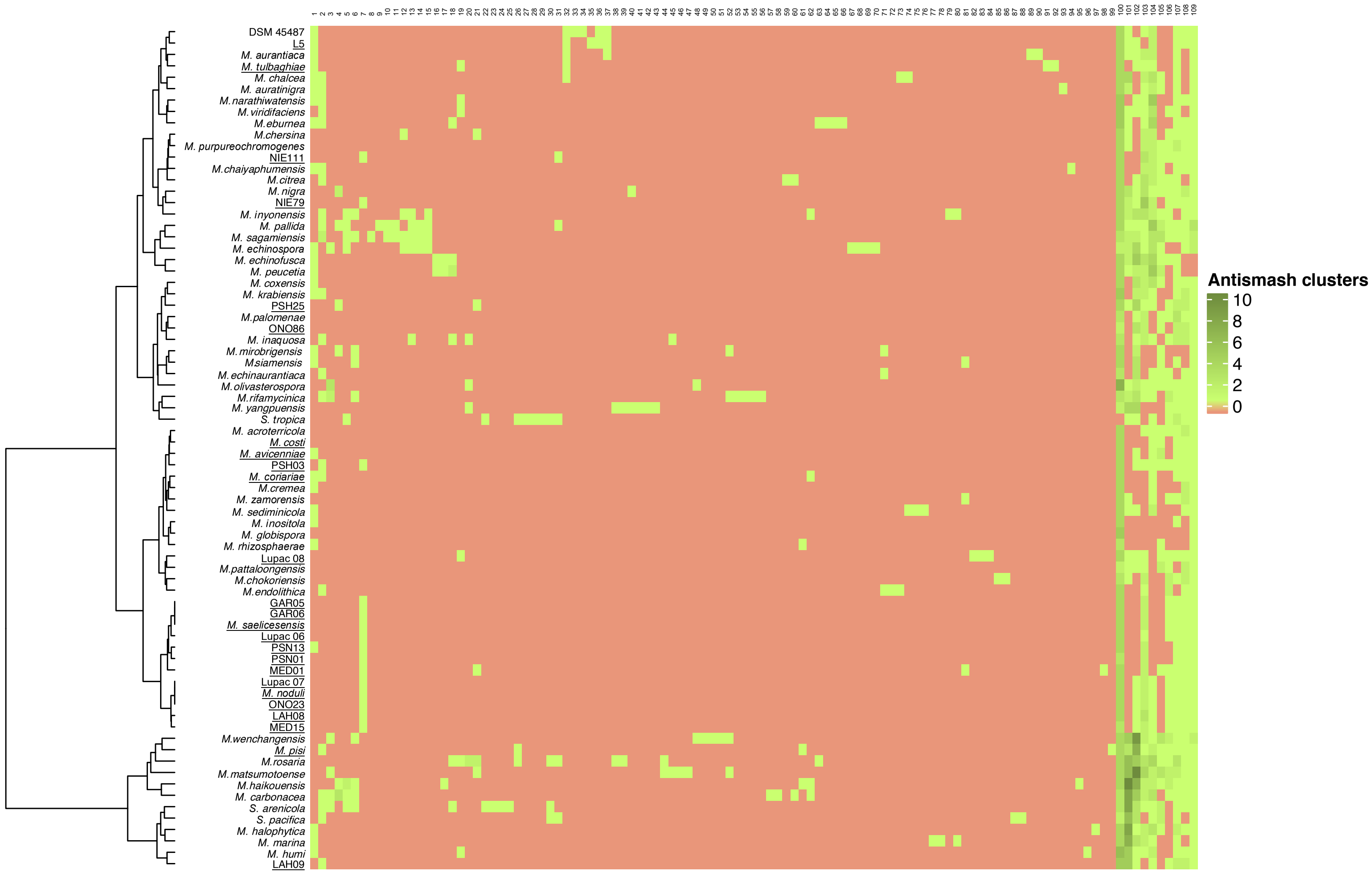


Figure 10: Heatmap of all biosynthetic clusters found in antiSMASH. Endophytic *Micromonospora* strains are shown underlined. Abbreviations according to antiSMASH glossary (https://docs.antismash.secondarymetabolites.org/glossary/): lassopeptide (lasso peptide), amglyccycl (aminoglycoside/aminocyclitol), tXpks (type X polyketide synthase), others (Other types of polyketide synthase), nrps (non-ribosomal peptide synthetase), transatpks (Trans-AT polyketide synthase), fused (phegan-omycin-style protein ligase-containing cluster), blactam (β-lactam), arylpolyene (aryl polyene), head_to_tail (Head-to-tail cyclised (subtilosin-like) cluster), other (Cluster containing a secondary metabolite-related protein that does not fit into any other category), bacteriocin (Bacteriocin or other unspecified ribosomally synthesised and post-translationally modified peptide product (RiPP)).

1	bacteriocin-terpene
2	thiopeptide
3	indole
4	lassopeptide
5	amglyccycl
6	t1pks-otherks
7	t2pks-siderophore
8	nrps-otherks
9	bacteriocin-t1pks-nrps
10	t1pks-butyrolactone
11	t1pks-nucleoside-nrps
12	t3pks-nrps
13	t3pks-t1pks-nrps
14	bacteriocin-lanthipeptide
15	t2pks-t1pks-otherks
16	thiopeptide-t3pks
17	siderophore-t1pks-nrps
18	lanthipeptide-nrps
19	arylpolyene-nrps
20	linaridin
21	otherks
22	bacteriocin-t1pks
23	oligosaccharide-otherks-t1pks-nrps
24	ladderane
25	ladderane-t1pks-arylpolyene-nrps
26	transatpks-t1pks-nrps
27	thiopeptide-terpene
28	t1pks-arylpolyene-nrps
29	terpene-nrps
30	butyrolactone
31	fused
32	oligosaccharide-terpene-nrps
33	otherks-nrps
34	transatpks
35	transatpks-otherks-nrps
36	nrps-t1pks-siderophore-otherks
37	t2pks-oligosaccharide
38	blactam
39	resorcinol
40	lanthipeptide-t1pks-nrps
41	t2pks-terpene
42	indole-phenazine
43	lanthipeptide-terpene
44	phenazine
45	lassopeptide-nrps
46	phosphoglycolipid-t1pks-nucleoside-otherks
47	thiopeptide-nrps
48	phosphoglycolipid-nucleoside
49	nucleoside-otherks
50	bacteriocin-t2pks-nrps
51	linaridin-nrps
52	terpene-oligosaccharide-t1pks-nrps
53	t3pks-otherks-t1pks-nrps
54	t3pks-blactam-t1pks-nrps
55	bacteriocin-t3pks
56	t2pks-butyrolactone
57	t2pks-lanthipeptide-t1pks-nrps
58	terpene-t1pks
59	nrps-t1pks-siderophore
60	oligosaccharide-t1pks-otherks
61	oligosaccharide
62	arylpolyene
63	ectoine
64	otherks-t1pks-ladderane
65	lanthipeptide-t1pks-otherks
66	oligosaccharide-t1pks-otherks-amglyccycl-nrps
67	t2pks-nrps
68	otherks-t1pks-lassopeptide-nrps
69	bacteriocin-t1pks-otherks
70	nucleoside
71	t2pks-arylpolyene
72	linaridin-t1pks-nrps
73	nrps-t1pks-otherks
74	siderophore-nrps
75	t2pks-ladderane
76	thiopeptide-t2pks
77	nrps-siderophore-t1pks-otherks
78	oligosaccharide-terpene-siderophore-nrps-transatpks-otherks
79	t1pks-nucleoside
80	t2pks-t1pks-nrps
81	thiopeptide-t1pks-nrps
82	t2pks-otherks
83	butyrolactone-otherks
84	otherks-t1pks-nrps
85	terpene-t1pks-nrps
86	thiopeptide-t2pks-t1pks
87	oligosaccharide-nrps
88	thiopeptide-siderophore
89	nrps-t1pks-otherks-siderophore
90	transatpks-nrps-otherks
91	nrps-siderophore
92	transatpks-blactam-nrps
93	t2pks-oligosaccharide-otherks
94	butyrolactone-otherks-t1pks-nrps
95	aminocoumarin-t1pks-nucleoside-nrps
96	butyrolactone-lassopeptide
97	t1pks-arylpolyene
98	head_to_tail
99	oligosaccharide-terpene
100	terpene
101	t1pks
102	nrps
103	lanthipeptide
104	t1pks-nrps
105	other
106	bacteriocin
107	t2pks
108	siderophore
109	t3pks

2.4.4- Selection of plant-related *Micromonospora* genes

2.4.4.1- Identification and removal of highly-conserved, non-differential genes in the *Micromonospora* database

Sequence identity distribution analysis of the 92 bacterial core genes described in Na et al., (2018) revealed high stability, being the most stable the *rps*J gene (30S ribosomal protein S10), with a mean identity value of 99%. However, the genes *arg*S (arginine tRNA ligase) and *ile*S (isoleucine tRNA ligase) presented not only a high difference between the minimum and the maximum identity values but high dispersion in the overall values (Figure 11).

Mean and median identity values of the 92 bacterial core genes correlate well in all identity progresion analyses suggesting a simetrical distribution of identies. All *Micromonospora* genomes included in the database had a similarity ≥40% for all the genes in the identity study, with the exception of two genes (*ctg*A and *ctg*X, with minimal values of 21.3%). Considering the overall identity progression, a cut-off value of 70% was selected as threshold for the core genome calculation, as it covered the mean identity of 98.9% of the genes (91/92). The selected threshold was above the minimum value of identity for 87.0% of the genes (80/92) (Figure 12).

With the selected identity threshold, the core genome was calculated to be 992 genes, representing 15.5% of the genome, with an average genome of 6407 genes. The calculated pangenome was 73500 genes (Figure 13). All the genes in the core genome were deleted from the final plant-related gene database to facilitate recognition of the differential genes potentially involved in the *Micromonospora*-plant interaction.

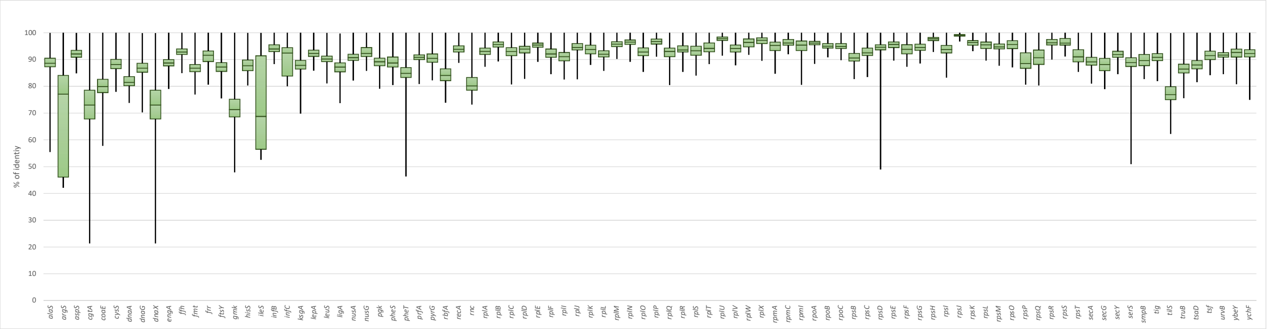


Figure 11: Identity distribution of the 92 bacterial core genes as recommended by Na et al., 2018.

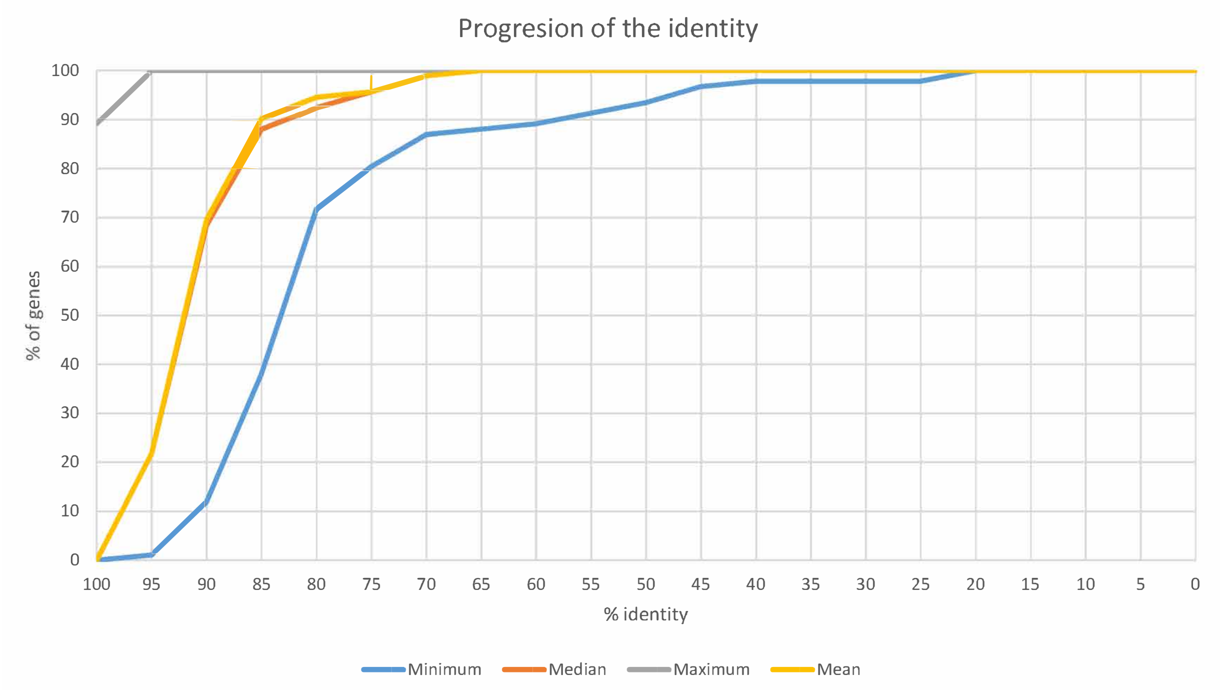


Figure 12: Identity progression of the 92 bacterial core genes. Minimum values (blue), maximum values (grey), mean values (yellow) and median values (orange).

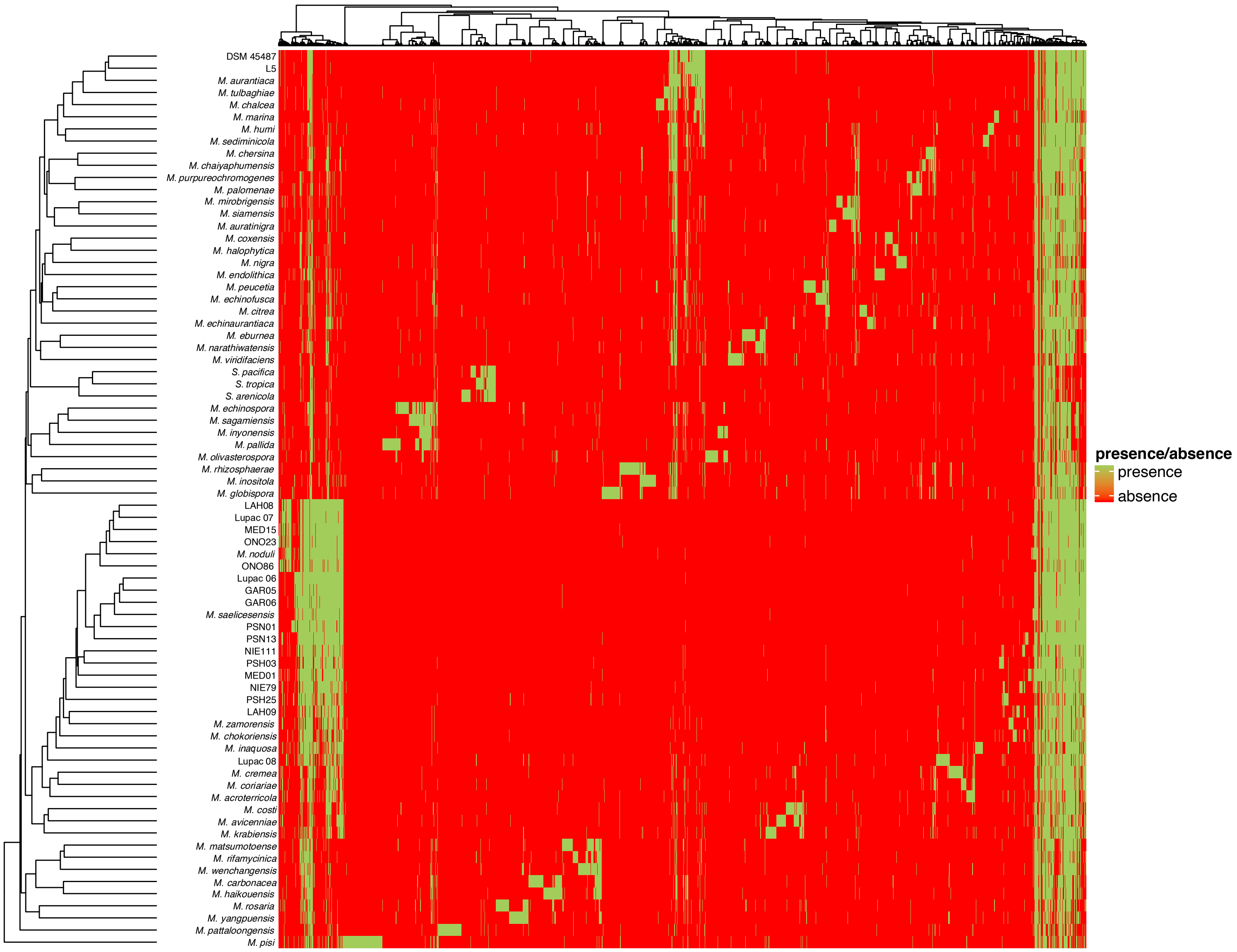


Figure 13: Visual representation of the pangenome. Each column in the heatmap represents a gene in the pangenome. Colored boxes in each column of the heatmap represent the presence (green) or absence (red) of a pangenome gene in a genome (rows).

2.4.4.2- Selection of plant-associated and root-associated genes

Levy et al. (2018) divided their bacterial genome database in several sections, according to their separation in a multilocus phylogenetic tree using 31 universal single-copy genes (Levy et al., 2018). The phylum *Actinobacteria* was divided in two sections (*Actinobacteria*1 and *Actinobacteria*2), however, neither group included any representatives from the genus *Micromonospora.* A recent genome-based taxonomic analysis of the phylum *Actinobacteria* (Nouioui et al., 2018), positioned *Micromonospora* within the *Actinobacteria*1 section database. Consequentially, only elements in *Actinobacteria*1 were selected for the search analysis.

Considering only the features that were supported by two or more statistical analyses, 431 Pfam, 298 KEGG Orthology (KO) and 120 TIGRFAM annotation features were selected for plant-associated (PA) gene screening. In addition, to cover all unannotated or poorly annotated genes, 273 Orthofinder-generated orthologs in GFOBAP *Actinobacteria*1 database were also added to the database. Finally, the database was completed with 86 Pfam, 70 KO and 38 TIGRFAM features to represent root-associated (RA) genes and complemented with 122 Orthofinder-generated orthologs.

Identification of the genes in PA and RA categories were not mutually exclusive, and therefore, these were labeled as unique PA, unique RA or belonging to more than one category (Table 6). Unique PA genes ranged from 675 genes in *M. pisi* DSM 45175^T to 364 in *M. inyonensis* DSM 46123^T, being the endophytes the strains with more PA genes. Unique RA genes ranged from 52 in *M. inositola* DSM 43819^T, to 12 in *M. pattaloongensis* DSM 45245^T, with the soil, rhizospheric soil and endophytes containing the highest number of RA genes. As expected, *Salinispora* strains presented the lowest values for both categories with a range of 312 to 273 for unique PA related genes and a range of 10 to 12 for unique RA genes.

2.4.4.3- Plant-resembling bacterial proteins

Comparative analysis between available host plant proteomes *(Cicer, Glycine, Lupinus, Medicago, Oryza, Phaseolus* and *Trifolium* proteomes in a database of 731325 proteins) and the bacterial genome database revealed a range of 300-550 plant-resembling bacterial proteins coding genes per strain. A high percentage of these genes were part of the calculated core-genome and therefore were deleted from the final database. This deletion mainly affected very conserved genes in all domains of life, such as ribosomal related protein coding genes. Several of the remaining genes also matched with the Plant associated (PA) and Root associated (RA) genes previously identified (section 2.3.4.2). Thus, a range of 167 to 295 unique plant-resembling bacterial proteins coding genes per strain were added to the plant-related gene database (Table 6).

2.4.4.4- Final bacterial plant-related gene database

The information derived from the plant-associated (PA), root-associated (RA) and plant-resembling bacterial proteins (PRPB) were combined in a final database of 69046 putative plant-related genes. Not all the genes were unique for one of the three categories defined (PA, RA or PRPB), therefore elements belonging to two or three of the previously defined categories were distinguished as separate groups to make the final gene count for each genome (Table 6).

The total number of potential plant-related genes varied greatly among the study strains, with *M. pisi* and *S. tropica* CNB-440^T containing the highest and lowest numbers respectively (1137 and 570). Nevertheless, a clear correlation between the environment and the number of genes retrieved in the analysis was observed. Specifically, the endophytic strains had the highest number of genes with a mean of 1036 ± 57 potentially plant-related genes, followed by soil (914 ± 85), sediments (841 ± 172) and saline environment (714 ± 14).

Except for the endophytic strains, it was not possible to make a significant correlation ($p< 0.05$) between the number of plant-related genes found and the environment, due to the high dispersion of the sediment isolates. Interestingly, the soil isolates *M. cremea* DSM 45599^T and *M. carbonacea* DSM 43168^T were recovered as outliers given their high number of potentially plant-related genes identified (1121 and 1100 respectively); these values were above the mean of the endophytic strains. Soil isolate *M. inyonensis* DSM 46123^T was also recovered as an outlier due to the low number of genes identified (704) (Figure 14).

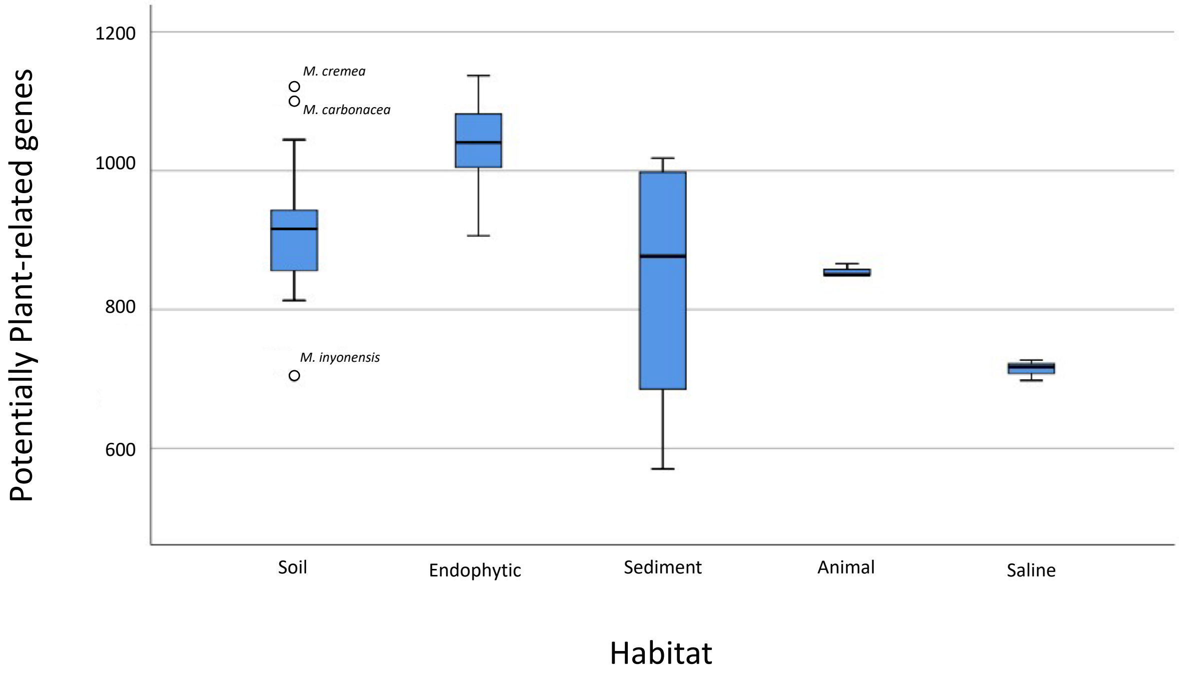


Figure 14: Box plot distribution of all strains according to the number of genes potentially related to plants found in the analysis with respect to their habitat.

Table 6: Number of genes included in the plant-related gene database. All genes in the core genome were deleted from the final database and are not included in the table. Plant associated genes (PA genes); Root associated genes (RA genes); Plant resembling bacterial protein associated genes (PRBP genes).

Strain	PA unique genes	RA unique genes	PRBP unique genes	Genes in two categories	Genes in the three categories	Total genes found
M. acroterricola 5R2A7^T	563	40	186	133	21	**943**
M. aurantiaca DSM 45487	513	23	218	132	23	**909**
M. aurantiaca ATCC 27029^T	505	26	215	131	24	**901**
M. aurantiaca L5	506	25	218	136	25	**910**
M. auratinigra DSM 44815^T	522	21	196	131	14	**884**
M. avicenniae DSM 45758^T	566	31	207	138	27	**969**
M. carbonacea DSM 43168^T	603	14	286	158	39	**1100**
M. chaiyaphumensis DSM 45246^T	526	39	211	127	18	**921**
M. chalcea DSM 43026^T	512	24	239	138	31	**944**
M. chersina DSM 44151^T	538	36	204	141	18	**937**
M. chokoriensis DSM 45160^T	553	34	195	131	22	**935**
M. citrea DSM 43903^T	501	22	187	116	15	**841**
M. coriariae DSM 44875^T	583	37	211	138	25	**994**
M. costi CS1-12^T	595	50	200	135	25	**1005**
M. coxensis DSM 45161^T	517	22	193	118	22	**872**
M. cremea DSM 45599^T	608	51	281	145	36	**1121**
M. eburnea DSM 44814^T	481	22	215	120	16	**854**
M. echinaurantiaca DSM 43094^T	556	33	206	130	18	**943**
M. echinofusca DSM 43913^T	501	16	197	122	14	**850**
M. echinospora DSM 43816^T	595	24	229	132	30	**1010**
M. endolithica DSM 44398^T	535	25	214	125	17	**916**
M. globispora S2901^T	531	33	209	158	35	**966**
M. haikouensis DSM 45626^T	622	19	204	130	22	**997**
M. halophytica DSM 43171^T	394	21	185	108	19	**727**
M. humi DSM 45647^T	546	25	215	140	19	**945**
M. inaquosa LB39^T	615	36	219	147	28	**1045**
M. inositola DSM 43819^T	514	52	205	145	22	**938**
M. inyonensis DSM 46123^T	364	20	198	102	20	**704**
M. krabiensis DSM 45344^T	618	44	201	132	23	**1018**

M. lupini Lupac 08	607	29	284	154	26	**1100**
M. marina DSM 45555^{T}	383	16	186	97	16	**698**
M. matsumotoense DSM 44100^{T}	590	22	218	131	17	**978**
M. mirobrigensis DSM 44830^{T}	507	26	193	138	20	**884**
M. narathiwatensis DSM 45248^{T}	480	24	196	121	20	**841**
M. nigra DSM 43818^{T}	419	15	167	100	16	**717**
M. noduli GUI43^{T}	619	40	243	159	21	**1082**
M. noduli LAH08	634	39	224	151	21	**1069**
M. noduli Lupac 07	632	38	220	146	22	**1058**
M. noduli MED15	636	38	242	155	23	**1094**
M. noduli ONO23	650	38	219	149	21	**1077**
M. noduli ONO86	611	37	209	147	22	**1026**
M. olivasterospora DSM 43868^{T}	424	24	229	112	24	**813**
M. pallida DSM 43817^{T}	482	21	208	130	29	**870**
M. palomenae DSM 102131^{T}	487	31	195	115	21	**849**
M. pattaloongensis DSM 45245^{T}	421	12	172	114	23	**742**
M. peucetia DSM 43363^{T}	484	17	219	120	23	**863**
M. pisi DSM 45175^{T}	675	39	227	170	26	**1137**
M. purpureochromogenes DSM 43821^{T}	480	32	201	120	23	**856**
M. rhizosphaerae DSM 45131^{T}	542	45	236	147	29	**999**
M. rifamycinica DSM 44983^{T}	501	17	189	112	16	**835**
M. rosaria DSM 803^{T}	567	24	208	132	25	**956**
M. saelicesensis DSM 44871^{T}	614	38	235	146	23	**1056**
M. saelicesensis GAR05	625	36	214	140	23	**1038**
M. saelicesensis GAR06	617	38	220	141	25	**1041**
M. saelicesensis Lupac 06	615	38	221	144	24	**1042**
M. saelicesensis PSN01	605	38	205	138	22	**1008**
M. saelicesensis PSN13	656	41	225	152	25	**1099**
M. sagamiensis DSM 43912^{T}	498	22	194	119	23	**856**
M. sediminicola DSM 45794^{T}	529	29	206	130	18	**912**
M. siamensis DSM 45097^{T}	452	25	196	127	16	**816**
Micromonospora sp. LAH09	586	36	201	131	18	**972**
Micromonospora sp. MED01	636	39	239	158	19	**1091**
Micromonospora sp. NIE111	590	33	238	141	20	**1022**
Micromonospora sp. NIE79	634	39	241	153	18	**1085**
Micromonospora sp. PSH03	596	33	211	139	22	**1001**

Micromonospora sp. PSH25	603	36	213	148	18	**1018**
M. tulbaghiae DSM 45142^T	519	23	211	128	25	**906**
M. viridifaciens DSM 43909^T	471	34	215	105	28	**853**
M. wenchangensis CCTCC AA 2012002^T	529	20	295	136	32	**1012**
M. yangpuensis DSM 45577^T	453	14	249	124	26	**866**
M. zamorensis DSM 45600^T	542	35	211	131	12	**931**
S. arenicola CNH-643^T	312	12	194	87	24	**629**
S. pacifica CNR-114^T	274	12	184	91	18	**579**
S. tropica CNB-440^T	273	10	181	88	18	**570**

2.4.5- Functional characterization of the plant-related gene database

2.4.5.1- Clusters of Orthologous Groups (COGs)

Principal Component Analysis of the COGs of the putative plant-related genes that composed the final database (69046) revealed a distribution mainly dependent of the carbohydrate metabolism and transport (G, ~60%), transcription (K, ~20%), secondary metabolism (Q, ~10%) and inorganic ion transport and metabolism (P, ~5%) categories (Figure 15a).

Based on the COG results, the strains were divided in three clusters: the first one comprised twenty-nine strains, being twenty-two of them *Micromonospora* with endophytic origin (76% of the cluster), six soil-related strains (21%) and one sediment isolate (3%). This cluster was characterized with a high percentage of genes included in the transcription (K), carbohydrate metabolism and transport (G) and inorganic ion transport and metabolism (P) categories (Figure 15b). The second cluster contained thirty-four strains, with a more heterogeneous origin. The main habitats in the second cluster were soil, with eighteen isolates (53% of the cluster) and sediments, with seven isolates (20%). The remaining nine strains were of endophytic (3 strains, 9%), animal (3, 8%), marine (1, 3%), air (1, 3%) and fresh water (1, 3%) origin. The second cluster was highly influenced by secondary metabolism (Q) related genes. Finally, the third cluster contained only ten isolates, two marine (20%), and five soil (50%) related *Micromonospora* strains, and the three *Salinispora* representatives (30%), with high values of Q and low values of G, P and K categories (Figure 15b). Individuals in cluster 1, mainly related to endophytic strains, grouped tightly together, suggesting a close relationship in the functional role of the potentially plant-related genes. Clusters 2 and 3 representing most of the soil isolates were loosely recovered, indicating an important degree of heterogeneity in their gene composition and the complexity of their environment. In a similar manner, cluster 3 was not highly correlated and showed a complex relationship.

Approximately 17.8% of the genes included in the plant-related gene database were categorized in the S category (Unknown function).

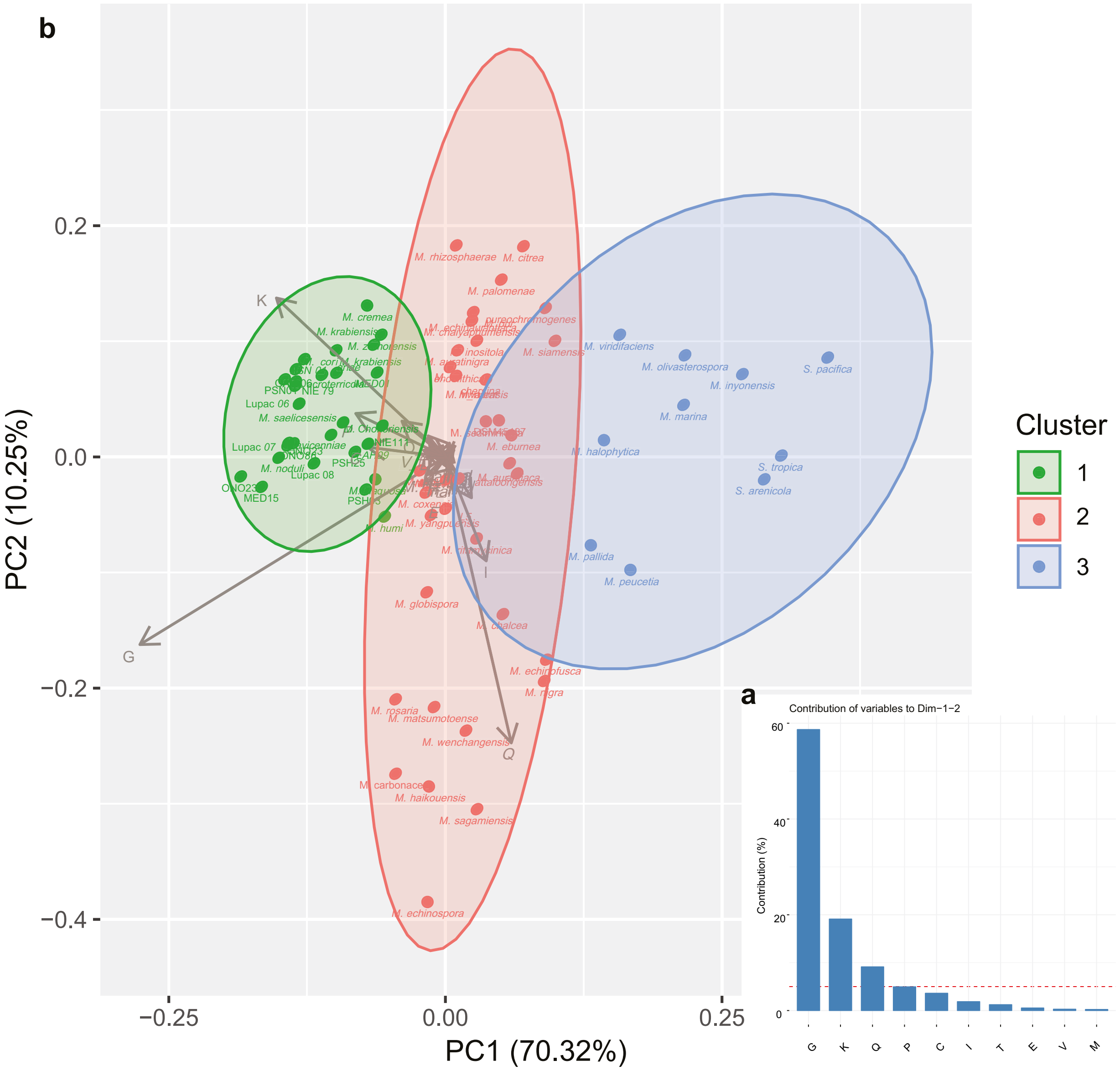


Figure 15: a: Contribution of each variable to the Principal Component Analysis (%). **b**: Principal component analysis of the genes putatively related to plants according to COG categories for each strain. The PCA shows the strains recovered in three clusters: cluster 1, 2 and 3 in green, red and blue respectively.

2.4.5.2- Functional KEGG Ortholog analysis

KEGG Orthology annotations of the putative plant-related genes were also compared, clustered and plotted to determine any differential traits that separated plant-associated *Micromonospora* strains from the other environments (Figure 16). In this case, the PCA analysis based on KEGG data, three clusters were also obtained with almost identical composition as the COG clustering analysis (Section 2.4.5.1). The first cluster was composed of thirty isolates, mainly of endophytic (23 strains, 77 % of the cluster) and soil (5, 17%) origin. The second cluster contained thirty-two strains, mainly soil (21, 66%) and sediment (5, 16%) related strains. Finally, the third cluster included twelve strains, mainly isolated from soil (4, 33%), marine water (3, 25%) and sediments (4, 33%), including the *Salinispora* isolates.

Pearson chi-square test revealed a strong correlation between the clusters 1 and 2 and the isolation source of their strains. Cluster 1 was highly correlated with the endophytic habitat (3.3 x 10^{-9} q value). Cluster 2 was significatively influenced by the soil habitat (3.0 x 10^{-3} q value). Cluster 3 contained all marine isolates (three strains) but did not correlate with any other habitat significantly. Accordingly, from this point the three groups of strains will be referred as cluster 1 (endophytes), cluster 2 (soil) and cluster 3 (mixed). The detailed composition of the three clusters is given in Table 7.

Comparative analysis of the KEGG Orthology annotations revealed significant differences between the three clusters (mean of genes with the annotation varied significantly from the overall mean in all strains (q value < 0.05)). Specifically, 105 significantly over-represented and twenty under-represented KEGG annotations were found in endophytes cluster (cluster 1); twenty-two over- and twenty-four under-represented annotations corresponded to the soil cluster (cluster 2) while only two and sixteen over- and under-represented KEGG features were found for the mixed group (cluster 3). All significantly over-represented and under-represented KEGG annotations identified in the clustering analysis are found in Appendix II.

An interesting result was the fact that strain *M. pattaloongensis* DSM 45245^T contained 75 unique elements and according to their annotation, coded for an almost complete flagellar assembly system (Morimoto and Minamino, 2014) (Figure 17). Hitherto, the presence of a similar flagellar system has not been reported before for *Micromonospora* strains. These genes were not considered in the functional comparative analysis, as they were unique to this strain. Their removal from the database did not affect the position of *M. pattaloongensis* DSM 45245^T as a component of cluster 3.

Table 7: Distribution of KEGG Orthology derived clusters, according to the strain habitat. Endophytes cluster (1, green), soil cluster (2, red) and mixed cluster (3, black).

	Strain	Habitat	Strain	Habitat
Endophytes cluster	GAR05	Endophyte	*M. noduli*	Endophyte
	GAR06	Endophyte	*M. pisi*	Endophyte
	LAH08	Endophyte	*M. rhizosphaerae*	Sediment
	LAH09	Endophyte	*M. saelicesensis*	Endophyte
	Lupac 06	Endophyte	*M. zamorensis*	Soil
	Lupac 07	Endophyte	MED01	Endophyte
	Lupac 08	Endophyte	MED15	Endophyte
	M. acroterricola	Soil	NIE111	Endophyte
	M. avicenniae	Endophyte	NIE79	Endophyte
	M. chokoriensis	Soil	ONO23	Endophyte
	M. coriariae	Endophyte	ONO86	Endophyte
	M. costi	Endophyte	PSH03	Endophyte
	M. cremea	Soil	PSH25	Endophyte
	M. inaquosa	Soil	PSN01	Endophyte
	M. krabiensis	Sediment	PSN13	Endophyte
Soil cluster	DSM45487	Soil	*M. haikouensis*	Sediment
	L5	Endophyte	*M. humi*	Soil
	M. aurantiaca	Soil	*M. inositola*	Soil
	M. auratinigra	Soil	*M. matsumotoense*	Soil
	M. carbonacea	Soil	*M. mirobrigensis*	Water
	M. chaiyaphumensis	Soil	*M. narathiwatensis*	Soil
	M. chalcea	Air	*M. pallida*	Soil
	M. chersina	Soil	*M. peucetia*	Soil
	M. citrea	Sediment	*M. rifamycinica*	Sediment
	M. coxensis	Soil	*M. rosaria*	Soil
	M. eburnea	Soil	*M. sagamiensis*	Soil
	M. echinaurantiaca	Soil	*M. sediminicola*	Sediment
	M. echinofusca	Animal	*M. siamensis*	Soil
	M. echinospora	Soil	*M. tulbaghiae*	Endophyte
	M. endolithica	Soil	*M. viridifaciens*	Soil
	M. globispora	Sediment	*M. yangpuensis*	Animal
Mixed cluster	*M. halophytica*	Saline	*M. pattaloongensis*	Sediment
	M. inyonensis	Soil	*M. purpureochromogenes*	Soil
	M. marina	Saline	*M. wenchangensis*	Soil
	M. nigra	Saline	*S. arenicola*	Sediment
	M. olivasterospora	Soil	*S. pacifica*	Sediment
	M. palomenae	Animal	*S. tropica*	Sediment

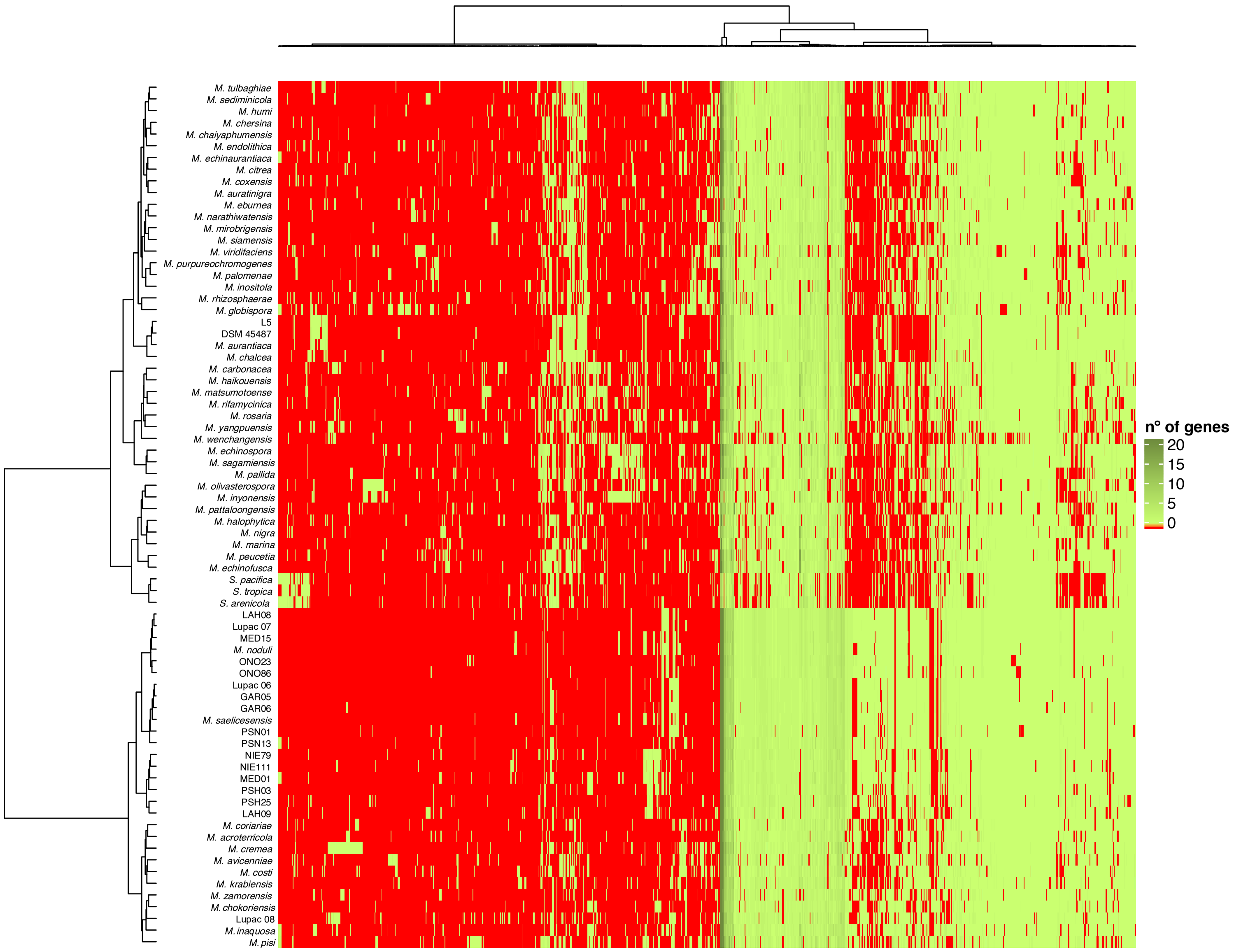


Figure 16: Visual representation of all potentially plant-related KEGG Orthologs (KO) found in each genome. Each column in the heatmap represent a KO. Colored boxes in each column of the heatmap represent the number of KO found (green) or its absence (red) in a genome (rows).

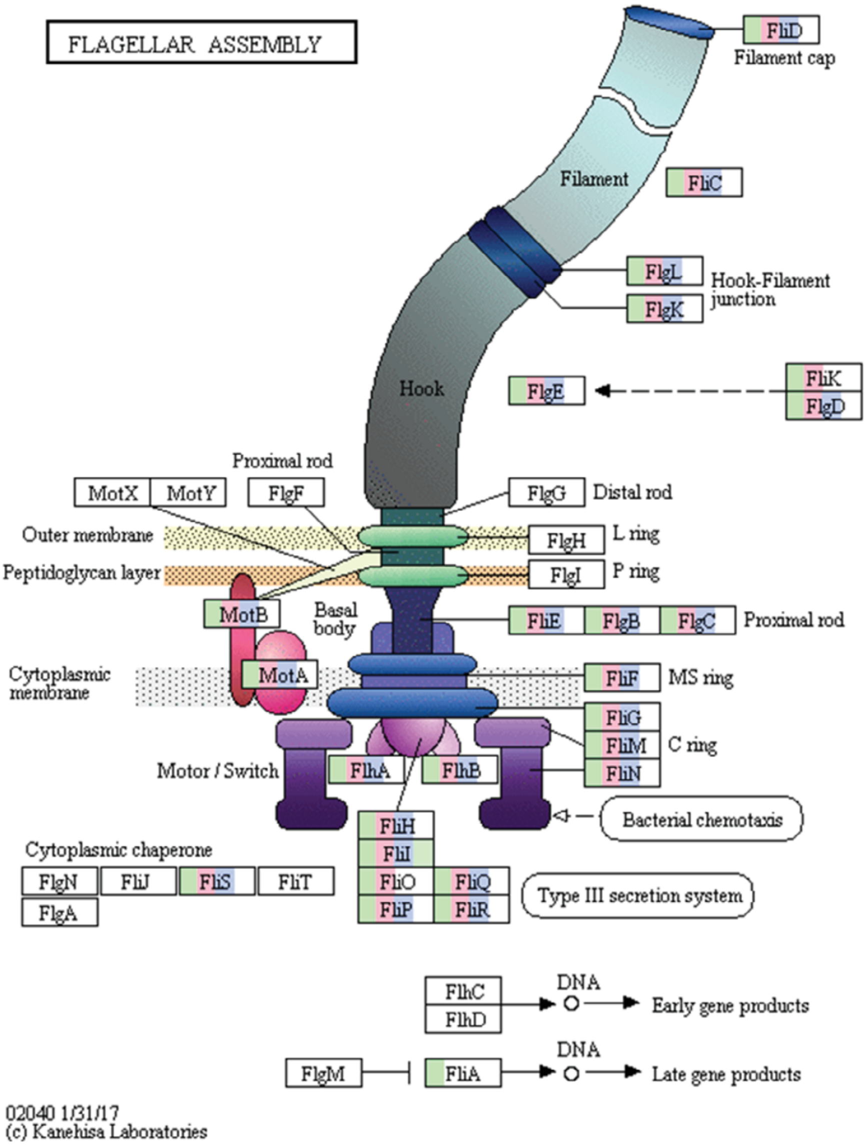


Figure 17: Flagellar assembly KEGG map (02040). In green, general genome annotation; in red, plant-associated genes (PA); in blue, root-associated (RA), in light green, plant-resembling bacterial proteins. The box is not colored if the genes was not found.

2.4.5.2.1- Endophytes Cluster - Main features

The cluster containing the endophyte-associated strains was characterized by a high number of over-represented plant-related genes, many of them involved in use of carbohydrates (Appendix II). It has been reported that plant-related bacterial genomes encode for more carbohydrate metabolic functions than phylogenetically related, but non-plant-associated bacterial genomes (Levy et al., 2018).

Plant-related bacteria evolved to take advantage of sugars usually present in the rhizosphere (raffinose, melibiose, galactose, etc.), reinforcing their transport systems to secure these sugars in the competitive bacterial community of this environment

(Dennis et al., 2010). In this context, several oligosaccharide transporters were found to be over-represented in cluster 1. Multiple sugar transporters coding genes (*msm*X, K, E, F and G), responsible for the transport of sugars such as raffinose, stachyose and melibiose (Tao et al., 2010) were found highly over-represented, in some cases in more than two-fold with respect to the overall mean (*msm*X). In addition, *mal*Z, *sac*A and *gal*A genes, coding for several sugar interconversions from raffinose, sucrose, stachyose, manninotriose and melibiose to glucose, galactose and fructose, were found over-represented in this cluster.

Part of the ribose ABC transport system coding genes (*rbs*A*,* B and C) were found over-represented in cluster 1, with four to five genes per strain. However, the auxiliary component *rbs*D gene was not present in any of the plant-related strains and was found only in seven genomes of the 74 analyzed. *rbs* transporters have been reported to interact with the Autoinducer-2, being a key factor in several quorum sensing mechanisms (Rezzonico et al., 2012; Shimada et al., 2013). *Rbs*D, catalyzes the conversion between the *β* -pyran and *β*-furan forms of *D*-ribose (Shimada et al., 2013) and could be dispensable if used only in the transport of the Autoinducer II and not ribose.

The presence of plant cell-wall degrading enzymes was reported earlier in the genus *Micromonospora* (de Menezes et al., 2012; Trujillo et al., 2014, 2015). *β*- glucosidases (EC: 3.2.1.21) which participate in cellulose degradation systems, hydrolyzing the cellobiose released during the initial hydrolysis of this polymer (Medie et al., 2012), were found in almost six copies in all strains in the endophyte cluster. However, the degradation of plant polymers (celluloses, hemilcelluloses, pectin, etc.) requires multiple intermediate steps for their complete hydrolysis. Thus, in different hemicelluloses such as arabinoxylan, the main *β*-d-(1,4)-linked xylopyrano backbone is substituted with *L*-arabinose residues (Dimarogona and Topakas, 2016), which require the action of specific hydrolytic bacterial enzymes. The arabinofuranohydrolases are enzymes that cleave these *L*-arabinose residues enabling the action of other hydrolytic enzymes such as arabinosyl hydrolases for the decomposition of arabinan, a major pectin polysaccharide (Kim, 2008). Genes coding for *L*-arabonate dehydratase (*ara*C) and arabinoxylan arabinofuranohydrolase (*xyn*D) were found over-represented in the endophyte cluster, with more than a two-fold difference with respect to the overall mean in the whole genome database.

Siderophore production is one of the most frequently plant-growth promoting systems screened in bacteria (Afzal et al., 2019; Crowley, 2006). Interestingly, siderophore secondary metabolic clusters (Figure 9) and genes encoding for transporters associated with siderophores (iron complex transporters coding genes *fhu*B, C and E) were found in almost all *Micromonospora* strains analyzed without a significant difference with respect to the endophytes cluster. Moreover, a ferric transporter coding gene *afu*B was found highly under-represented in the cluster 1 (fold change <0.5).

Vitamins have been reported to be involved in rhizosphere colonization (Babalola, 2010), quorum sensing and cellular signaling (Miret and Munné-Bosch, 2014; Rajamani et al., 2008), gene regulation (Miret and Munné-Bosch, 2014), redox modulation (Mooney et al., 2009) and as cofactors in several biological reactions (Vanderschuren et al., 2013). Complete metabolic pathways for production of thiamine (B1), riboflavin (B2), niacin (B3), pantothenate (B5), pyridoxine (B6), biotin (B7) and folate (B9) were found in almost all *Micromonospora* genomes analyzed. However, only the genes *thi*D, *ilv*D and another one coding for a pyridoxine 4-dehydrogenase (involved in B1, B6 and B5 biogenesis respectively) were found significantly over-represented in the endophyte cluster (q value < 0.05).

Amino-acid transport has been reported to have several roles in the relation between the bacteria and its host. Amino-acids can serve as nitrogen sources for the plant or the bacteria, and even as a regulation factor of the rhizobial bacteroids inside the plant nodule (Prell et al., 2009b). Genes coding for branched-chain amino-acid transporters (*liv*) were found to be over-represented in all strains in the endophyte cluster, presenting a mean of ten genes per genome encoding for the genes *liv*G and *liv*F. Branched chain aminoacid transporters have been reported to have a broad substrate specificity, being able to transport γ-amino-butyric acid (GABA) (Hosie et al., 2002). GABA is an abundant aminoacid in legume nodules, and serves as an additional carbon source for bacteroids in the rhizobium-legume symbiosis (Cooper et al., 2018; Prell et al., 2009a). The transport of GABA to bacterial cells has also been described as one of the quorum sensing factors associated with the control of *Rhizobium radiobacter* growth (previously known as *Agrobacterium tumefaciens*), regulating the colonization of plant cells (Chevrot et al., 2006).

Urate is one of the main end products of rhizobial infected cells in legumes. It is transported to uninfected nodular cells where it is transformed into ureides that are transported in the xylem to the rest of the plant. The N:C ratio of 1:1 confers a major advantage for ureides as N-transport molecules, providing N to plants at a minimal reduced C cost (Baral et al., 2016; Izaguirre-Mayoral et al., 2018). Interestingly, two genes coding for xanthine dehydrogenases (*xdh*G and *yag*T), involved in the metabolization of urates, were highly over-represented in the endophytes cluster, with more than two-fold difference with respect to the overall mean.

2.4.5.2.2- Soil cluster - Main features

The soil cluster (cluster 2) was characterized by an almost equal number of over-represented and under-represented genes, with only a few being over-represented (fold change > 2).

Among the highly over-represented, an acetyl-CoA synthetase and a propionyl-CoA synthetase coding genes were found. These genes are known to participate in multiple metabolic routes, including the degradation of acetate, metabolization of pyruvate and propionate and lipid biosynthesis. Acetate is an essential element in soil, being one of the main carbon sources available (Lanoil and Han, 2006). On the other hand, many genes related to lipid biosynthesis have been found over-represented in cluster 2, including *ACSF2, ACSL* and *fad*D36, coding for fatty acid CoA and acyl CoA synthetases.

The number of under-represented (fold change < 0.5), or even absent genes in the soil cluster strains was significantly high. These genes were related to carbohydrate metabolism and transport, and genetic information processing, as predicted in the COG analysis (Appendix II and Figure 15a and b). Specifically, the absence of *malZ* and *ara*C (coding for an *alpha*-glucosidase and a *L*-arabonate dehydrase) and the low number of *xyn*D (coding for an arabinoxylan arabinofuranohydrolase) are of special importance due to their involvement in the hydrolysis of plant polymers.

Fe acquisition through siderophore production and capture plays an essential role in the colonization of soils and it has been widely studied in rhizospheric bacteria (Crowley, 2006; Martínez-Viveros et al., 2010). Iron acquisition genes *afu*A and B, part of an ABC transport system responsible for the reception of iron (III) (Chin et al., 1996) were found over-represented in the soil cluster.

2.4.5.2.3- Mixed cluster - Main features

The mixed cluster (cluster 3) was characterized by the low number of plant-related features, presenting only eighteen differential features, sixteen of them under-represented (Appendix II). Most of the highly under-represented features (fold < 0.5) were involved in carbon source metabolism and transport (*ara*A, *msm*FG, and several multiple sugar transport permease coding genes). Clearly, these results highly correlate with the origin of the strains.

2.5- DISCUSSION

In the present work, the genomes of seventeen new endophytic *Micromonospora* strains isolated from different legume tissues were sequenced. This data was combined with other publicly available *Micromonospora* sequences to build a database of 74 genomes representing strains from several environments, with endophytic and soil strains, highly represented. The database was used in a comparative analysis to identify the genomic features of adaptation of *Micromonospora* to its host plant.

2.5.1- Features of plant-associated bacterial genomes

Recent data suggests that bacterial adaptation to plants is partially reflected in an increase in genome size as compared to non-plant related bacteria. It is also suggested that these genomes contain an enriched number of genes involved in carbohydrate metabolism and a loss of genomic information related to mobile elements (Levy et al., 2018).

In this study, no significant correlation between genome size and different environments was found. Furthermore, genome size in the two main environments (soil and plant) was very similar (7.1 ± 0.4 Mbp). As expected, the genome sizes of the *Micromonospora* and *Salinispora* representative strains varied greatly, with a mean difference of 1.5 Mb, indicating that while these two microorganisms are often difficult to distinguish phenotypically, important differences can be found at the genomic level. *Salinispora* is a marine obligate bacterium that cannot grow in the absence of NaCl (Mincer et al., 2002) and its reduced genome strongly suggests an adaptation to marine environments. *Micromonospora* on the other hand, appears to have evolved to adapt to multiples niches which could be translated in larger genomes to accommodate different life styles (Trujillo et al., 2014).

Data derived from COG principal component analysis further revealed that *Micromonospora* genomes of plant-associated strains had an increased gene pool involved in the metabolism of carbohydrates (G) and transcription (K) in comparison with the *Micromonospora* isolates recovered from other habitats. On the other hand, marine and soil isolates contained a genome rich in replication, recombination and repair systems (L), and secondary metabolism (Q) functions. Although in some cases, it is not easy to determine the real origin of some of the strains included in this study, clear differences were observed between the genomes of plant-associated and non-plant associated bacteria (Levy et al., 2018).

2.5.2 Construction of a dedicated *Micromonospora* genome database associated to plants

While many *Micromonospora* strains have been isolated from plant tissues, they are not considered obligate endophytes (Carro et al., 2012b; Trujillo et al., 2010). Recent studies singled out several strategies necessary to lead a successful lifestyle as a saprophyte in the rhizosphere, a competitive and harsh environment, and as an endophyte capable of colonizing the internal plant tissues (Afzal et al., 2019; Brader et al., 2017; Compant et al., 2010; Trujillo et al., 2014).

Considering that genomic information must be related to niche adaptation, an analysis based on GWAS (Genome-wide association study) to correlate different bacterial lifestyles (including a plant-associated) with the genomic differences found in the *Micromonospora* genome database, would be a good choice. However, for the correct application of the GWAS method, it is necessary to know before hand, if the *Micromonospora* strains included in our database have a close relationship with the plant. As *Micromonospora* is not an obligate endophyte, it is not always possible to stablish a direct correlation between the bacteria and the plant based only on the isolation origin. Thus, with the available data the GWAS approach cannot be applied with confidence. As such, it was necessary to develop an alternative approach that used genomic data to group strains associated with plants, stablishing the functional differences that separate them from the rest of the strains.

In this work, a new bioinformatic pipeline based on three different stages was developed. First, a search for all genes potentially related to the plant-bacteria interaction was carried out, second, a functional comparative analysis was done and finally a correlation between the results of the functional analysis with the habitats of origin was carried out. A general overview of this pipeline and its comparison with a GWAS approach is presented in Figure 18.

The first step was to use a curated version of the plant-related features database released in 2018 (Levy et al., 2018) as a starting point and added plant-resembling bacterial proteins to identify all potential plant-related genes contained in the genomic database of *Micromonospora.* This data compilation generated a raw database of 69046 potentially plant-related genes. In the second step of this pipeline, all genes with functional annotation (based on COG and KEGG annotations) were included in a comparative analysis which resulted in the distribution of the study strains into three well-defined groups, based on specific functional characteristics. To try to establish a relationship between the groups generated in the functional comparative analysis and a potential habitat, in the last step of the pipeline, these groups were correlated with the original strain isolation habitats.

By following this approach, the group that included all strains sequenced in this work were linked with the endophytic environment (Cluster 1). This cluster was also composed by several strains isolated from rhizospheric (*M. cremea* DSM 45599^T and *M. zamorensis* DSM 45600^T) (Carro et al., 2012a), and desert soils (*M. inaquosa* LB39^T and *M. acroterricola* 5R2A7^T) (Carro et al., 2019a, 2019b). Surprisingly, these four strains shared an important number of functions identified as plant-related functions.

The sequence of steps in the new pipeline follows a completely different order than in a conventional GWAS analysis (Falush and Bowden, 2006; San et al., 2020). A database was not constructed on functions directly related to a phenotype (e.g. habitat). On the contrary, starting from a database of functions previously correlated with bacterial adaptation to the plant, the strains are grouped based on the differential presence of these functional features in the genome (Figure 18). The method developed in this thesis may seem indirect, but it has the advantage of identifying the strains most likely to have a relationship associated with the plant, using only genomic material as the baseline. Under this approach, the information available for the original habitat is only used to characterize the groups generated at the genomic level, but not for the distribution of the strains into their respective groups. It is hoped that in the future, this new pipeline can be used to select new *Micromonospora* strains with agro-biotechnological potential, regardless of their habitat.

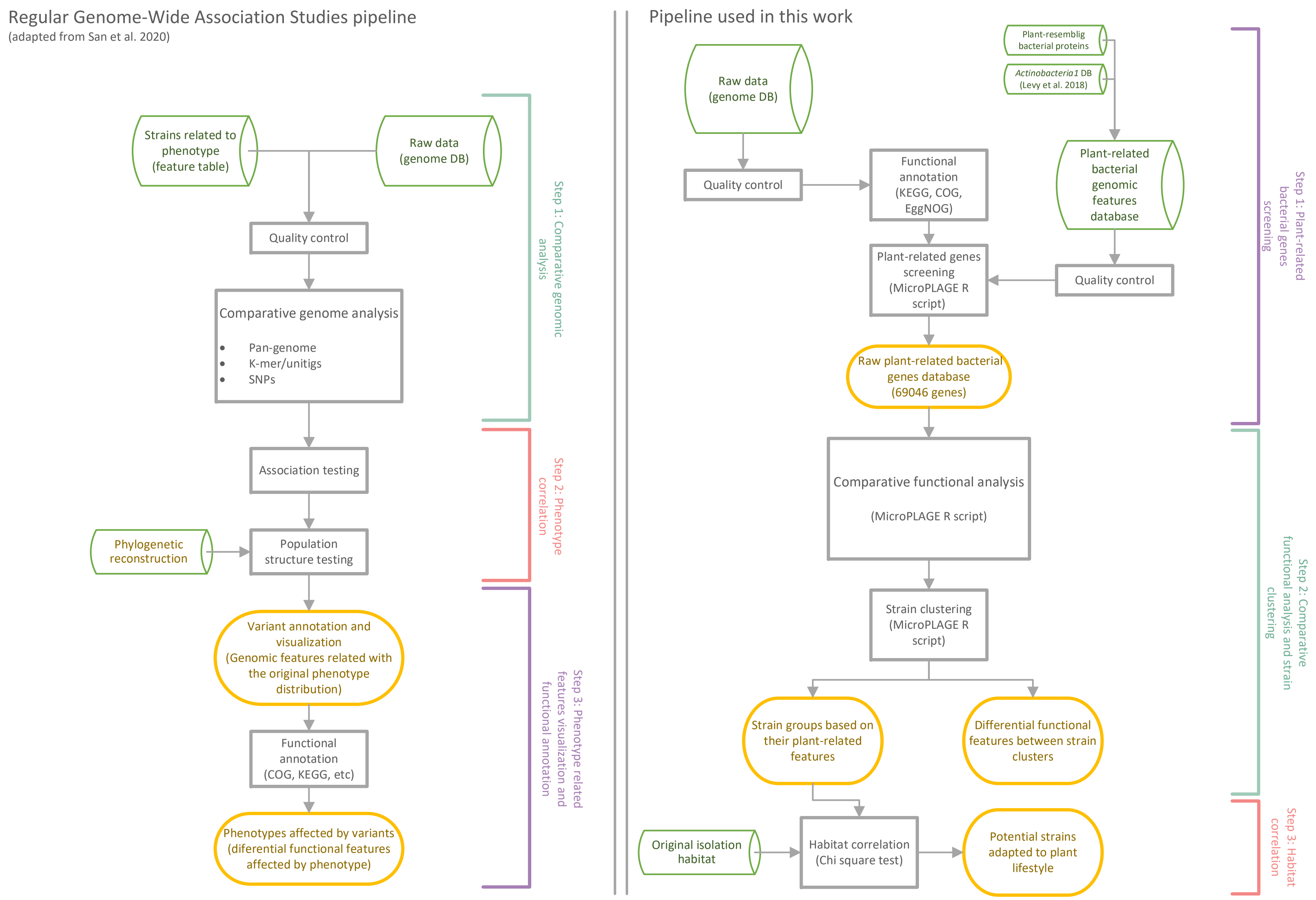

Regular Genome-Wide Association Studies pipeline
(adapted from San et al. 2020)
Strains related to phenotype (feature table)
Raw data (genome DB)
Quality control
Comparative genome analysis
Pan-genome
K-mer/unitigs
SNPs
Association testing
Phylogenetic reconstruction
Population structure testing
Variant annotation and visualization (Genomic features related with the original phenotype distribution)
Functional annotation (COG, KEGG, etc)
Phenotypes affected by variants (diferential functional features affected by phenotype)
Step 1: Comparative genomic analysis
Step 2: Phenotype correlation
Step 3: Phenotype related features visualization and functional annotation
Pipeline used in this work
Raw data (genome DB)
Quality control
Functional annotation (KEGG, COG, EggNOG)
Plant-resemblig bacterial proteins
Actinobacteria1 DB (Levy et al. 2018)
Plant-related bacterial genomic features database
Quality control
Plant-related genes screening (MicroPLAGE R script)
Raw plant-related bacterial genes database (69046 genes)
Comparative functional analysis
(MicroPLAGE R script)
Strain clustering (MicroPLAGE R script)
Strain groups based on their plant-related features
Differential functional features between strain clusters
Original isolation habitat
Habitat correlation (Chi square test)
Potential strains adapted to plant lifestyle
Step 1: Plant-related bacterial genes screening
Step 2: Comparative functional analysis and strain clustering
Step 3: Habitat correlation

Figure 18: Comparation between a genome wide association study workflow (San et al., 2020) and the pipeline used in this work. Green cylinders represent databases and gray squares represent process of the pipeline. Results are outlined in orange.

2.5.3- *Micromonospora* and the host plant: from the rhizosphere to the nodule

The bioinformatic pipeline developed in this work was used to identify *Micromonospora* strains with the highest association to plants, using a comparative functional analysis of 69046 genes. The aim is thus, to single out those genomic functions that appear to be closely related to bacteria-plant interactions and are unique with respect to non-plant related bacteria.

Interestingly, although many of the genomic characteristics commonly related to PGPB bacteria were identified in the initial search for plant-related genes (e.g. siderophore production, phytohormone production, etc.), for the most part, they were not found to be differential in the present analysis. In fact, most of the functional differences were not due to the presence or absence of a specific functional characteristic, but to the overrepresentation of specific systems, increasing the number of genes dedicated to specific functions. Particularly noteworthy was the increase in the number of genes related to the metabolism and carbohydrate transport in the strains most closely associated to an endophyte lifestyle, a result that coincides with data previously described for *Micromonospora* and other plant-associated bacteria (Levy et al., 2018; Trujillo et al., 2014).

The colonization of the rhizosphere, especially by filamentous and non-motile bacteria such as *Micromonospora*, is closely linked to its growth rate, which in turn is related to the ability to capture essential growth elements (Dennis et al., 2010). Growth capacity has been correlated, among other factors, with genes involved in nutrient uptake (de Weert et al., 2006; Garcia-Fraile et al., 2015), with the production and transport of amino acids (Hosie et al., 2002; Prell et al., 2009b; Simons et al., 1997) and the production of vitamin B1 (Simons et al., 1996). These biological processes stood out as differential and were found over-represented in the *Micromonospora* strains most closely related to an endophytic or plant-associated lifestyle (cluster 1).

In the rhizosphere, it is important that the strains are able to use and metabolize with great efficiency the sugars exuded by the plant (raffinose, melibiose, galactose, etc.) in order to compete with other bacteria in the same environment (Dennis et al., 2010). Many of the genes related to the assimilation of these sugars (*mal*Z, *sac*A and *gal*A) and their transport (*msm*EFGKX and *rbs*ABC) were overrepresented in the genomes of plant-associated strains. The regulation of the expression of this set of genes also seems to be important in *Micromonospora*, especially in relation to genes coding for sugar transporters. Expression assays in contact with plant exudates of the strains *Micromonospora cremea* CR30^T, *M. lupini* Lupac 08 and *M. saelicesensis* Lupac 09^T, demonstrated that the *msm* and *rsb* systems, identified as differential in this work, strongly regulated their expression in contact with exudates from the plant (Benito, 2020). Some of these carbohydrate transport coding genes, such as the case of the ABC transporter for ribose (*rbs*ABC), may be useful in the regulation of the bacterial community in the rhizosphere, since they can also transport *quorum sensing* molecules (Polkade et al., 2016;

Rezzonico et al., 2012). Since *Micromonospora* can establish a relationship with rhizobial bacteria both in the rhizosphere and inside plant tissues (Benito et al., 2017; Trujillo et al., 2015), the presence and regulation of possible bacterial *quorum sensing* systems is of great importance.

The presence of genes encoding for hydrolytic enzymes with the capacity to degrade the plant cell wall was previously described in *Micromonospora* (Trujillo et al., 2014, 2015). In this study, overrepresentation of these genes in the bacterial genomes associated with plants was also confirmed. Furthermore, transcriptomic data strongly showed that several genes related to hydrolytic enzymes were strongly regulated when several *Micromonospora* strains were confronted against root exudates (Benito, 2020). In the case of the genes coding for *β*-glucosidases, their expression was increased four to ten-fold in *Micromonospora* strains considered endophytic. Interestingly, the presence of these enzymes does not appear to have a negative effect on the colonization of the plant by the bacterium (Trujillo et al., 2014, 2015).

In *Medicago* and *Trifolium, Micromonospora* has been observed to adhere to the root hairs of the host plant, causing deformation of the roots hairs but without interfering in the infection by rhizobia, suggesting a close relationship between the two bacteria and the host plant (Benito et al., 2017). The presence of plant-cell wall hydrolytic genes is not uncommon in non-pathogenic bacteria and does not necessarily mean that these strains are involved in plant tissue degradation (Mastronunzio et al., 2008; Reinhold-Hurek et al., 2006; Taghavi et al., 2010). In addition, the presence of genes coding for cellulases has been related in some bacteria with the production of biofilms and these molecules are essential for root colonization of rhizobia (Medie et al., 2012; Robledo et al., 2012).

The role of amino acids is essential in the interaction between legumes and *Rhizobium* (Day et al., 2001). These molecules have been described as important factors in the bacteroid-legume relationship and may serve as nitrogen sources for the bacteroid and the plant (Prell et al., 2009b). Amino acids could also act as regulation factors for the growth and proliferation of bacteroids in the plant cell (Delmotte et al., 2010; Prell et al., 2009b). Transporters such as branched-chain amino acid transporters (*liv*), are not only important in the relationship between the plant and the bacteroid after infection (Cooper et al., 2018; Prell et al., 2009a, 2009b), they also appear to involved in plant-bacteria communication prior to colonization, allowing the plant to control the bacterial growth in the root (Chevrot et al., 2006). In this study, branched chain transporters (*liv*GF) were found overrepresented with nearly 10 copies on average, in genomes of strains associated with an endophytic lifestyle. These genes also appear to be strongly regulated in the presence of plant exudates as recently shown with several endophytic *Micromonospora* strains (Benito, 2020). Similar to *Rhizobium*, branched chain transporters may also play a role in colonization. Since the presence of *Micromonospora* increases the efficiency of the nodule (Cerda, 2008; Martínez-Hidalgo et al., 2015; Trujillo et al., 2015), it is possible that LIV transporters are also important for *Micromonospora.* However, further studies are needed to clarify the role of these genes inside plant tissues.

Ureides are one of the most important nitrogen transport molecules in legumes. Along with amino acids (asparagine and arginine), ureides constitute one of the main end products produced in nitrogen fixing nodules (Baral et al., 2016; Izaguirre-Mayoral et al., 2018). These molecules are formed from urates produced in rhizobium-infected cells which are subsequently transported to uninfected cells within the nodule to finally transform into ureides. Ureides are subsequently transported through the xylem to the host plant's leaves (Baral et al., 2016; Izaguirre-Mayoral et al., 2018). Interestingly, genes related to ureide metabolism were also overrepresented in the cluster of plant-associated strains (Cluster 1). It is unclear how *Micromonospora* and ureides may interact, but this could partially explain the presence of *Micromonospora* in plant tissues other than nodules, such as leaves (e.g. strains LAH08, LAH09, PSH03, PSH25).

In this work, several differential genomic characteristics with potential relationship to an endophytic lifestyle were identified. Some features, such as the production of hydrolytic enzymes and an increase in the number of genes related to carbohydrate metabolism had been previously described in *Micromonospora* (Trujillo et al., 2014). However, this work constitutes the first large-scale search for plant-associated genes using 71 *Micromonospora* strains, without focusing exclusively for genes commonly related to PGPB traits.

When compared, many of the genes often identified with plant growth promotion functions (production of siderophores, plant hormones, etc.) did not have any weight as differential characteristics in our database. The present data suggest that adaptation at a genomic level to an endophytic environment is more likely due to the reinforcement of central systems, such as those related to carbohydrate metabolism and transport, and not necessarily due to the presence of unique genes or functional characteristics that could act as a trump card in the relationship with the plant (like the production of an antibiotic or a phytohormone). The reinforced systems provide a competitive advantage in the rhizospheric environment and enhances the possibility to interact both with the plant and other rhizobacteria.

The new pipeline proposed has proved to be an efficient method for the selection of strains with a genomic profile that indicates they have potential to interact with plants. New *Micromonospora* genomes need to be tested and if successful, this method could be very useful for the selection of new candidates for biotechnological applications in agriculture and the environment.

2.6- CONCLUSIONS

1. Genome-size and habitat distribution did not correlate in a significative level in members of the genus *Micromonospora* of our database. However, a significative difference in genome size has been found between the members of the genus *Salinispora* and the members of the genus *Micromonospora,* being the first ones significative smaller.
2. A final database of 69046 potentially plant-related genes has been obtained in this study. Endophyte *Micromonospora* contributed with a mean of 1036 ± 57 potentially plant-related genes per strain to the overall database.
3. The new bioinformatic pipeline has revealed several strains that have the genomic potential to have a close relationship with the plant. More studies must be done in the future to test if these bacteria could colonize the plant as endophytes.
4. Carbon source metabolism and transport systems are highly reinforced in plant-related *Micromonospora.*
5. A list of the most significant functional differences between potentially plant-related *Micromonospora* and the other strains in the database have been inferred. These differences were mostly due to a reinforcement in the genomic content for certain functional systems and not so much for the absence/presence of specific genes.

2.7- REFERENCES

Afzal, I., Shinwari, Z. K., Sikandar, S., and Shahzad, S. (2019). Plant beneficial endophytic bacteria: Mechanisms, diversity, host range and genetic determinants. *Microbiol. Res.* 221, 36–49. doi:10.1016/j.micres.2019.02.001.

Alonso-Vega, P., Normand, P., Bacigalupe, R., Pujic, P., Lajus, A., Vallenet, D., et al. (2012). Genome Sequence of *Micromonospora lupini* Lupac 08, Isolated from Root Nodules of *Lupinus angustifolius. J. Bacteriol.* 194, 4135–4135. doi:10.1128/JB.00628-12.

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* 36, 566. Available at: https://doi.org/10.1038/nbt.4163.

Babalola, O. O. (2010). Beneficial bacteria of agricultural importance. *Biotechnol. Lett.* 32, 1559–1570. doi:10.1007/s10529-010-0347-0.

Bandoy, D. D. R., and Weimer, B. C. (2019). Biological machine learning combined with bacterial population genomics reveals common and rare allelic variants of genes to cause disease. *bioRxiv,* 739540. doi:10.1101/739540.

Baral, B., Teixeira da Silva, J. A., and Izaguirre-Mayoral, M. L. (2016). Early signaling, synthesis, transport and metabolism of ureides. *J. Plant Physiol.* 193, 97–109. doi:10.1016/j.jplph.2016.01.013.

Beasley, T. M., and Schumacker, R. E. (1995). Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures. *J. Exp. Educ.* 64, 79–93. doi:10.1080/00220973.1995.9943797.

Benito, P. (2020). Comparative proteomic and transcriptomic profiling of *Micromonospora* strains associated with legumes.

Benito, P., Alonso-Vega, P., Aguado, C., Luján, R., Anzai, Y., Hirsch, A. M., et al. (2017). Monitoring the colonization and infection of legume nodules by *Micromonospora* in co-inoculation experiments with rhizobia. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-11428-1.

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., et al. (2017). AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* 45, W36–W41. doi:10.1093/nar/gkx319.

Brader, G., Compant, S., Vescio, K., Mitter, B., Trognitz, F., Ma, L.-J., et al. (2017). Ecology and Genomic Insights into Plant-Pathogenic and Plant-Nonpathogenic Endophytes. *Annu. Rev. Phytopathol.* 55, 61–83. doi:10.1146/annurev-phyto-080516-035641.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Carro, L. (2009). Avances en la Sistemática del Género *Micromonospora:* Estudio de Cepas aisladas de la Rizosfera y Nódulos de *Pisum sativum.*

Carro, L., Castro, J. F., Razmilic, V., Nouioui, I., Pan, C., Igual, J. M., et al. (2019a). Uncovering the potential of novel *micromonosporae* isolated from an extreme hyper-arid Atacama Desert soil. *Sci. Rep.* 9, 4678. doi:10.1038/s41598-019-38789-z.

Carro, L., Golinska, P., Nouioui, I., Bull, A. T., Igual, J. M., Andrews, B. A., et al. (2019b). *Micromonospora acroterricola* sp. nov., a novel actinobacterium isolated from a high altitude Atacama Desert soil. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/ijsem.0.003634.

Carro, L., Nouioui, I., Sangal, V., Meier-Kolthoff, J. P., Trujillo, M. E., Montero-Calasanz, M. del C., et al. (2018). Genome-based classification of *micromonosporae* with a focus on their biotechnological and ecological potential. *Sci. Rep.* 8, 525. doi:10.1038/s41598-017-17392-0.

Carro, L., Pujic, P., Trujillo, M. E., and Normand, P. (2013). *Micromonospora* is a normal occupant of actinorhizal nodules. *J. Biosci.* 38, 685–693. doi:10.1007/s12038-013-9359-y.

Carro, L., Pukall, R., Sproër, C., Kroppenstedt, R. M., and Trujillo, M. E. (2012a). *Micromonospora cremea* sp. nov. and *Micromonospora zamorensis* sp. nov., isolated from the rhizosphere of *Pisum sativum. Int. J. Syst. Evol. Microbiol.* 62, 2971–2977. doi:10.1099/ijs.0.038695-0.

Carro, L., Spröer, C., Alonso, P., and Trujillo, M. E. (2012b). Diversity of *Micromonospora* strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. *Syst. Appl. Microbiol.* 35, 73–80. doi:10.1016/j.syapm.2011.11.003.

Cerda, M. E. (2008). Aislamiento de *Micromonospora* de Nódulos de Leguminosas Tropicales y Análisis de Su Interés Como Promotor del Crecimiento Vegetal.

Chevrot, R., Rosen, R., Haudecoeur, E., Cirou, A., Shelp, B. J., Ron, E., et al. (2006). GABA controls the level of *quorum-sensing* signal in *Agrobacterium tumefaciens. Proc. Natl. Acad. Sci. U. S. A.* 103, 7460–7464. doi:10.1073/pnas.0600313103.

Chin, N., Frey, J., Chang, C.-F., and Chang, Y.-F. (1996). Identification of a locus involved in the utilization of iron by *Actinobacillus pleuropneumoniae. FEMS Microbiol. Lett.* 143, 1–6. doi:https://doi.org/10.1016/0378-1097(96)00296-0.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/nar/gkv1276.

Compant, S., Clément, C., and Sessitsch, A. (2010). Plant growth-promoting bacteria in the rhizo- and endosphere of plants: Their role, colonization, mechanisms involved and prospects for utilization. *Soil Biol. Biochem.* 42, 669–678. doi:10.1016/j.soilbio.2009.11.024.

Cooper, B., Campbell, K. B., Beard, H. S., Garrett, W. M., Mowery, J., Bauchan, G. R., et al. (2018). A proteomic network for symbiotic nitrogen fixation efficiency in *Bradyrhizobium elkanii. Mol. Plant-Microbe Interact.* 31, 334–343. doi:10.1094/MPMI-10-17-0243-R.

Crowley, D. E. (2006). "Microbial Siderophores in the Plant Rhizosphere," in *Iron Nutrition in Plants and Rhizospheric Microorganisms*, eds. L. L. Barton and J. Abadia (Dordrecht: Springer Netherlands), 169–198. doi:10.1007/1-4020-4743-6_8.

Day, D. A., Poole, P. S., Tyerman, S. D., and Rosendahl, L. (2001). Ammonia and amino acid transport across symbiotic membranes in nitrogen-fixing legume nodules. *Cell. Mol. Life Sci.* doi:10.1007/PL00000778.

de la Vega, P. A. (2010). Distribución, caracterización e importancia ecologica de *Micromonospora* en nódulos fijadores de nitrogeno de *Lupinus.*

de Menezes, A. B., McDonald, J. E., Allison, H. E., and McCarthy, A. J. (2012). Importance of *Micromonospora* spp. As colonizers of cellulose in freshwater lakes as demonstrated by quantitative reverse transcriptase PCR of 16s rRNA. *Appl. Environ. Microbiol.* 78, 3495–3499. doi:10.1128/AEM.07314-11.

de Weert, S., Dekkers, L. C., Bitter, W., Tuinman, S., Wijfjes, A. H. M., van Boxtel, R., et al. (2006). The two-component colR/S system of *Pseudomonas fluorescens* WCS365 plays a role in rhizosphere competence through maintaining the structure and function of the outer membrane. *FEMS Microbiol. Ecol.* 58, 205–213. doi:10.1111/j.1574-6941.2006.00158.x.

Delmotte, N., Ahrens, C. H., Knief, C., Qeli, E., Koch, M., Fischer, H.-M., et al. (2010). An integrated proteomics and transcriptomics reference data set provides new insights into the *Bradyrhizobium japonicum* bacteroid metabolism in soybean root nodules. *Proteomics* 10, 1391–1400. doi:10.1002/pmic.200900710.

Dennis, P. G., Miller, A. J., and Hirsch, P. R. (2010). Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? *FEMS Microbiol. Ecol.* 72, 313–327. Available at: http://dx.doi.org/10.1111/j.1574-6941.2010.00860.x.

Dimarogona, M., and Topakas, E. (2016). “Regulation and Heterologous Expression of Lignocellulosic Enzymes in *Aspergillus*,” in *New and Future Developments in Microbial Biotechnology and Bioengineering: Aspergillus System Properties and Applications* (Elsevier B.V.), 171–190. doi:10.1016/B978-0-444-63505-1.00012-9.

Dowle, M., and Srinivasan, A. (2018). *data.table:* Extension of `data.frame`. Available at: https://cran.r-project.org/package=data.table.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.

Falush, D., and Bowden, R. (2006). Genome-wide association mapping in bacteria? *Trends Microbiol.* 14, 353–355. doi:10.1016/j.tim.2006.06.003.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. Available at: http://dx.doi.org/10.1093/nar/gkv1344.

Garcia-Fraile, P., Seaman, J. C., Karunakaran, R., Edwards, A., Poole, P. S., and Downie, J. A. (2015). Arabinose and protocatechuate catabolism genes are important for growth of *Rhizobium leguminosarum* biovar viciae in the pea rhizosphere. *Plant Soil* 390, 251–264. doi:10.1007/s11104-015-2389-5.

Garcia, L. C., Martinez-Molina, E., and Trujillo, M. E. (2010). *Micromonospora pisi* sp. nov., isolated from root nodules of *Pisum sativum. Int. J. Syst. Evol. Microbiol.* 60, 331–337. doi:10.1099/ijs.0.012708-0.

Genilloud, O. (2015). *Micromonospora. Bergey's Man. Syst. Archaea Bact.* doi:-doi:10.1002/9781118960608.gbm00148.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, W52–W57. doi:10.1093/nar/gkm360.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.*

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., et al. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–43.

Hosie, A. H. F. F., Allaway, D., Galloway, C. S., Dunsby, H. A., and Poole, P. S. (2002). *Rhizobium leguminosarum* has a second general amino acid permease with unusually broad substrate specificity and high similarity to branched-chain amino acid transporters (*Bra/LIV*) of the ABC family. *J. Bacteriol.* 184, 4071–4080. doi:10.1128/JB.184.15.4071-4080.2002.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi:10.1093/molbev/msx148.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi:10.1093/nar/gkv1248.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.

Izaguirre-Mayoral, M. L., Lazarovits, G., and Baral, B. (2018). Ureide metabolism in plant-associated bacteria: purine plant-bacteria interactive scenarios under nitrogen deficiency. *Plant Soil* 428, 1–34. doi:10.1007/s11104-018-3674-x.

Jafari, M., and Ansari-Pour, N. (2019). Why, when and how to adjust your P values? *Cell J.* 20, 604–607. doi:10.22074/cellj.2019.5992.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi:10.1093/nar/gkr988.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 42, 199–205. doi:10.1093/nar/gkt1076.

Kassambara, A., and Mundt, F. (2017). *factoextra:* Extract and Visualize the Results of Multivariate Data Analyses. Available at: https://cran.r-project.org/package=factoextra.

Kim, T. J. (2008). *Microbial Exo-and Endo-Arabinosyl Hydrolases: Structure, Function, and Application in L-Arabinose Production.* Woodhead Publishing Limited doi:10.1533/9781845695750.2.229.

Kirby, B. M., and Meyers, P. R. (2010). *Micromonospora tulbaghiae* sp. nov., isolated from the leaves of wild garlic, *Tulbaghia violacea*. *Int. J. Syst. Evol. Microbiol.* 60, 1328–1333. doi:10.1099/ijs.0.013243-0.

Lanoil, B. D., and Han, S. (2006). Identification of Microbes Responsible for Acetate Consumption in Soils Under Different Wetting Regimes. *Water,* 1–17. doi:10.1027/2151-2604/a000162.

Lê, S., Josse, J., and Husson, F. (2008). *{FactoMineR}*: A Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi:10.18637/jss.v025.i01.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018). Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50, 138–150. doi:10.1038/s41588-017-0012-9.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). *cluster:* Cluster Analysis Basics and Extensions.

Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi:10.1093/nar/gkr1044.

Martínez-Hidalgo, P., Galindo-Villardón, P., Trujillo, M. E., Igual, J. M., and Martínez-Molina, E. (2015). *Micromonospora* from nitrogen fixing nodules of alfalfa (*Medicago sativa L.*). A new promising Plant Probiotic Bacteria. *Sci. Rep.* 4, 6389. doi:10.1038/srep06389.

Martínez-Viveros, O., Jorquera, M., Crowley, D., Gajardo, G., and Mora, M. (2010). Mechanisms and Practical Considerations Involved in Plant Growth Promotion By *Rhizobacteria. J. soil Sci. plant Nutr.* 10, 293–319. doi:10.4067/S0718-95162010000100006.

Mastronunzio, J. E., Tisa, L. S., Normand, P., and Benson, D. R. (2008). Comparative secretome analysis suggests low plant cell wall degrading capacity in *Frankia* symbionts. *BMC Genomics* 9, 47. doi:10.1186/1471-2164-9-47.

Medie, F. M., Davies, G. J., Drancourt, M., and Henrissat, B. (2012). Genome analyses highlight the different biological roles of cellulases. *Nat. Rev. Microbiol.* 10, 227. Available at: https://doi.org/10.1038/nrmicro2729.

Mincer, T. J., Jensen, P. R., Kauffman, C. A., and Fenical, W. (2002). Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. *Appl. Environ. Microbiol.* 68, 5005 LP – 5011. Available at: http://aem.asm.org/content/68/10/5005.abstract.

Miret, J. A., and Munné-Bosch, S. (2014). Plant amino acid-derived vitamins: biosynthesis and function. *Amino Acids* 46, 809–824. doi:10.1007/s00726-013-1653-3.

Mooney, S., Leuendorf, J. E., Hendrickson, C., and Hellmann, H. (2009). Vitamin B6: A long known compound of surprising complexity. *Molecules* 14, 329–351. doi:10.3390/molecules14010329.

Morimoto, Y., and Minamino, T. (2014). Structure and Function of the Bi-Directional Bacterial Flagellar Motor. *Biomolecules* 4, 217–234. doi:10.3390/biom4010217.

Na, S.-I., Kim, Y. O., Yoon, S.-H., Ha, S., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi:10.1007/s12275-018-8014-6.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* 43, D130–D137. doi:10.1093/nar/gku1063.

Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., et al. (2018). Genome-Based Taxonomic Classification of the Phylum *Actinobacteria. Front. Microbiol.* 9, 2007. doi:10.3389/fmicb.2018.02007.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi:10.1093/nar/gki866.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi:10.1093/bioinformatics/btv421.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. Available at: http://genome.cshlp.org/content/25/7/1043.abstract.

Polkade, A. V., Mantri, S. S., Patwekar, U. J., and Jangid, K. (2016). *Quorum sensing*: An under-explored phenomenon in the phylum *Actinobacteria. Front. Microbiol.* 7, 131. doi:10.3389/fmicb.2016.00131.

Prell, J., Bourdès, A., Karunakaran, R., Lopez-Gomez, M., and Poole, P. (2009a). Pathway of gamma-aminobutyrate metabolism in *Rhizobium leguminosarum* 3841 and its role in symbiosis. *J. Bacteriol.* 191, 2177–2186. doi:10.1128/JB.01714-08.

Prell, J., White, J. P., Bourdes, A., Bunnewell, S., Bongaerts, R. J., and Poole, P. S. (2009b). Legumes regulate *Rhizobium* bacteroid development and persistence by the supply of branched-chain amino acids. *Proc. Natl. Acad. Sci.* 106, 12477 LP – 12482. Available at: http://www.pnas.org/content/106/30/12477.abstract.

R Development Core Team, R., and R Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria doi:10.1007/978-3-540-74686-7.

Rajamani, S., Bauer, W. D., Robinson, J. B., Farrow, J. M., Pesci, E. C., Teplitski, M., et al. (2008). The Vitamin Riboflavin and Its Derivative Lumichrome Activate the *Las*R Bacterial *Quorum-Sensing* Receptor. *Mol. Plant-Microbe Interact.* 21, 1184–1192. doi:10.1094/MPMI-21-9-1184.

Reinhold-Hurek, B., Maes, T., Gemmer, S., Van Montagu, M., and Hurek, T. (2006). An endoglucanase is involved in infection of rice roots by the not-cellulose-metabolizing endophyte *Azoarcus* Sp. strain BH72. *Mol. Plant-Microbe Interact.* 19, 181–188. doi:10.1094/MPMI-19-0181.

Rezzonico, F., Smits, T. H. M., and Duffy, B. (2012). Detection of AI-2 Receptors in Genomes of *Enterobacteriaceae* Suggests a Role of Type-2 Quorum Sensing in Closed Ecosystems. *Sensors* 12. doi:10.3390/s120506645.

Riesco, R., Carro, L., Román-Ponce, B., Prieto, C., Blom, J., Klenk, H. P., et al. (2018). Defining the species *Micromonospora saelicesensis* and *Micromonospora noduli* under the framework of genomics. *Front. Microbiol.* 9, 1360. doi:10.3389/fmicb.2018.01360.

Rigaud, J. R., and Puppo, A. (1975). Indole 3 acetic acid catabolism by soybean bacteroids. *J. Gen. Microbiol.* doi:10.1099/00221287-88-2-223.

Robledo, M., Rivera, L., Jiménez-Zurdo, J. I., Rivas, R., Dazzo, F., Velázquez, E., et al. (2012). Role of *Rhizobium endoglucanase* CelC2 in cellulose biosynthesis and biofilm formation on plant roots and abiotic surfaces. *Microb. Cell Fact.* 11, 125. doi:10.1186/1475-2859-11-125.

Saber, M. M., and Shapiro, B. J. (2020). Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb. genomics* 6. doi:10.1099/mgen.0.000337.

San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., et al. (2020). Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.03119.

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, 686–689. doi:10.1093/nar/gki366.

Shimada, T., Kori, A., and Ishihama, A. (2013). Involvement of the ribose operon repressor *Rbs*R in regulation of purine nucleotide synthesis in *Escherichia coli. FEMS Microbiol. Lett.* 344, 159–165. doi:10.1111/1574-6968.12172.

Shirling, E. B., and Gottlieb, D. (1966). Methods for characterization of *Streptomyces* species. *Int. J. Syst. Bacteriol.* 16, 313–340. doi:10.1099/00207713-16-3-313.

Sillanpää, M. J., and Corander, J. (2002). Model choice in gene mapping: what and why. *Trends Genet.* 18, 301–307. doi:10.1016/S0168-9525(02)02688-4.

Simons, M., Permentier, H., De Weger, L., Wijffelman, C., and Lugtenberg, B. (1997). Amino Acid Synthesis Is Necessary for Tomato Root Colonization by *Pseudomonas fluorescens* Strain WCS365. *Mol. Plant-microbe Interact.* 10, 102–106. doi:10.1094/MPMI.1997.10.1.102.

Simons, M., van der Bij, A. J., Brand, I., de Weger, L. A., Wijffelman, C. A., and Lugtenberg, B. J. (1996). Gnotobiotic system for studying rhizosphere colonization by plant growth-promoting *Pseudomonas* bacteria. *Mol. Plant. Microbe. Interact.* 9 7, 600–607. Available at: http://europepmc.org/abstract/MED/8810075.

Sutton, D., Livingstone, P. G., Furness, E., Swain, M. T., and Whitworth, D. E. (2019). Genome-Wide Identification of Myxobacterial Predation Genes and Demonstration of Formaldehyde Secretion as a Potentially Predation-Resistant Trait of Pseudomonas aeruginosa. *Front. Microbiol.* 10, 1–9. doi:10.3389/fmicb.2019.02650.

Taghavi, S., van der Lelie, D., Hoffman, A., Zhang, Y.-B., Walla, M. D., Vangronsveld, J., et al. (2010). Genome Sequence of the Plant Growth Promoting Endophytic Bacterium *Enterobacter* sp. 638. *PLOS Genet.* 6, e1000943. Available at: https://doi.org/10.1371/journal.pgen.1000943.

Tang, Y., Horikoshi, M., and Li, W. (2016). *ggfortify:* Unified Interface to Visualize Statistical Results of Popular R Packages. *R J. 8.2* 8, 478–489. doi:10.1016/j.cattod.2005.03.010.

Tao, L., Sutcliffe, I. C., Russell, R. R. B., and Ferretti, J. J. (2010). Transport of Sugars, Including Sucrose, by the *msm* Transport System of *Streptococcus mutans. J. Dent. Res.* 72, 1386–1390. doi:10.1177/00220345930720100701.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. Available at: http://dx.doi.org/10.1093/nar/gkw1099.

Thuleau, S., and Husson, F. (2018). *FactoInvestigate:* Automatic Description of Factorial Analysis. Available at: https://cran.r-project.org/package=FactoInvestigate.

Trujillo, M. E., Alonso-Vega, P., Rodríguez, R., Carro, L., Cerda, E., Alonso, P., et al. (2010). The genus *Micromonospora* is widespread in legume root nodules: The example of *Lupinus angustifolius. ISME J.* 4, 1265–1281. doi:10.1038/ismej.2010.55.

Trujillo, M. E., Bacigalupe, R., Pujic, P., Igarashi, Y., Benito, P., Riesco, R., et al. (2014). Genome Features of the Endophytic Actinobacterium *Micromonospora lupini* Strain Lupac 08: On the Process of Adaptation to an Endophytic Life Style? *PLoS One* 9, e108522. doi:10.1371/journal.pone.0108522.

Trujillo, M. E., Kroppenstedt, R. M., Fernández-Molinero, C., Schumann, P., and Martínez-Molina, E. (2007). *Micromonospora lupini* sp. nov. and *Micromonospora saelicesensis* sp. nov., isolated from root nodules of Lupinus angustifolius. *Int. J. Syst. Evol. Microbiol.* 57, 2799–2804. doi:10.1099/ijs.0.65192-0.

Trujillo, M. E., Riesco, R., Benito, P., and Carro, L. (2015). Endophytic actinobacteria and the interaction of *Micromonospora* and nitrogen fixing plants. *Front. Microbiol.* 6, 1–15. doi:10.3389/fmicb.2015.01341.

Valdés, M., Pérez, N.-O., Estrada-de Los Santos, P., Caballero-Mellado, J., Peña-Cabriales, J. J., Normand, P., et al. (2005). Non-*Frankia* actinomycetes isolated from surface-sterilized roots of *Casuarina equisetifolia* fix nitrogen. *Appl. Environ. Microbiol.* 71, 460–466. doi:10.1128/AEM.71.1.460-466.2005.

Vanderschuren, H., Boycheva, S., Li, K.-T., Szydlowski, N., Gruissem, W., and Fitzpatrick, T. (2013). Strategies for vitamin B6 biofortification of plants: a dual role as a micronutrient and a stress protectant. *Front. Plant Sci.* 4, 143. Available at: https://www.frontiersin.org/article/10.3389/fpls.2013.00143.

Veyisoglu, A., Carro, L., Cetin, D., Guven, K., Spröer, C., Pötter, G., et al. (2016). *Micromonospora profundi* sp. nov., isolated from deep marine sediment. *Int. J. Syst. Evol. Microbiol.* 66, 4735–4743. doi:10.1099/ijsem.0.001419.

Vincent, J. M. (1970). "The cultivation, isolation and maintenance of rhizobia," in *A Manual for the Practical Study of Root-nodule Bacteria* (Oxford: Blackwell Scientific Publications), 1–13.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York Available at: http://ggplot2.org.

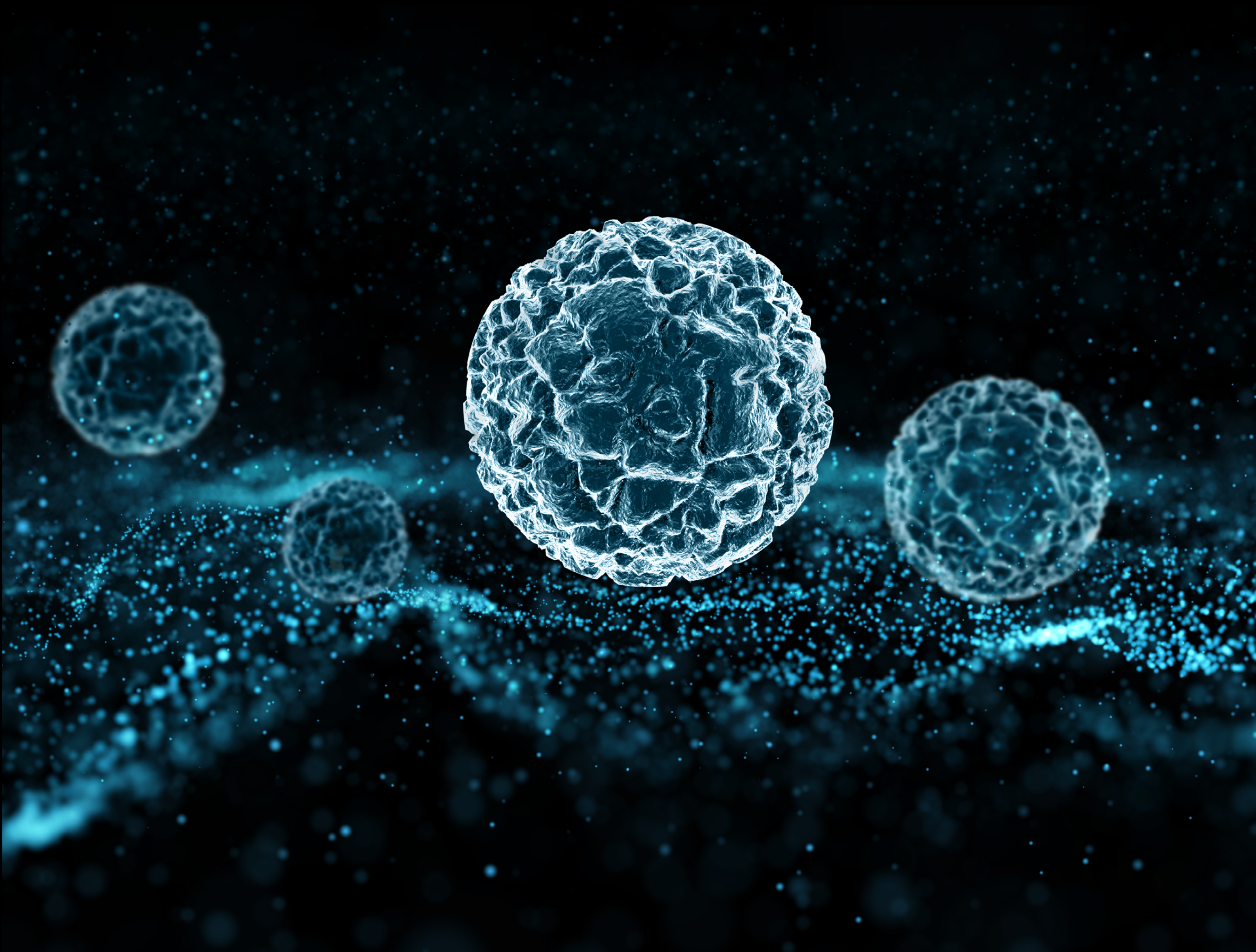
Wickham, H., and Henry, L. (2018). *tidyr:* Easily Tidy Data with "spread()" and "gather()" Functions. Available at: https://cran.r-project.org/package=tidyr.

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017). Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* 67, 1613–1617. doi:10.1099/ijsem.0.001755.

Zhang, L., Xi, L., Ruan, J., and Huang, Y. (2012). *Micromonospora yangpuensis* sp. nov., isolated from a sponge. *Int. J. Syst. Evol. Microbiol.* 62, 272–278. doi:10.1099/ijs.0.029439-0.

CHAPTER III

Defining the species *Micromonospora saelicesensis* and *Micromonospora noduli* under the framework of genomics.



3.1- INTRODUCTION

To define a species, current prokaryotic taxonomy relies on the polyphasic approach, integrating multiple aspects of a microorganism including phenotypic, chemotaxonomic and genomic data (Chun and Rainey, 2014; Colwell, 1970; Vandamme et al., 1996). This approach has contributed for decades to improve classification and identification schemes. However, its limitations and pitfalls, particularly in relation to reproducibility of some methods and the difficulty of data storage, have been addressed multiple times (Sutcliffe et al., 2012; Thompson et al., 2015; Vandamme and Peeters, 2014).

The introduction of cost-effective whole genome sequencing and the improvement of genome-based methodologies provide a new working framework for species delineation. Unlike DNA-DNA hybridization, that was considered the “golden standard” for delineating genomic species in 1987 (Wayne et al., 1987), genomic data can be stored and made available to the scientific community. Furthermore, bioinformatic results are also highly reproducible (Chun and Rainey, 2014). Genomic data can also be used to predict phenotypic profiles that can be then tested in the laboratory, reducing the need to perform labor-intensive and non-reproducible tests (Amaral et al., 2014; Sutcliffe et al., 2012). It is even possible to reconstruct the complete metabolic profile of a microorganism using genomes, expanding the physiological characterization beyond the conventional wet-lab analysis (Durot et al., 2009; Mendoza et al., 2019).

Another problem is the definition of species based on single-strain representatives. This approach does not allow the recognition of intra-species diversity and limits the proposal for a sound and testable definition of a prokaryotic species (Oren and Garrity, 2014; Sutcliffe et al., 2012). Unfortunately, for most species, only one strain is described (the type strain) or one genome is deposited in the repositories, hindering the possibility to study intra-species variation.

The genus *Micromonospora* represented by Gram-stain positive, filamentous and sporulating actinobacteria, belongs to the family *Micromonosporaceae* of the order *Micromonosporales* in the phylum *Actinobacteria* (Genilloud, 2015a, 2015b, 2015c). As reviewed in the previous chapter of this thesis, the genus *Micromonospora* has been isolated from diverse ecosystems such as soil (Lee and Whang, 2017; Thawai et al., 2016), aquatic habitats (de Menezes et al., 2012; Trujillo et al., 2005), plant tissues (Carro et al., 2013, 2016; Trujillo et al., 2007) and other environments (Veyisoglu et al., 2016; Zhang et al., 2012). Recently, a revised classification of the genus *Micromonospora* based on genome sequence data has been proposed (Carro et al., 2018).

Micromonospora saelicesensis was formally described in 2007, with three strains isolated from internal nodule tissues of *Lupinus angustifolius*, Lupac 06, Lupac 07 and Lupac 09^T (Trujillo et al., 2007). This species has also been found in other leguminous and actinorhizal plants including *Medicago* sp. (Martínez-Hidalgo et al., 2014), *Pisum* sp. (Carro, 2009; Carro et al., 2012), *Alnus sp., Coriaria myrtifolia and Morelia pensylvanica* (Carro et al., 2013). The species *M. saelicesensis* is the most frequently isolated species in legumes and actinorhizal plants (Carro, 2009; Carro et al., 2012; Cerda, 2008; de la Vega, 2010; Trujillo et al., 2010, 2015).

Micromonospora noduli described with a single representative, strain GUI43^{T}, was isolated from the nodular tissue of *Pisum sativum*. It was found to be closely related to the species *M. saelicesensis,* with a DNA-DNA hybridization (DDH) value of 63.4% (62.3% reciprocal) (Carro et al., 2016). While the DDH value is below the proposed threshold for species delineation of 70% (Wayne et al., 1987), these two species share many features, and the question arises whether they should be merged into a single species.

In our laboratory, we maintain more than 2000 *Micromonospora* strains isolated from legume nodules, with 13% of them having an almost complete 16S rRNA gene sequence. Thirty-three percent of the strains sequenced have been found to be closely related to the species *Micromonospora noduli and Micromonospora saelicesensis* (with 17% and 16% of the isolates respectively). As the most frequently isolated *Micromonospora*, these two species probably play an important role in the relationship with their legume host.

3.2- OBJECTIVES

The present study was designed to study the taxonomic relationship between the species *Micromonospora saelicesensis* and *Micromonospora noduli* using genome-based data to determine if they should be considered the same species. To achieve this, several specific aims were established:

1. To characterize the species *Micromonospora saelicesensis* and *Micromonospora noduli*, using a genomic approach based on a combination of overall genomic relatedness indices (OGRI) and phylogenomic analysis.
2. To analyze and compare different genomic approaches for the correct identification of *Micromonospora* species based on a genomic framework using the latest accepted classification of the genus.
3. To infer phenotypic profiles based on the genomic information and to confirm these with physiological and biochemical data obtained in the laboratory.

3.3- MATERIALS AND METHODS

3.3.1- Selection of strains

Sixteen strains closely related to *Micromonospora saelicesensis* and *Micromonospora noduli*, isolated from nitrogen fixing nodules of six different legumes were selected. The list of all isolates is given in chapter II, with details of their isolation source and their relatedness with both *M. saelicesensis* and *M. noduli* according to the 16S rRNA gene (Table 2, Chapter II). Genomic DNA of the selected strains was extracted, sequenced and annotated as specified previously (Section 2.3.4, Chapter II).

3.2.2- 16S rRNA gene analysis

DNA extraction using "REDExtract-N-Amp Plant PCR Kit" (Sigma™), 16S rRNA gene amplification and sequencing of the strains MED01 and MED15 were carried out as previously explained (Section 2.3.2, Chapter II). 16S rRNA gene sequencing of strains GAR05, GAR06, LAH08, LAH09, Lupac 06, Lupac 07, NIE79, NIE111, ONO23, ONO86, PSH03, PSH25, PSN01 and PSN13 was previously carried out in the laboratory (Benito, 2020; Cerda, 2008; Trujillo et al., 2007).

16S rRNA gene sequences were compared against the EzBiocloud Database (Yoon et al., 2017) and other public platforms (Genbank, EMBL, etc.). A database was constructed with eighty-one 16S rRNA gene sequences of *Micromonospora* type strains downloaded from the GenBank database (Clark et al., 2016) and the sixteen study strains. *Catellatospora citrea* IMSNU 22008^T, a member of the family *Micromonosporaceae* was selected as an outgroup for the 16S rRNA gene phylogeny, and therefore its 16S rRNA coding gene was added to the database. The final database contained 98 16S rRNA sequences.

All sequences included in the database were aligned with ClustalX v. 2.0 (Thompson et al., 1997). Phylogenetic analyses were performed using MEGA (v 7.0.14) (Kumar et al., 2016). Distances were calculated with the Kimura 2-parameter (Kimura, 1980) and tree topologies were based on the Maximum Likelihood algorithm (Felsenstein, 1981). Total analysis included 1319 positions and a bootstrap sampling of 1000 (Felsenstein, 1985).

3.3.3- *gyr*B gene phylogeny

Complete *Micromonospora* β subunit of DNA gyrase coding gene (*gyr*B) sequences were extracted from whole genome sequence data when available. *gyr*B gene sequences were screened in the genomes using HMMER (hmmer.org), and a precomputed HMM profile. The following script was used for batch gene search in a Linux environment, using Prodigal v. 2.6.1 (Hyatt et al., 2010), HMMER v. 3.2.1 and HMMER associated toolkit *easel* (hmmer.org). E value for the search was stablished in 1 x10^{-150}.

```
#!/bin/bash.
for filename in ./fasta/*.fasta
do
prodigal -i “$filename” -t hmm/Micromonospora.trn -d “./$(basename “$filename” .fasta).fasta”;
done
mkdir output
for fasta in ./*.fasta
do
	hmmsearch -E 1e-150 --tblout myhits.tbl hmm/gyrb.hmm “$fasta”
	esl-sfetch --index “$fasta”
	grep -v “^#” myhits.tbl | awk NR==1’{print $1}’ | esl-sfetch -f “$fasta” - > out_gyr.fasta
	sed ‘1 d’ out_gyr.fasta > out_gyr2.fasta
	sed -i “1i >”$(basename “$fasta” .fasta)”_gyrb” out_gyr2.fasta
	mv “out_gyr2.fasta” “”$(basename “$fasta” .fasta)”_gyrb.fasta”
	rm -r out_gyr*.fasta *.ssi *.tbl
done
mkdir output/gyrb
cat *_gyrb.fasta> output/gyrb/secuencias_gyrb.fa
rm *_gyrb.fasta
```

*gyr*B complete gene sequences of the sixteen isolates and forty-three *Micromonospora* type strains were screened from whole genomic sequences and included in a database. When whole genome sequences were not available, partial *gyr*B gene sequences were directly downloaded from GenBank nucleotide database (Clark et al., 2016). A total of thirty-three partial sequences were included in the *gyr*B gene database. *Catellatospora citrea* DSM 44097^T was selected as an outgroup for *gyr*B phylogeny, and therefore its *gyr*B coding gene was screened from its genome and added to the final database. The final *gyr*B gene database contained 93 sequences.

All sequences included in the *gyr*B database were aligned with ClustalX v. 2.0 and used to construct a Maximum-Likelihood phylogenetic tree based on Kimura 2-parameter, using 1001 nucleotide positions and a bootstrap value of 1000 using MEGA (v 7.0.14) (Kumar et al., 2016).

3.3.4- Multi-Locus Sequence Analysis (MLSA)

Sequences of the β-subunit of RNA polymerase (*rpo*B), ATP synthase β-subunit (*atp*D) and recombination protein A (*rec*A) coding genes were retrieved from the genomes of the sixteen isolates and forty-seven *Micromonospora* genomes, using HMMER v. 3.2.1 and a precomputed HMM profile as previously explained for *gyr*B gene. In addition, partial gene sequences for eight *Micromonospora* type strains were directly retrieved from GenBank database. The final MLSA database contained sequences for 72 strains, including *Catellatospora citrea* DSM 44097^T that was selected as an outgroup for the phylogenetic analysis.

All sequences were aligned using ClustalX v2.0 and then trimmed in the flanking regions of the alignment to the first and last position that covers all sequences using Gblocks v0.91b (Talavera and Castresana, 2007). Sequences were concatenated using Geneious® v. 10 in the following order 16S rRNA-*gyr*B-*rpo*B-*atp*D-*rec*A. Maximum-Likelihood phylogenetic tree based on Kimura 2-parameter model was inferred using MEGA (v 7.0.14) (Kumar et al., 2016) with a bootstrap value of 1000. A total of 4340 nucleotide positions were used for the tree construction.

3.3.5- OGRI analysis and core genome analysis

To delimit the closest strains to the species *Micromonospora saelicesensis and Micromonospora noduli*, four different overall genomic relatedness indices (OGRI) were used: BLAST based Average Nucleotide Identity (ANI_b) (Goris et al., 2007), OrthoANI (Lee et al., 2016), Digital DNA-DNA hybridization (dDDH) (Meier-Kolthoff et al., 2013) and G+C content differences (Meier-Kolthoff et al., 2014b). The genomes of the sixteen study strains (Section 3.3.1, Chapter III) were compared against the genomes of *M. saelicesensis* DSM 44871^T (Carro et al., 2018) and *M. noduli* GUI43^T (Chapter II, Table 3).

ANI and OrthoANI comparisons were made with the Orthologous Average Nucleotide Identity Tool (OAT) v0.93 (https://www.ezbiocloud.net/tools/orthoani). Digital DNA-DNA hybridizations (dDDH) and G+C content differences were obtained with Genome to Genome Distance Calculator (GGDC) v2.0 available at https://ggdc.dsmz.de/ggdc.php#. BLAST+ local alignment and formula 2 output (optimized for draft genome sequences) were used as GGDC preferences (Meier-Kolthoff et al., 2013b). dDDH heatmap was made using GGDC Output Management Assistant (GOMA) described in the first chapter of this thesis.

EDGAR 2.0 platform (Blom et al., 2016) was used to calculate the core genome, dispensable genome and singleton genes of the ten strains closest to the type strains of *M. saelicesensis* and *M. noduli*.

3.3.6- Whole-Genome phylogenomic Analyses

Two approaches were used to infer phylogenomic relationships: Genome Blast Distance Phylogeny approach (GBDP) (Auch et al., 2010) and the Up to date Bacterial Core Gene (UBCG) based on the multilocus analysis of 92 universal bacterial core genes (Na et al., 2018). Genomes for the phylogenetic tree reconstruction were downloaded from NCBI GenBank database (Clark et al., 2016). All genomes used for the tree and their accession numbers are provided in Appendix III.

1. Whole genome phylogenetic tree based on Genome Blast Distance Phylogeny approach (GBDP): Distance matrices were calculated using GGDC online tool v2.0 (Meier-Kolthoff et al., 2013), with the recommended settings of BLAST+ and formula 2 (optimized for incompletely sequenced genomes at contig level). GOMA script was used to construct

distance matrices both in Phylip and comma separated values (.*csv*) formats. Phylip formatted distance table was submitted to FastME v2.0 online tool (Lefort et al., 2015) for phylogenetic reconstruction. As GGDC bootstrapping and jackknifing tool is not freely available, the final tree was based on a unique distance matrix, and therefore no statistical support (bootstrap) was provided.

2. Up to date Bacterial Core Gene (UBCG): standalone UBCG v3.0 program (Na et al., 2018) in combination with UBCG_iTOL_maker script (see Chapter I) were used to infer phylogenetic reconstruction. Default parameters for UBCG program were used (codon alignment method, filtering cut-off for gap containing positions of 50% and FastTree phylogeny reconstruction program). Bootstrap and Gene Support Index (number of individual gene trees that support the node) are given in the tree.

iTOL v3.0 online tool (Letunic and Bork, 2016) was used for the visualization and manual editing of the final annotation details of the inferred trees.

3.3.7- Physiology

A set of physiological and biochemical tests reported to differentiate between *M. saelicesensis* and *M. noduli* were carried out. These tests included carbon source utilization, determination of enzymatic activities, NaCl and pH tolerance, and degradation of starch, Tween 20, Tween 80, tyrosine and urea. All tests were done in triplicate.

Nineteen carbon sources (*D*-maltose, *D*-trehalose, *D*-cellobiose, sucrose, *D*-raffinose, *D*-mannose, *D*-galactose, *L*-rhamnose, *D*-serine, *L*-alanine, *L*-arginine, *L*-histidine, *L*-arabinose, lysine, *L*-proline, sorbose, valine, xylitol and *myo*-inositol) were tested *in vitro* at different times in the same conditions (2016 and 2017), using the method described in Williams et al., (1983). These results were compared with the original description of *Micromonospora saelicesensis* (Trujillo et al., 2007), that contained physiological data of two of the strains studied (Lupac 06 and Lupac 07). The following protocol was used to prepare the media:

1. Sugar carbon sources were prepared at a final concentration of 1% p/v and amino acid carbon sources at a final concentration of 0.1% p/v. Each carbon source was tindalized in 20 ml of distilled water (5X concentration) for 30 min, 100 °C and 1atm, during three consecutive days.

2. 10 ml of Difco^{TM} Yeast Nitrogen Base was prepared at a 10X concentration following the manufacturer indications and filtered with a 0.2 µm sterile filter.

3. A basal agar medium was prepared with a final concentration of 0.5 g/l of KH_2PO_4, 0.5g/l of $MgSO_4$ x7H_2O and 18 g/l of agar. This medium was prepared in 70 ml at a concentration of 1.42X and autoclaved (20 min, 120 °C, 1atm). The media was maintained at 50 °C until use.

4. Sugar carbon source preparation, nitrogen base and basal agar preparation (20:10:70) were mixed to a final volume of 100 ml and plated. The media was stored at 5 °C until use.

Seven-day old bacterial cultures were suspended in 0.8% NaCl for a final concentration 1.8 x 10^9 cells/ml (MacFarland 6). 20 μl were inoculated on each carbon source agar plate in triplicate. Strains were incubated at 28°C for twenty-one days. The test strains were also inoculated on the basal medium alone (negative control) and the medium supplemented with glucose (positive control). Growth on the test media was compared with that on both positive and negative controls. Strains were scored positive if growth on the test plate was greater than that on the negative control. On the contrary, negative results were recorded when growth was less than or equal to the negative control plate. Genome data of the test strains was screened for genes coding for proteins for carbon metabolism of the carbon sources assayed.

API ZYM system (bioMérieux SA) was used for determination of enzymatic activities. All strains were cultures on M65 agar at 28°C for 7 days. Bacterial suspensions were prepared in 0.8% NaCl for a final concentration 1.8 x 10^9 cells/ml (MacFarland 6) and inoculated in the API ZYM microplates. After 48 hours of incubation at 28°C, one drop of ZYMA and ZYMB reagents (bioMérieux SA) were used to develop the enzymatic test. Positive results were assessed by development of color in the microplate and following manufacturer's instructions for interpretation of results

NaCl tolerance was assessed using M65 agar plates supplemented with NaCl at a final concentration of 1%, 2%, 3% 4%, 5%, 6% and 7% (Kutzner, 1981). pH tolerance was evaluated on M65 agar supplemented with appropriate buffers to adjust pH to pH 2, 3, 4, 5, 6, 7, 8, 9 and 10 (Kutzner, 1981). Strains were inoculated in the same manner as explained above and incubated at 28°C for 21 days.

M65 agar plates supplemented with starch (0.5%) were used to evaluate amylase activity. Plates were incubated at 28°C for 10 days. Positive results were developed with addition of lugol, revealing translucent halos within a blue background.

Degradation of Tweens 20 and 80 were made using a bacto-pectone based agar medium (bacto-pectone 10g/l, NaCl 5 g/l, $CaCl_2$ 0.1 g/l and agar 15 g/l) supplemented with the corresponding Tween substrate for a final concentration of 1% (v/v) (Sierra, 1957). Strains were incubated at 28°C for 21 days. Positive results were assessed by the apparition of an opaque degradation halo.

Tyrosine degradation was measured using M65 agar supplemented with *L (-)* tyrosine at a final concentration of 0.4%. Strains inoculated as previously explained and incubated at 28°C for 21 days. Positive results were assessed by the development of a translucent degradation halo. For the urea degradation test, strains were incubated in Urea Broth (Fluka) at 28°C for 14 days (Christensen, 1946). Positive results were visible as a change of the fuchsia original color to purple.

3.3.8- Biolog characterization

GEN III Microplates in an Omnilog device (BIOLOG Inc., Hayward, CA, USA) were used to generate a phenotypic fingerprint of seventy-one carbon sources and twenty-three chemical sensitivity assays.

Strains selected in OGRI analysis and reference strains *Micromonospora saelicesensis* Lupac 09^T and *Micromonospora noduli* GUI43^T were tested at 28°C. One-week old cells were suspended in an inoculating fluid (IF C) provided by the manufacturer and inoculated in the GEN III Microplates at a cell density of 80% transmittance. Phenotype microarray mode was used to measure respiration rates yielding a total running time of 7 days using two independent replicates for each strain. Data were recovered and analyzed using the *opm* package for R, v.1.0.6 (Vaas et al., 2012, 2013). Clustering analyses of the phenotypic microarrays were constructed using the *pvclust* package for R v.1.2.2 (Suzuki and Shimodaira, 2015). Distinct behaviors between the two repetitions in the reactions were regarded as ambiguous.

3.4- RESULTS

3.4.1- 16S rRNA gene phylogeny

The 16S rRNA gene sequences were used to determine the nearest phylogenetic neighbors based on overall sequence similarity in relation to currently described *Micromonospora* species. In all cases, the closest species were *M. saelicesensis* and *M. noduli* with similarity values of 99.3-100% (Chapter II, Table 2.2).

A phylogenetic tree constructed with the study strain sequences and those of 81 *Micromonospora* type strains distributed the sixteen strains into two groups: Group I contained the type strain *M. saelicesensis* Lupac 09^T and the isolates PSN13, GAR06, PSN01, Lupac 06, GAR05 and Lupac 07. Group II was formed with ONO86, ONO23, LAH08 and MED15 and *M. noduli* GUI43^T (Figure 19). The strains LAH09, PSH25, PSH03, NIE79 and NIE111 form an independent cluster, characterized by longer distance branches separating its members from *M. saelicesensis* and showing a closer relationship to *M. zamorensis, M. luteifusca* and *M. vinacea.* Strain MED01 was recovered as a separate strain, between groups I and II.

Tree topology showed the close relationship between the strains in groups I, II and MED01 with very low distances represented by almost inexistent branches. Group II (*M. noduli*), Group I (*M. saelicesensis*) and MED01 also showed a close relationship with the type strains of *Micromonospora profundi* isolated from a deep marine sediment (Veyisoglu et al., 2016) and *Micromonospora ureilytica* isolated from *Pisum sativum* (Carro et al., 2016). Reported DDH values between *M. saelicesensis* and *M. ureilytica* and *M. profundi* (Veyisoglu et al., 2016) were 28.4% and 56.9%, respectively. A DDH value of 50.9% was found between *M. noduli* GUI 43^T and *M. ureilytica* GUI23^T (Carro et al., 2016).

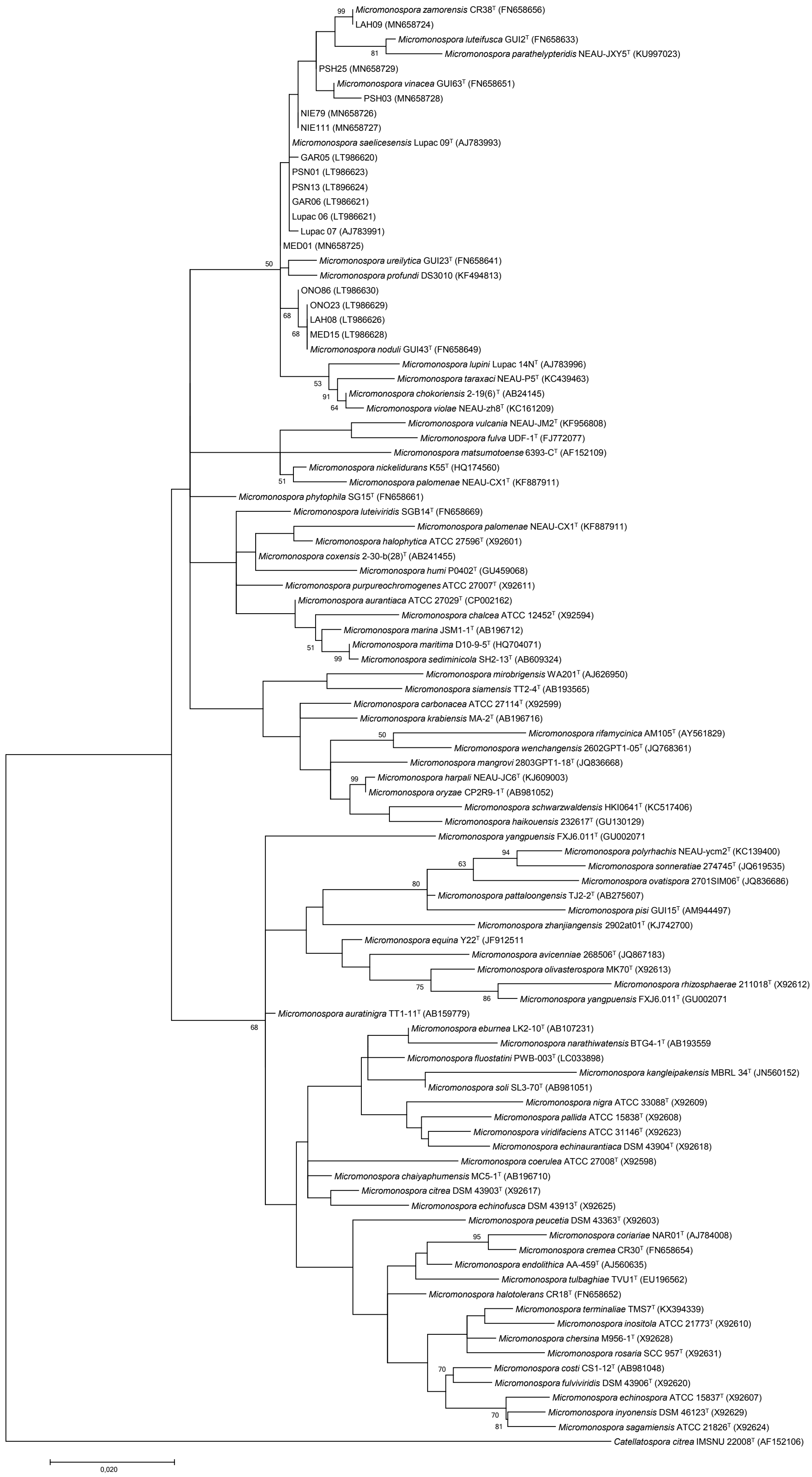

99
Micromonospora zamorensis CR38^{T} (FN658656)
LAH09 (MN658724)
Micromonospora luteifusca GUI2^{T} (FN658633)
81
Micromonospora parathelypteridis NEAU-JXY5^{T} (KU997023)
PSH25 (MN658729)
Micromonospora vinacea GUI63^{T} (FN658651)
PSH03 (MN658728)
NIE79 (MN658726)
NIE111 (MN658727)
Micromonospora saelicesensis Lupac 09^{T} (AJ783993)
GAR05 (LT986620)
PSN01 (LT986623)
PSN13 (LT896624)
GAR06 (LT986621)
Lupac 06 (LT986621)
Lupac 07 (AJ783991)
MED01 (MN658725)
50
Micromonospora ureilytica GUI23^{T} (FN658641)
Micromonospora profundi DS3010 (KF494813)
ONO86 (LT986630)
ONO23 (LT986629)
68
LAH08 (LT986626)
68
MED15 (LT986628)
Micromonospora noduli GUI43^{T} (FN658649)
Micromonospora lupini Lupac 14N^{T} (AJ783996)
53
Micromonospora taraxaci NEAU-P5^{T} (KC439463)
91
Micromonospora chokoriensis 2-19(6)^{T} (AB24145)
64
Micromonospora violae NEAU-zh8^{T} (KC161209)
Micromonospora vulcania NEAU-JM2^{T} (KF956808)
Micromonospora fulva UDF-1^{T} (FJ772077)
Micromonospora matsumotoense 6393-C^{T} (AF152109)
Micromonospora nickelidurans K55^{T} (HQ174560)
51
Micromonospora palomenae NEAU-CX1^{T} (KF887911)
Micromonospora phytophila SG15^{T} (FN658661)
Micromonospora luteiviridis SGB14^{T} (FN658669)
Micromonospora palomenae NEAU-CX1^{T} (KF887911)
Micromonospora halophytica ATCC 27596^{T} (X92601)
Micromonospora coxensis 2-30-b(28)^{T} (AB241455)
Micromonospora humi P0402^{T} (GU459068)
Micromonospora purpureochromogenes ATCC 27007^{T} (X92611)
Micromonospora aurantiaca ATCC 27029^{T} (CP002162)
Micromonospora chalcea ATCC 12452^{T} (X92594)
Micromonospora marina JSM1-1^{T} (AB196712)
51
Micromonospora maritima D10-9-5^{T} (HQ704071)
99
Micromonospora sediminicola SH2-13^{T} (AB609324)
Micromonospora mirobrigensis WA201^{T} (AJ626950)
Micromonospora siamensis TT2-4^{T} (AB193565)
Micromonospora carbonacea ATCC 27114^{T} (X92599)
Micromonospora krabiensis MA-2^{T} (AB196716)
50
Micromonospora rifamycinica AM105^{T} (AY561829)
Micromonospora wenchangensis 2602GPT1-05^{T} (JQ768361)
Micromonospora mangrovi 2803GPT1-18^{T} (JQ836668)
99
Micromonospora harpali NEAU-JC6^{T} (KJ609003)
Micromonospora oryzae CP2R9-1^{T} (AB981052)
Micromonospora schwarzwaldensis HKI0641^{T} (KC517406)
Micromonospora haikouensis 232617^{T} (GU130129)
Micromonospora yangpuensis FXJ6.011^{T} (GU002071)
94
Micromonospora polyrhachis NEAU-ycm2^{T} (KC139400)
63
Micromonospora sonneratiae 274745^{T} (JQ619535)
80
Micromonospora ovatispora 2701SIM06^{T} (JQ836686)
Micromonospora pattaloongensis TJ2-2^{T} (AB275607)
Micromonospora pisi GUI15^{T} (AM944497)
Micromonospora zhanjiangensis 2902at01^{T} (KJ742700)
Micromonospora equina Y22^{T} (JF912511
Micromonospora avicenniae 268506^{T} (JQ867183)
Micromonospora olivasterospora MK70^{T} (X92613)
75
Micromonospora rhizosphaerae 211018^{T} (X92612)
86
Micromonospora yangpuensis FXJ6.011^{T} (GU002071
Micromonospora auratinigra TT1-11^{T} (AB159779)
68
Micromonospora eburnea LK2-10^{T} (AB107231)
Micromonospora narathiwatensis BTG4-1^{T} (AB193559
Micromonospora fluostatini PWB-003^{T} (LC033898)
Micromonospora kangleipakensis MBRL 34^{T} (JN560152)
Micromonospora soli SL3-70^{T} (AB981051)
Micromonospora nigra ATCC 33088^{T} (X92609)
Micromonospora pallida ATCC 15838^{T} (X92608)
Micromonospora viridifaciens ATCC 31146^{T} (X92623)
Micromonospora echinaurantiaca DSM 43904^{T} (X92618)
Micromonospora coerulea ATCC 27008^{T} (X92598)
Micromonospora chaiyaphumensis MC5-1^{T} (AB196710)
Micromonospora citrea DSM 43903^{T} (X92617)
Micromonospora echinofusca DSM 43913^{T} (X92625)
Micromonospora peucetia DSM 43363^{T} (X92603)
95
Micromonospora coriariae NAR01^{T} (AJ784008)
Micromonospora cremea CR30^{T} (FN658654)
Micromonospora endolithica AA-459^{T} (AJ560635)
Micromonospora tulbaghiae TVU1^{T} (EU196562)
Micromonospora halotolerans CR18^{T} (FN658652)
Micromonospora terminaliae TMS7^{T} (KX394339)
Micromonospora inositola ATCC 21773^{T} (X92610)
Micromonospora chersina M956-1^{T} (X92628)
Micromonospora rosaria SCC 957^{T} (X92631)
70
Micromonospora costi CS1-12^{T} (AB981048)
Micromonospora fulviviridis DSM 43906^{T} (X92620)
Micromonospora echinospora ATCC 15837^{T} (X92607)
70
Micromonospora inyonensis DSM 46123^{T} (X92629)
81
Micromonospora sagamiensis ATCC 21826^{T} (X92624)
Catellatospora citrea IMSNU 22008^{T} (AF152106)
0,020

Figure 19: Maximum-likelihood phylogenetic tree based on 16S rRNA gene sequences showing the relationships between 81 *Micromonospora* type and the study strains. Distances were calculated with the Kimura 2-parameter. The tree is based on 1,319 nt. Bootstrap percentages ≥50% (1,000 samplings) are shown at nodes. Bar, 0.02 substitutions per nucleotide.

3.4.2- *gyr*B phylogeny

The phylogenetic tree constructed with the *gyr*B gene sequences showed a similar topology to the 16S rRNA gene tree with respect to the members of group I (*M. saelicesensis*) and II (*M. noduli*) (Figure 20). Again, two groups were recovered, with almost similar composition. The exception was strain Lupac 07, recovered in the *M. noduli* cluster (Group II) with a support bootstrap value of 99%. Interestingly, this strain was originally classified as *M. saelicesensis* (Trujillo et al., 2007). Strains PSH03, MED01, NIE111, NIE79 and PSH25 were recovered as separate strains and appear less related to the species *M. noduli* and *M. saelicesensis*. Isolate LAH09 is positioned at a significant distance from the rest of the study strains and appears more related to *M. zamorensis* DSM 45600^T.

The positions of the type strains *M. profundi* DS 3010^T and *M. ureilytica* GUI23^T also changed and moved out of the *M. noduli* and *M. saelicesensis* clusters. As previously noted (Carro et al., 2012; Garcia et al., 2010), the *gyr*B gene phylogeny yielded a better resolution as observed by slightly larger distance branches, however, the topology of the remaining type strains was very different from that obtained using the 16S rRNA gene.

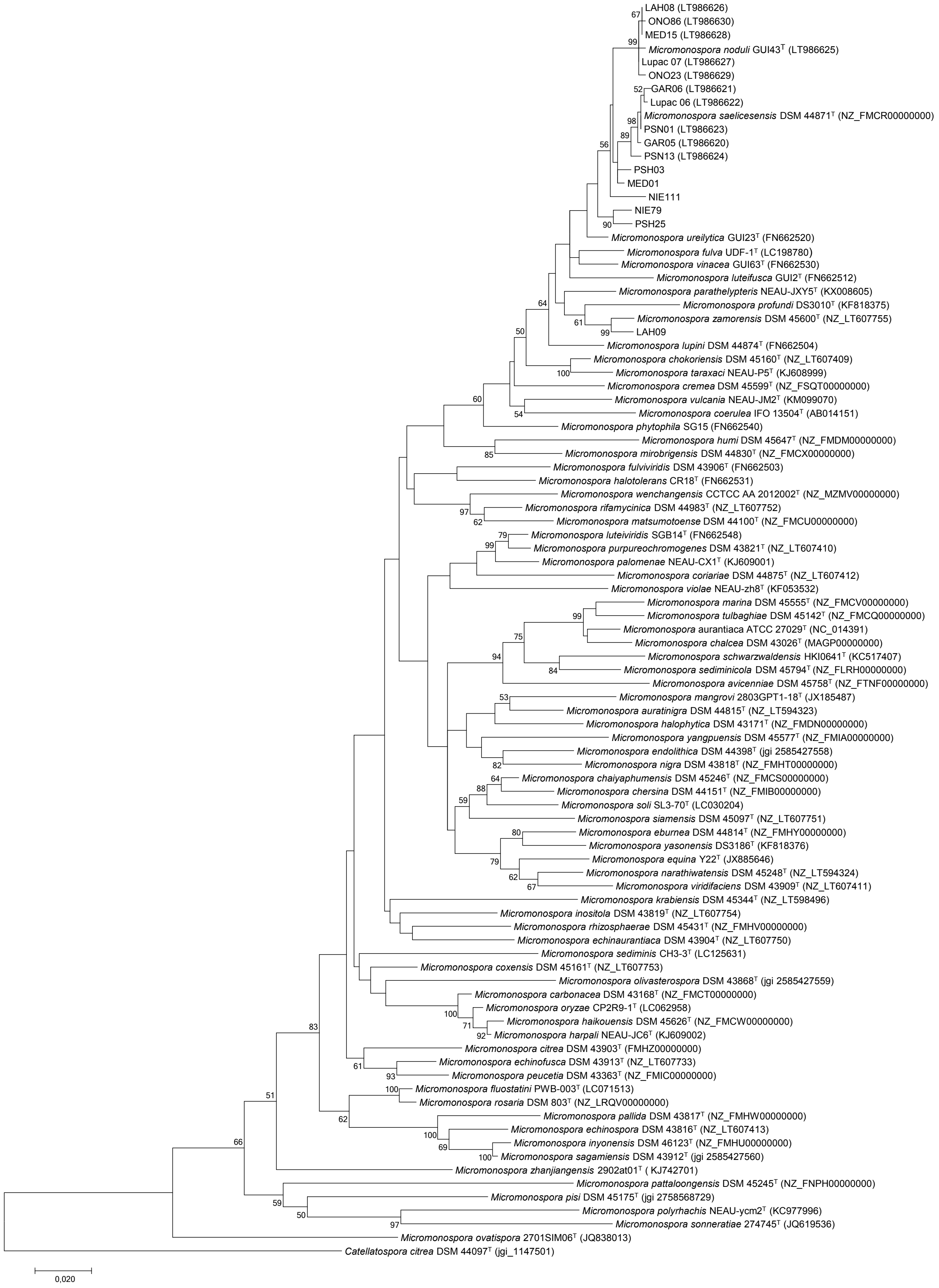

LAH08 (LT986626)
ONO86 (LT986630)
MED15 (LT986628)
Micromonospora noduli GUI43^{T} (LT986625)
Lupac 07 (LT986627)
ONO23 (LT986629)
GAR06 (LT986621)
Lupac 06 (LT986622)
Micromonospora saelicesensis DSM 44871^{T} (NZ_FMCR00000000)
PSN01 (LT986623)
GAR05 (LT986620)
PSN13 (LT986624)
PSH03
MED01
NIE111
NIE79
PSH25
Micromonospora ureilytica GUI23^{T} (FN662520)
Micromonospora fulva UDF-1^{T} (LC198780)
Micromonospora vinacea GUI63^{T} (FN662530)
Micromonospora luteifusca GUI2^{T} (FN662512)
Micromonospora parathelypteris NEAU-JXY5^{T} (KX008605)
Micromonospora profundi DS3010^{T} (KF818375)
Micromonospora zamorensis DSM 45600^{T} (NZ_LT607755)
LAH09
Micromonospora lupini DSM 44874^{T} (FN662504)
Micromonospora chokoriensis DSM 45160^{T} (NZ_LT607409)
Micromonospora taraxaci NEAU-P5^{T} (KJ608999)
Micromonospora cremea DSM 45599^{T} (NZ_FSQT00000000)
Micromonospora vulcania NEAU-JM2^{T} (KM099070)
Micromonospora coerulea IFO 13504^{T} (AB014151)
Micromonospora phytophila SG15 (FN662540)
Micromonospora humi DSM 45647^{T} (NZ_FMDM00000000)
Micromonospora mirobrigensis DSM 44830^{T} (NZ_FMCX00000000)
Micromonospora fulviviridis DSM 43906^{T} (FN662503)
Micromonospora halotolerans CR18^{T} (FN662531)
Micromonospora wenchangensis CCTCC AA 2012002^{T} (NZ_MZMV00000000)
Micromonospora rifamycinica DSM 44983^{T} (NZ_LT607752)
Micromonospora matsumotoense DSM 44100^{T} (NZ_FMCU00000000)
Micromonospora luteiviridis SGB14^{T} (FN662548)
Micromonospora purpureochromogenes DSM 43821^{T} (NZ_LT607410)
Micromonospora palomenae NEAU-CX1^{T} (KJ609001)
Micromonospora coriariae DSM 44875^{T} (NZ_LT607412)
Micromonospora violae NEAU-zh8^{T} (KF053532)
Micromonospora marina DSM 45555^{T} (NZ_FMCV00000000)
Micromonospora tulbaghiae DSM 45142^{T} (NZ_FMCQ00000000)
Micromonospora aurantiaca ATCC 27029^{T} (NC_014391)
Micromonospora chalcea DSM 43026^{T} (MAGP00000000)
Micromonospora schwarzwaldensis HKI0641^{T} (KC517407)
Micromonospora sediminicola DSM 45794^{T} (NZ_FLRH00000000)
Micromonospora avicenniae DSM 45758^{T} (NZ_FTNF00000000)
Micromonospora mangrovi 2803GPT1-18^{T} (JX185487)
Micromonospora auratinigra DSM 44815^{T} (NZ_LT594323)
Micromonospora halophytica DSM 43171^{T} (NZ_FMDN00000000)
Micromonospora yangpuensis DSM 45577^{T} (NZ_FMIA00000000)
Micromonospora endolithica DSM 44398^{T} (jgi 2585427558)
Micromonospora nigra DSM 43818^{T} (NZ_FMHT00000000)
Micromonospora chaiyaphumensis DSM 45246^{T} (NZ_FMCS00000000)
Micromonospora chersina DSM 44151^{T} (NZ_FMIB00000000)
Micromonospora soli SL3-70^{T} (LC030204)
Micromonospora siamensis DSM 45097^{T} (NZ_LT607751)
Micromonospora eburnea DSM 44814^{T} (NZ_FMHY00000000)
Micromonospora yasonensis DS3186^{T} (KF818376)
Micromonospora equina Y22^{T} (JX885646)
Micromonospora narathiwatensis DSM 45248^{T} (NZ_LT594324)
Micromonospora viridifaciens DSM 43909^{T} (NZ_LT607411)
Micromonospora krabiensis DSM 45344^{T} (NZ_LT598496)
Micromonospora inositola DSM 43819^{T} (NZ_LT607754)
Micromonospora rhizosphaerae DSM 45431^{T} (NZ_FMHV00000000)
Micromonospora echinaurantiaca DSM 43904^{T} (NZ_LT607750)
Micromonospora sediminis CH3-3^{T} (LC125631)
Micromonospora coxensis DSM 45161^{T} (NZ_LT607753)
Micromonospora olivasterospora DSM 43868^{T} (jgi 2585427559)
Micromonospora carbonacea DSM 43168^{T} (NZ_FMCT00000000)
Micromonospora oryzae CP2R9-1^{T} (LC062958)
Micromonospora haikouensis DSM 45626^{T} (NZ_FMCW00000000)
Micromonospora harpali NEAU-JC6^{T} (KJ609002)
Micromonospora citrea DSM 43903^{T} (FMHZ00000000)
Micromonospora echinofusca DSM 43913^{T} (NZ_LT607733)
Micromonospora peucetia DSM 43363^{T} (NZ_FMIC00000000)
Micromonospora fluostatini PWB-003^{T} (LC071513)
Micromonospora rosaria DSM 803^{T} (NZ_LRQV00000000)
Micromonospora pallida DSM 43817^{T} (NZ_FMHW00000000)
Micromonospora echinospora DSM 43816^{T} (NZ_LT607413)
Micromonospora inyonensis DSM 46123^{T} (NZ_FMHU00000000)
Micromonospora sagamiensis DSM 43912^{T} (jgi 2585427560)
Micromonospora zhanjiangensis 2902at01^{T} (KJ742701)
Micromonospora pattaloongensis DSM 45245^{T} (NZ_FNPH00000000)
Micromonospora pisi DSM 45175^{T} (jgi 2758568729)
Micromonospora polyrhachis NEAU-ycm2^{T} (KC977996)
Micromonospora sonneratiae 274745^{T} (JQ619536)
Micromonospora ovatispora 2701SIM06^{T} (JQ838013)
Catellatospora citrea DSM 44097^{T} (jgi_1147501)
0,020

Figure 20: Maximum-likelihood phylogenetic tree based on *gyr*B gene sequences. A total of 76 *Micromonospora* type strains and 10 non-type strains have been used for the analysis. Distances were calculated with the Kimura 2-parameter, using 1001 nucleotide positions and a bootstrap of 1,000. Bootstrap percentages ≥50% (1,000 samplings) are shown at nodes. Bar, 0.02 substitutions per nucleotide.

3.4.3- Multi-Locus Sequence Analysis (MLSA) phylogeny

Phylogenetic tree reconstruction based on the concatenation of five core genes (16S rRNA, *gyr*B, *rpo*B, *atp*D and *rec*A) showed a very similar topology to the *gyr*B gene tree with respect to position of the study strains. Again, two groups were visualized and isolate Lupac 07 was recovered in the *M. noduli* cluster (Group II) with 99% bootstrap support. The type strain *M. ureilytica* GUI23^T was recovered outside the *M. noduli* cluster in a similar topology as the *gyr*B gene phylogeny. Strains PSH03, NIE111 and MED01 grouped as a separate cluster, far from *M. saelicesensis* and *M. noduli* groups. Isolates PSH25 and NIE79 clustered with *M. ureilytica,* but with long distance branches between them while strain LAH09 clustered with *M. zamorensis* DSM 45600^T, but forming a long branch (Figure 21).

The remaining MLSA tree topology resembled the groups described in 2018 by Carro and colleagues (Carro et al., 2018). Pairwise similarity values for all strains are given in Appendix IV.

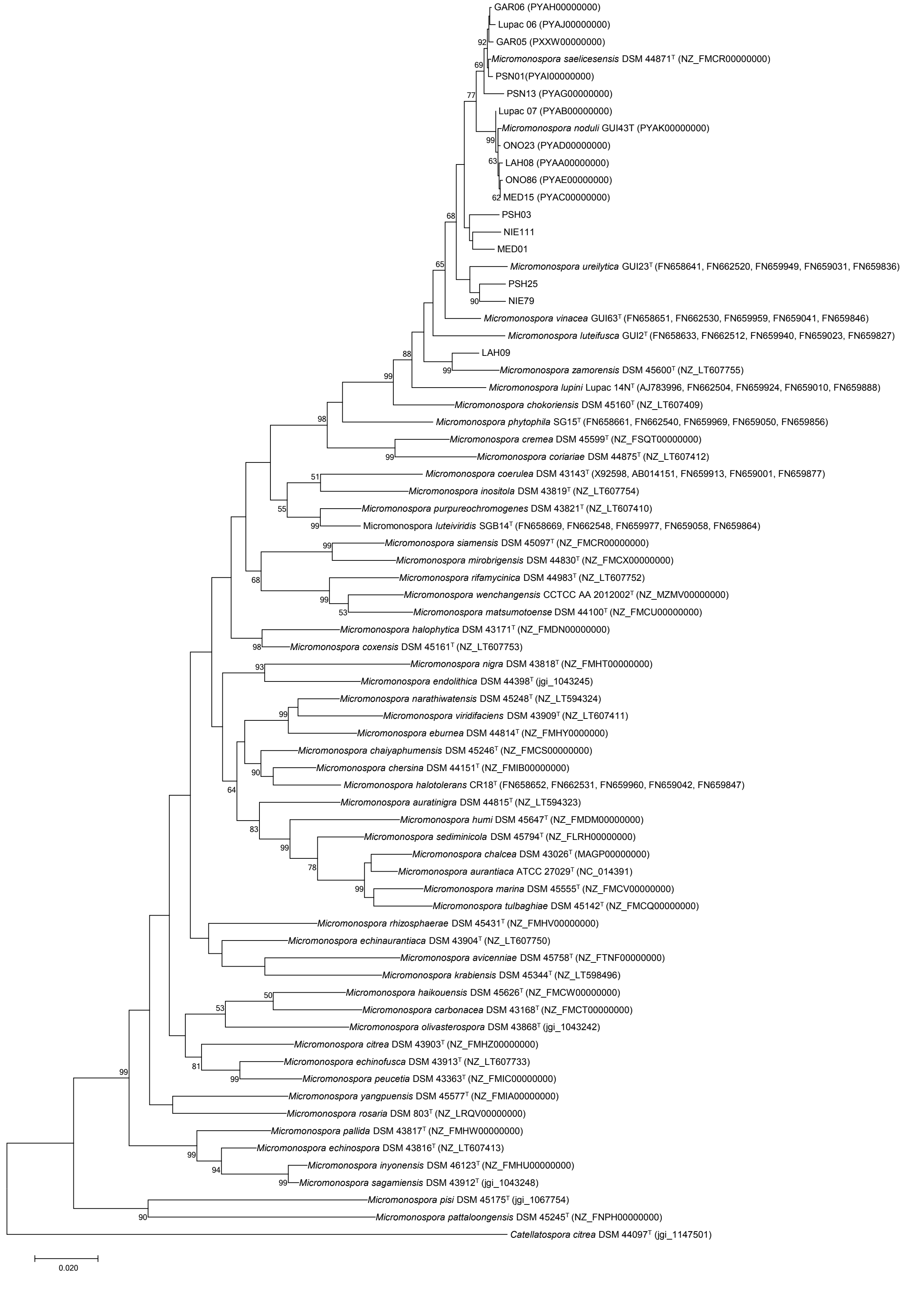

GAR06 (PYAH00000000)
Lupac 06 (PYAJ00000000)
92
GAR05 (PXXW00000000)
Micromonospora saelicesensis DSM 44871^T (NZ_FMCR00000000)
69
PSN01(PYAI00000000)
77
PSN13 (PYAG00000000)
Lupac 07 (PYAB00000000)
Micromonospora noduli GUI43T (PYAK00000000)
99
ONO23 (PYAD00000000)
63
LAH08 (PYAA00000000)
ONO86 (PYAE00000000)
62
MED15 (PYAC00000000)
68
PSH03
NIE111
MED01
65
Micromonospora ureilytica GUI23^T (FN658641, FN662520, FN659949, FN659031, FN659836)
PSH25
90
NIE79
Micromonospora vinacea GUI63^T (FN658651, FN662530, FN659959, FN659041, FN659846)
Micromonospora luteifusca GUI2^T (FN658633, FN662512, FN659940, FN659023, FN659827)
88
LAH09
99
Micromonospora zamorensis DSM 45600^T (NZ_LT607755)
99
Micromonospora lupini Lupac 14N^T (AJ783996, FN662504, FN659924, FN659010, FN659888)
Micromonospora chokoriensis DSM 45160^T (NZ_LT607409)
98
Micromonospora phytophila SG15^T (FN658661, FN662540, FN659969, FN659050, FN659856)
Micromonospora cremea DSM 45599^T (NZ_FSQT00000000)
99
Micromonospora coriariae DSM 44875^T (NZ_LT607412)
51
Micromonospora coerulea DSM 43143^T (X92598, AB014151, FN659913, FN659001, FN659877)
Micromonospora inositola DSM 43819^T (NZ_LT607754)
55
Micromonospora purpureochromogenes DSM 43821^T (NZ_LT607410)
99
Micromonospora *luteiviridis* SGB14^T (FN658669, FN662548, FN659977, FN659058, FN659864)
99
Micromonospora siamensis DSM 45097^T (NZ_FMCR00000000)
Micromonospora mirobrigensis DSM 44830^T (NZ_FMCX00000000)
68
Micromonospora rifamycinica DSM 44983^T (NZ_LT607752)
99
Micromonospora wenchangensis CCTCC AA 2012002^T (NZ_MZMV00000000)
53
Micromonospora matsumotoense DSM 44100^T (NZ_FMCU00000000)
Micromonospora halophytica DSM 43171^T (NZ_FMDN00000000)
98
Micromonospora coxensis DSM 45161^T (NZ_LT607753)
93
Micromonospora nigra DSM 43818^T (NZ_FMHT00000000)
Micromonospora endolithica DSM 44398^T (jgi_1043245)
Micromonospora narathiwatensis DSM 45248^T (NZ_LT594324)
99
Micromonospora viridifaciens DSM 43909^T (NZ_LT607411)
Micromonospora eburnea DSM 44814^T (NZ_FMHY0000000)
Micromonospora chaiyaphumensis DSM 45246^T (NZ_FMCS00000000)
90
Micromonospora chersina DSM 44151^T (NZ_FMIB00000000)
64
Micromonospora halotolerans CR18^T (FN658652, FN662531, FN659960, FN659042, FN659847)
Micromonospora auratinigra DSM 44815^T (NZ_LT594323)
83
Micromonospora humi DSM 45647^T (NZ_FMDM00000000)
99
Micromonospora sediminicola DSM 45794^T (NZ_FLRH00000000)
78
Micromonospora chalcea DSM 43026^T (MAGP00000000)
Micromonospora aurantiaca ATCC 27029^T (NC_014391)
99
Micromonospora marina DSM 45555^T (NZ_FMCV00000000)
Micromonospora tulbaghiae DSM 45142^T (NZ_FMCQ00000000)
Micromonospora rhizosphaerae DSM 45431^T (NZ_FMHV00000000)
Micromonospora echinaurantiaca DSM 43904^T (NZ_LT607750)
Micromonospora avicenniae DSM 45758^T (NZ_FTNF00000000)
Micromonospora krabiensis DSM 45344^T (NZ_LT598496)
50
Micromonospora haikouensis DSM 45626^T (NZ_FMCW00000000)
53
Micromonospora carbonacea DSM 43168^T (NZ_FMCT00000000)
Micromonospora olivasterospora DSM 43868^T (jgi_1043242)
Micromonospora citrea DSM 43903^T (NZ_FMHZ00000000)
81
Micromonospora echinofusca DSM 43913^T (NZ_LT607733)
99
99
Micromonospora peucetia DSM 43363^T (NZ_FMIC00000000)
Micromonospora yangpuensis DSM 45577^T (NZ_FMIA00000000)
Micromonospora rosaria DSM 803^T (NZ_LRQV00000000)
Micromonospora pallida DSM 43817^T (NZ_FMHW00000000)
Micromonospora echinospora DSM 43816^T (NZ_LT607413)
99
94
Micromonospora inyonensis DSM 46123^T (NZ_FMHU00000000)
99
Micromonospora sagamiensis DSM 43912^T (jgi_1043248)
Micromonospora pisi DSM 45175^T (jgi_1067754)
90
Micromonospora pattaloongensis DSM 45245^T (NZ_FNPH00000000)
Catellatospora citrea DSM 44097^T (jgi_1147501)
0.020

Figure 21: Maximum-likelihood phylogenetic tree based on a concatenation of 16S rRNA, *gyr*B, *rpo*B, *atp*D and *rec*A genes. Distances were calculated with the Kimura 2-parameter, using 4340 nucleotide positions and a bootstrap of 1,000. Bootstrap percentages ≥50% (1,000 samplings) are shown at nodes. Bar, 0.02 substitutions per nucleotide. Accession number for all genomes and individual sequences are given in the tree.

3.4.4- OGRI indices

Overall genomic relatedness indices (Chun and Rainey, 2014) were used to determine the relationship between each pair of genomes obtained previously (Section 2.3.4, Chapter II). The strains were distributed in eight groups according to their OGRI values that included ANI, ortho-ANI, dDDH and G+C values. Group I included the strains GAR05, GAR06, Lupac 06, PSN01 and PSN13 in addition to the type strain *Micromonospora saelicesensis* DSM 44871^{T}. Group II contained the isolates MED15, ONO23, ONO86, Lupac 07, LAH08 and the type strain *Micromonospora noduli* GUI43^{T}. The six remaining single strain groups were represented by the isolates MED01, NIE79, NIE111, PSH03, PSH25 and LAH09 with values below the recommended cut-off threshold for species boundaries and confirmed that they did not belong to the species *M. saelicesensis* or *M. noduli*.

In *M. saelicesensis* (Group I) ANI and OrthoANI values ranged from 97.82-99.13% and 97.96-99.19%, respectively, between the type and study strains. The *M. noduli* group (Group II) had ANI and OrthoANI values from 99.05-99.09% and 99.12-99.14% respectively (Table 8). In both cases, these values were above the recommended cut-off value of ~96% for species recognition (Richter and Rosselló-Móra, 2009).

The ANI and OrthoANI values between the type strains *M. saelicesensis* and *M. noduli* were 96.6% and 96.8% respectively. Overall, pairwise comparison between groups I and II showed ANI and OrthoANI values ranging from 96.2% to 96.6% and 96.2% to 96.9% for ANI and OrthoANI respectively. The highest ANI and OrthoANI values corresponding to strains GAR05 and GUI43^{T} (96.7%, ANI) and PSN01 and ONO23 (96.9%, OrthoANI) (Appendix V). Both results are slightly above the border line of 95-96% for the delineation of species. However, these results are comparable to OrthoANI values obtained for the genome pairs of *M. carbonacea* and *M. haikouensis* (95.2%), and *M. inyonensis* and *M. sagamiensis* (96.5%).

Members of the six single strain groups had ANI and OrthoANI values ranging from 90.1 to 95.7 and 90.5 to 95.9 with *M. noduli* and values of 90.3 to 95.7 and 90.7 and 95.9 with *M. saelicesensis* (Table 8). The single strain groups were below the recommended cut-off value of ~96% for species recognition.

Species delineation based on dDDH values ranged from 81.0 to 93.7% for the six strains in Group I (*M. saelicesensis*) and 92.3 to 93.8% for the strains contained in group II (*M. noduli*), all values clearly above the recommended threshold value of 70% (Appendix V). Similar to ANI and OrthoANI results, dDDH values between the two groups were slightly above the border limit threshold value of 70% (71.0 – 71.8%) (Appendix V). Overall pairwise comparisons of the study genomes and 48 additional *Micromonospora* type strains show the close relationship of the strains but clearly delineate each group within this 70-71% dDDH radius (Figure 22). A similar situation is observed between the species *Micromonospora sagamiensis* and *Micromonospora inyonensis* which share a dDDH value close to 70% (69.8%, dDDH; 61.3%, experimental DDH) (Kroppenstedt et al., 2005). Meier-Kolthoff and colleagues (Meier-Kolthoff et al., 2014a), recently proposed the delineation of subspecies using genomic data. Specifically, the authors recommended a dDDH threshold value of 79-80% to define subspecies in prokaryotic taxonomy. In the present study, values between strains in *M. noduli* and *M. saelicesensis* groups are much lower than this range, and therefore are better classified as separate species, rather than subspecies.

Internal dDDH values in the remaining single strain groups ranged from 39.6 to 61.0%, values which are clearly under the recommended threshold for species delineation of 70%. Furthermore, with respect to all strains in *M. noduli* and *M. saelicesensis* groups, the values ranged from 40.7 to 65.8%, being LAH09 the strain with the lowest dDDH values.

The G+C mol% among all strains was very homogeneous. The values of the *M. saelicesensis* group ranged from 71.1 – 71.2% while the strains in the *M. noduli* cluster varied from 70.9 to 71.1%. The six single group strain values ranged from 70.8 to 71.6%. As observed, the G+C mol% values between all genomes was less than 1% (0-0.8%).

Based on all OGRI values analyzed in this section, the strains MED01, NIE79, NIE111, PSH03, PSH25 and LAH09, representing single member groups had values below recommended thresholds for species delineation. Furthermore, *gyr*B gene and MLSA phylogenetic analysis also supported their status as separate species. As the scope of this study was to clarify the taxonomic relationship between the species *M. saelicesensis* and *M. noduli*, the isolates MED01, NIE79, NIE111, PSH03, PSH25 and LAH09, were not included in the core genome, phylogenomic and phenotypic studies.

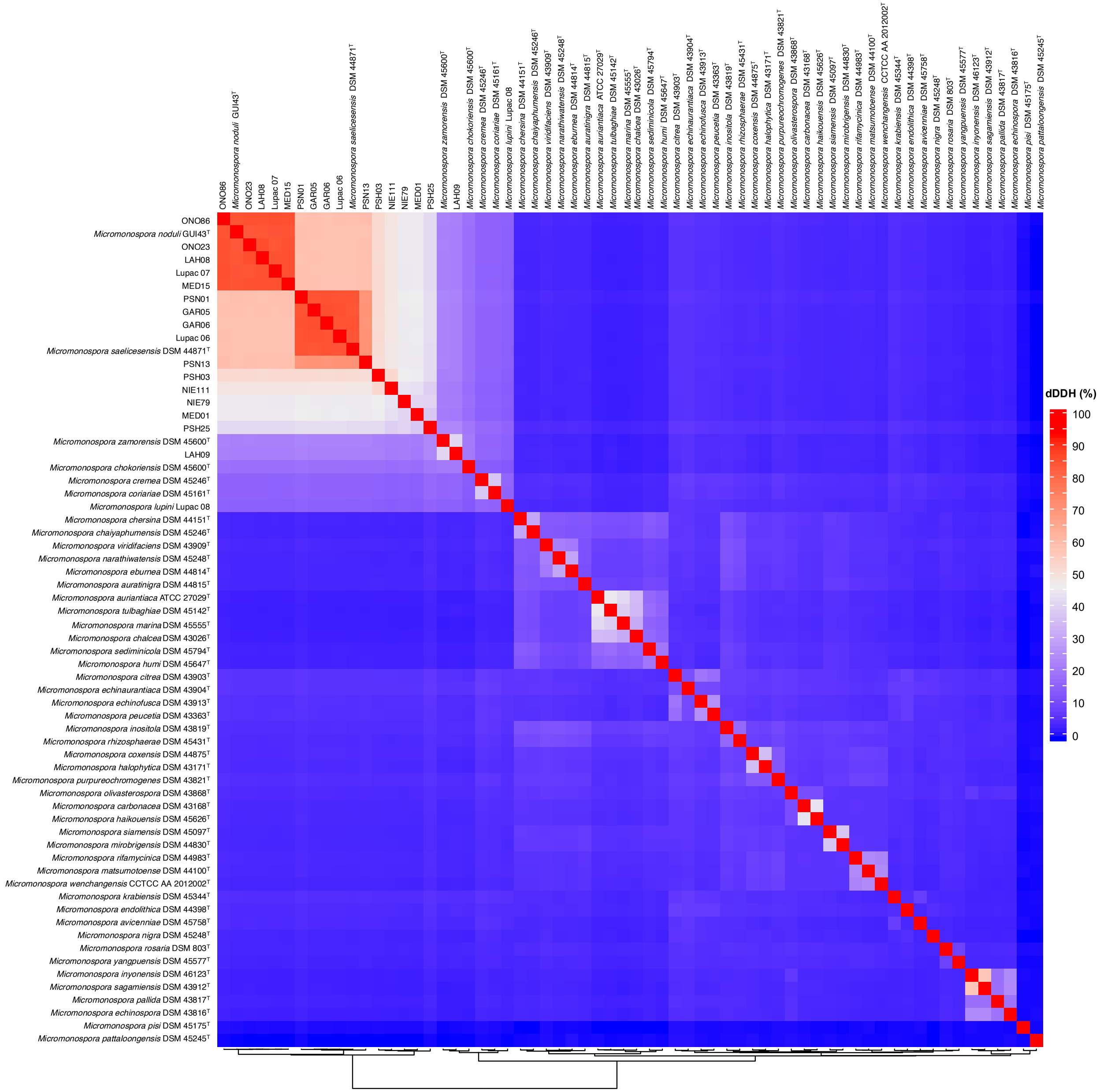


Figure 22: Digital DNA to DNA Hybridization (dDDH) pairwise comparison heatmap. Light red color denotes values close to species delineation threshold (70%).

Table 8: OGRI indices: ANI, OrthoANI and dDDH percentage values calculated between the type and study strains.

Strain	**ANI**	**OrthoANI**	**dDDH**
	M. saelicesensis DSM 44871^{T} / *M. noduli* GUI43^{T}	*M. saelicesensis* DSM 44871^{T} / *M. noduli* GUI43^{T}	*M. saelicesensis* DSM 44871^{T} / *M. noduli* GUI43^{T}
M. *saelicesensis* DSM 44871^{T}	100/96.6	100/96.8	100/71.2
GAR 05	99.1/96.7	99.2/96.8	92.5/71.5
GAR 06	99.1/96.6	99.2/96.8	92.9/71.3
Lupac 06	99.1/96.6	99.2/96.8	92.7/71.2
PSN 01	99.1/96.6	99.2/96.8	92.8/71.6
PSN 13	97.8/96.6	98.0/96.8	81.3/71.3
	M. noduli GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}	*M. noduli* GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}	*M. noduli* GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}
M. noduli GUI43^{T}	100/96.6	100/96.8	100/71.2
LAH 08	99.1/96.2	99.1/96.8	92.5/71.3
Lupac 07	99.1/96.6	99.1/96.8	92.7/71.3
MED 15	99.1/96.6	99.1/96.8	92.6/71.1
ONO 23	99.0/96.6	99.1/96.9	92.3/71.3
ONO 86	99.0/96.6	99.1/96.2	92.8/71.5
	M. noduli GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}	*M. noduli* GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}	*M. noduli* GUI43^{T}/ *M. saelicesensis* DSM 44871^{T}
MED01	94.9/94.9	95.1/95.2	60.3/60.6
NIE79	94.8/94.9	95.1/95.5	60.3/60.4
NIE111	95.3/95.3	95.6/95.9	63.1/62.7
PSH03	95.7/95.7	95.9/95.1	65.8/65.4
PSH25	94.3/94.3	94.6/94.7	57.4/57.9
LAH09	90.1/90.3	90.5/90.7	40.8/41.0

3.4.5- Core genome analysis

The core genome of the six strains in Group I (*M. saelicesensis*) was calculated to be 5313 genes that represents 81.3% of the genome considering an average genome of 6531 genes. The number of singletons ranged from 94 for GAR06 to 706 for PSN13. In *M. noduli* strains (group II) the core genome included 5759 genes, that represent 88.05% of an average genome of 6540 genes. In this group, strain Lupac 07 had the lowest number of singletons with 84 genes and ONO86 the largest number with 369 genes (Figure 23 and Table 9). The calculation of a core genome based on all strains dropped to 74.72% and contained 4884 genes (Table 9). The calculated pangenomes were 8405, 7857 and 9867 genes for *M. saelicesensis* (Group I), *M. noduli* (Group II) and the combination of both species, respectively (Table 9). As expected, an increase in the number of genes in the global pangenome was observed when all strains were considered for pangenome calculation, suggesting an important degree of variation between the genomes. The progression of the pan- and core genome can be seen in Figure 24.

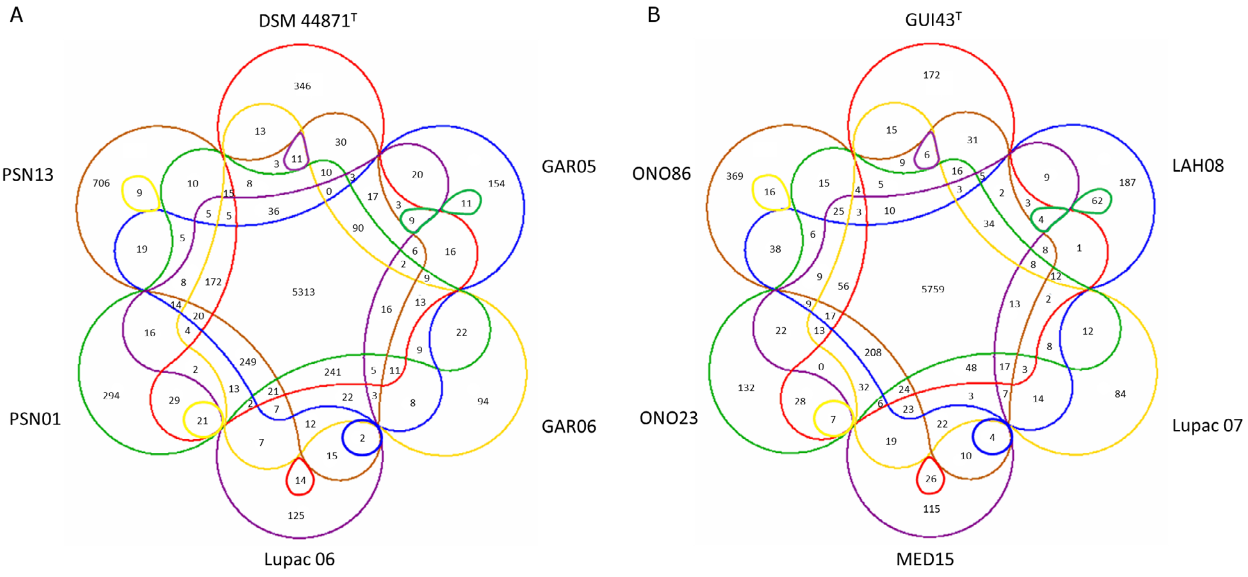


Figure 23: Venn diagram showing the number of clusters of orthologous genes, that conforms the core genome, the disposable genome, and singletons, **A**: between all strains in Group I (*Micromonospora saelicesensis*). **B:** between all strains in Group II (*Micromonospora noduli)*.

Table 9: Number of orthologous genes that conform the pan genome, core genome and singletons of *Micromonospora saelicesensis* (Group 1) and *Micromonospora noduli* (Group II). In parenthesis, values expressed as percentages based on an average genome of 6531 genes for *M. saelicesensis* and 6540 genes for *M. noduli.*

	Pan genome	Core genome	Singletons	
			Strain	Gene count (%)
***M. saelicesensis* (Group I)**	8405	5313 (81.35%)	DSM 44871^T	346 (5.29%)
			GAR05	154 (2.36%)
			GAR06	94 (1.44%))
			Lupac 06	125 (1.91%)
			PSN01	294 (4.50%)
			PSN13	706 (10.08%)
***M. noduli* (Group II)**	7857	5759 (88.05%)	GUI43^T	172 (2.63%)
			LAH08	187 (2.86%)
			Lupac 07	84 (1.28%)
			MED15	115 (1.76%)
			ONO23	132 (2.02%)
			ONO86	369 (5.64%)

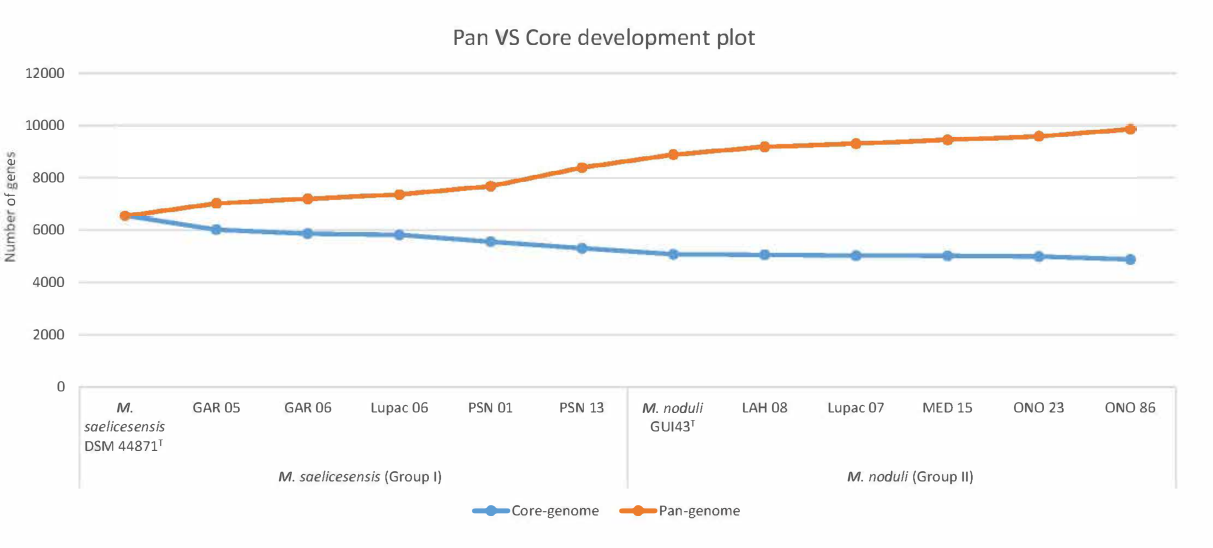


Figure 24: Pan- and Core genome development plot of *Micromonospora noduli* and *Micromonospora saelicesensis* strains. The orange and blue lines show the progression in the pan- and core genomes as more genomes are added.

3.4.6- Whole genome phylogenomic analysis (GBDP method)

Phylogenomic tree reconstruction based on whole-genome distances calculated with the GBDP tool is presented in Figure 25. This tree included the 10 study genomes, all *Micromonospora* strains (type and non-type) published previously (Carro et al., 2018) and the genome sequences of the type strains *M. noduli* GUI43^T, *M. avicenniae* DSM 45748^T, *M. pisi* DSM 45175^T**,** *M. pattaloongensis* DSM 45245^T, *M. rosaria* DSM 803^T and *M. wenchangensis* CCTCC AA 2012002^T (Appendix III). The composition of the *M. saelicesensis* and *M. noduli* groups defined in the *gyr*B gene and MLSA trees were identical, including the position of Lupac 07 as a member of *M. noduli* (Group II).

The overall topology of this tree and the one published by Carro et al. (Carro et al., 2018) was very similar, however, the inclusion of 17 additional genomes, as expected, influenced the distribution of the type strains, especially the inclusion of the six additional type strains. Nevertheless, three out of Carro's five defined groups (I, IV and V) were almost completely recovered in the present phylogenomic analysis, the major rearrangements were observed in Carro's groups II and III. In the present phylogenomic analysis, the strains in group II (*M. purpureochromogenes*, *M. coxensis* and *M. halophytica*) fused with *M. rifamycinica* and *M. matsumotoense* (group III) and were joined by *M. wenchangensis* (new to the analysis). The instability of group III was already highlighted (Carro et al., 2018). This rearrangement reduced group III to *M. olivasterospora, M. carbonacea and M. haikouensis*.

Micromonospora pattaloongensis DSM 45245^T and *M. pisi* DSM 45175^T, which were not included in the previous work (Carro et al., 2018), were recovered as a separate cluster, with high distance length between them and the rest of members of the genus *Micromonospora.* This separation was already seen in the MLSA analysis (Figure 21). *Salinispora arenicola* CNH643^T and *Salinispora pacifica* CNR114^T were also recovered as a separate cluster, between the *M. pisi-M. pattaloongensis* cluster and the remaining members of the genus *Micromonospora.*

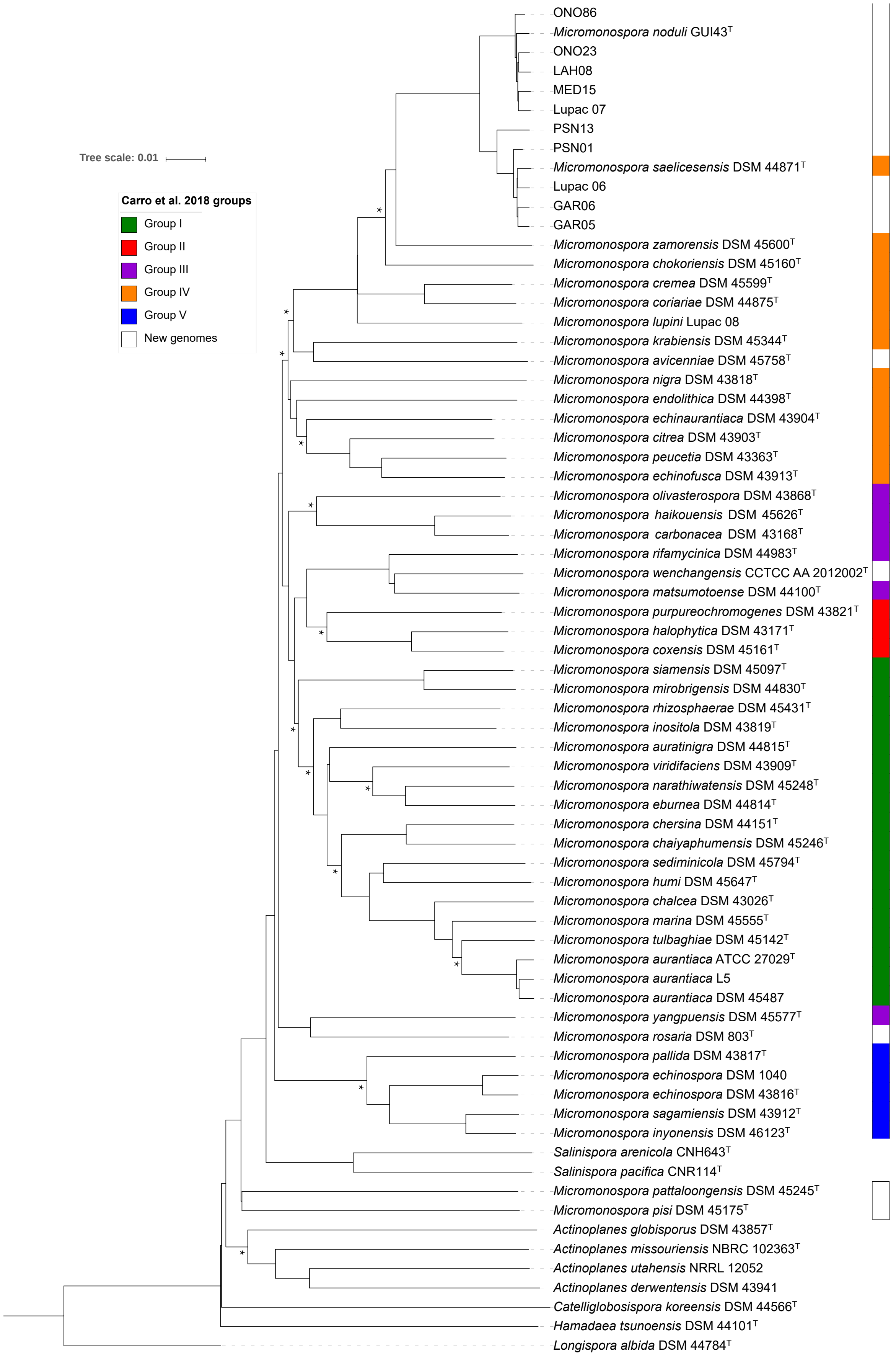

Tree scale: 0.01
Carro et al. 2018 groups
🟩 Group I
🟥 Group II
🟪 Group III
🟧 Group IV
🟦 Group V
□ New genomes
ONO86
Micromonospora noduli GUI43^{T}
ONO23
LAH08
MED15
Lupac 07
PSN13
PSN01
Micromonospora saelicesensis DSM 44871^{T}
Lupac 06
GAR06
GAR05
Micromonospora zamorensis DSM 45600^{T}
Micromonospora chokoriensis DSM 45160^{T}
Micromonospora cremea DSM 45599^{T}
Micromonospora coriariae DSM 44875^{T}
Micromonospora lupini Lupac 08
Micromonospora krabiensis DSM 45344^{T}
Micromonospora avicenniae DSM 45758^{T}
Micromonospora nigra DSM 43818^{T}
Micromonospora endolithica DSM 44398^{T}
Micromonospora echinaurantiaca DSM 43904^{T}
Micromonospora citrea DSM 43903^{T}
Micromonospora peucetia DSM 43363^{T}
Micromonospora echinofusca DSM 43913^{T}
Micromonospora olivasterospora DSM 43868^{T}
Micromonospora haikouensis DSM 45626^{T}
Micromonospora carbonacea DSM 43168^{T}
Micromonospora rifamycinica DSM 44983^{T}
Micromonospora wenchangensis CCTCC AA 2012002^{T}
Micromonospora matsumotoense DSM 44100^{T}
Micromonospora purpureochromogenes DSM 43821^{T}
Micromonospora halophytica DSM 43171^{T}
Micromonospora coxensis DSM 45161^{T}
Micromonospora siamensis DSM 45097^{T}
Micromonospora mirobrigensis DSM 44830^{T}
Micromonospora rhizosphaerae DSM 45431^{T}
Micromonospora inositola DSM 43819^{T}
Micromonospora auratinigra DSM 44815^{T}
Micromonospora viridifaciens DSM 43909^{T}
Micromonospora narathiwatensis DSM 45248^{T}
Micromonospora eburnea DSM 44814^{T}
Micromonospora chersina DSM 44151^{T}
Micromonospora chaiyaphumensis DSM 45246^{T}
Micromonospora sediminicola DSM 45794^{T}
Micromonospora humi DSM 45647^{T}
Micromonospora chalcea DSM 43026^{T}
Micromonospora marina DSM 45555^{T}
Micromonospora tulbaghiae DSM 45142^{T}
Micromonospora aurantiaca ATCC 27029^{T}
Micromonospora aurantiaca L5
Micromonospora aurantiaca DSM 45487
Micromonospora yangpuensis DSM 45577^{T}
Micromonospora rosaria DSM 803^{T}
Micromonospora pallida DSM 43817^{T}
Micromonospora echinospora DSM 1040
Micromonospora echinospora DSM 43816^{T}
Micromonospora sagamiensis DSM 43912^{T}
Micromonospora inyonensis DSM 46123^{T}
Salinispora arenicola CNH643^{T}
Salinispora pacifica CNR114^{T}
Micromonospora pattaloongensis DSM 45245^{T}
Micromonospora pisi DSM 45175^{T}
Actinoplanes globisporus DSM 43857^{T}
Actinoplanes missouriensis NBRC 102363^{T}
Actinoplanes utahensis NRRL 12052
Actinoplanes derwentensis DSM 43941
Catelliglobosispora koreensis DSM 44566^{T}
Hamadaea tsunoensis DSM 44101^{T}
Longispora albida DSM 44784^{T}

Figure 25: Whole genome-sequence based phylogenomic tree constructed with the GBDP tool (see main text for details). Colors on the right side represent groups described in Carro et al. (2018). Asterisks represent conserved nodes between this tree and the UBCG genome phylogenetic tree.

3.4.7- UBCG phylogenomic analysis

The same dataset as above (Appendix III) was used to construct a phylogenetic tree based on a core genome set of 92 genes using the UBCG tool (Na et al., 2018). Most of the UBCG selected genes (67/92) fall in the translation COG category (J), coding for ribosomal proteins (25/92, 50S and 18/92, 30S), aminoacid-tRNA ligases (10/92) and elongation and initiator factors (4/92).

Again, the ten strains were distributed in two groups of identical composition as that of the *gyr*B gene, MLSA and whole genome phylogenomic analyses with significant branch support as indicated by the bootstrap values and gene support indices (GSI) (Figure 26). GSI values indicate the number of individual gene trees that support a node (up to 92 genes) (Na et al., 2018).

The topology of this tree with respect to the composition of the two groups was the same as the whole genome, MLSA and *gyr*B gene trees, including the position of strain Lupac 07, recovered in the *M. noduli* group. The topology of the UBCG tree highly correlated to the topology of the whole genome phylogenomic tree of this study. Especially interesting was the fact that the new redefined groups II and III were recovered in their entirety together with groups I, IV and V. In this analysis, a new group that contained strains from Carro's groups I (*M. mirobrigensis* and *M. siamensis*), III (*M. yangpuensis*) and IV (*M. krabiensis*), in addition to the newly included type strains *M. avicenniae* and *M. rosaria*, was formed (Figure 26). Another important difference between the GBDP and UBCG trees of this study was the position of *Salinispora pacifica* and *Salinispora arenicola* which in the latter tree was found associated to group IV. In this case, the up-to-date bacterial core gene analysis was not resolutive.

The same topology as MLSA and whole genome trees with respect to *M. pattaloongensis* DSM 45245^T and *M. pisi* DS; 45175^T was observed in the UBCG tree. These two isolates were recovered in a separate distant cluster, with very high support values (90 GSI and 100% bootstrap).

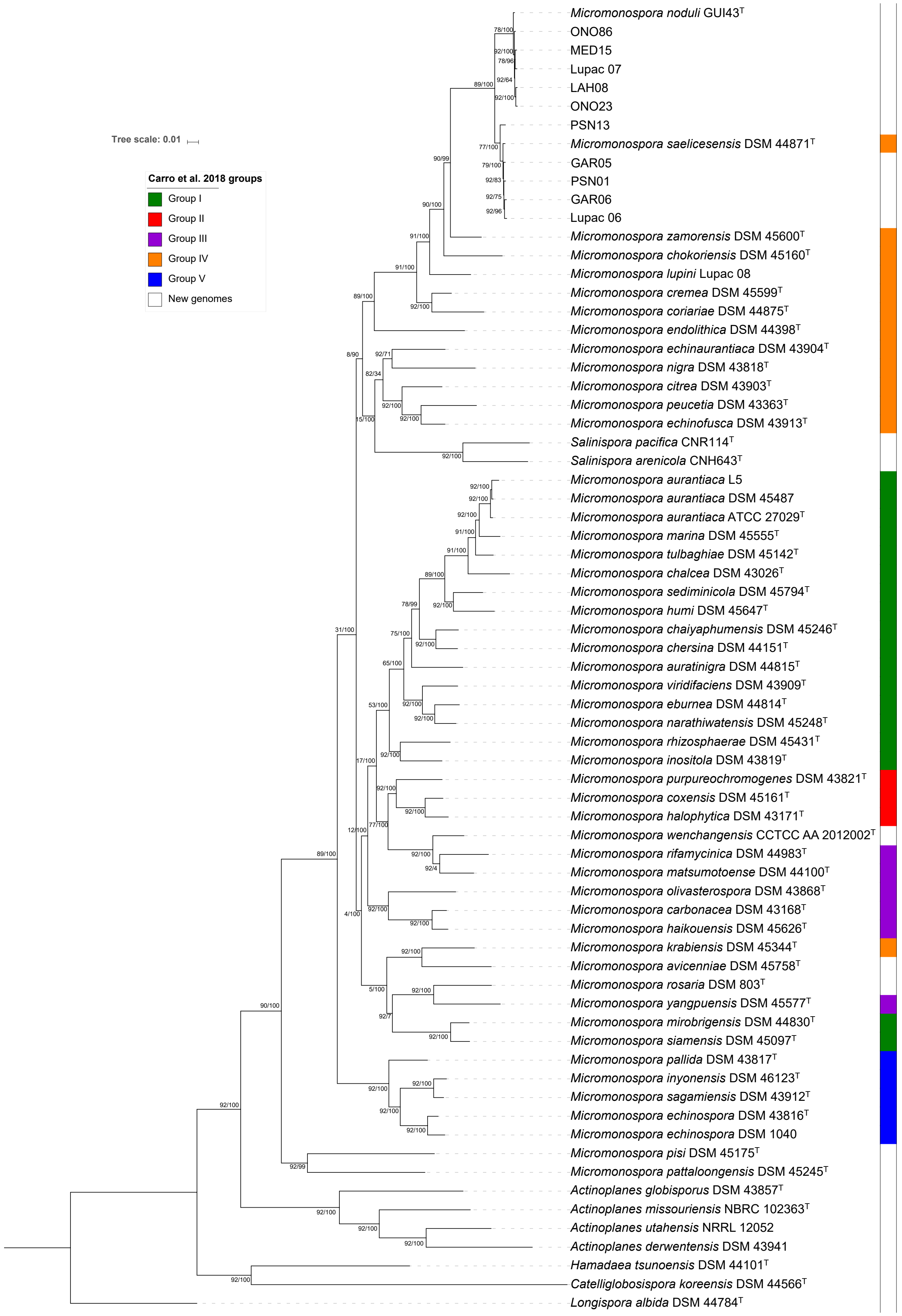

Tree scale: 0.01
Carro et al. 2018 groups
🟩 Group I
🟥 Group II
🟪 Group III
🟧 Group IV
🟦 Group V
□ New genomes
Micromonospora noduli GUI43^{T}
ONO86
MED15
Lupac 07
LAH08
ONO23
PSN13
Micromonospora saelicesensis DSM 44871^{T}
GAR05
PSN01
GAR06
Lupac 06
Micromonospora zamorensis DSM 45600^{T}
Micromonospora chokoriensis DSM 45160^{T}
Micromonospora lupini Lupac 08
Micromonospora cremea DSM 45599^{T}
Micromonospora coriariae DSM 44875^{T}
Micromonospora endolithica DSM 44398^{T}
Micromonospora echinaurantiaca DSM 43904^{T}
Micromonospora nigra DSM 43818^{T}
Micromonospora citrea DSM 43903^{T}
Micromonospora peucetia DSM 43363^{T}
Micromonospora echinofusca DSM 43913^{T}
Salinispora pacifica CNR114^{T}
Salinispora arenicola CNH643^{T}
Micromonospora aurantiaca L5
Micromonospora aurantiaca DSM 45487
Micromonospora aurantiaca ATCC 27029^{T}
Micromonospora marina DSM 45555^{T}
Micromonospora tulbaghiae DSM 45142^{T}
Micromonospora chalcea DSM 43026^{T}
Micromonospora sediminicola DSM 45794^{T}
Micromonospora humi DSM 45647^{T}
Micromonospora chaiyaphumensis DSM 45246^{T}
Micromonospora chersina DSM 44151^{T}
Micromonospora auratinigra DSM 44815^{T}
Micromonospora viridifaciens DSM 43909^{T}
Micromonospora eburnea DSM 44814^{T}
Micromonospora narathiwatensis DSM 45248^{T}
Micromonospora rhizosphaerae DSM 45431^{T}
Micromonospora inositola DSM 43819^{T}
Micromonospora purpureochromogenes DSM 43821^{T}
Micromonospora coxensis DSM 45161^{T}
Micromonospora halophytica DSM 43171^{T}
Micromonospora wenchangensis CCTCC AA 2012002^{T}
Micromonospora rifamycinica DSM 44983^{T}
Micromonospora matsumotoense DSM 44100^{T}
Micromonospora olivasterospora DSM 43868^{T}
Micromonospora carbonacea DSM 43168^{T}
Micromonospora haikouensis DSM 45626^{T}
Micromonospora krabiensis DSM 45344^{T}
Micromonospora avicenniae DSM 45758^{T}
Micromonospora rosaria DSM 803^{T}
Micromonospora yangpuensis DSM 45577^{T}
Micromonospora mirobrigensis DSM 44830^{T}
Micromonospora siamensis DSM 45097^{T}
Micromonospora pallida DSM 43817^{T}
Micromonospora inyonensis DSM 46123^{T}
Micromonospora sagamiensis DSM 43912^{T}
Micromonospora echinospora DSM 43816^{T}
Micromonospora echinospora DSM 1040
Micromonospora pisi DSM 45175^{T}
Micromonospora pattaloongensis DSM 45245^{T}
Actinoplanes globisporus DSM 43857^{T}
Actinoplanes missouriensis NBRC 102363^{T}
Actinoplanes utahensis NRRL 12052
Actinoplanes derwentensis DSM 43941
Hamadaea tsunoensis DSM 44101^{T}
Catelliglobosispora koreensis DSM 44566^{T}
Longispora albida DSM 44784^{T}
78/100
92/100
78/96
92/64
89/100
92/100
77/100
90/99
79/100
92/83
92/75
90/100
92/96
91/100
91/100
89/100
92/100
8/90
92/71
82/34
92/100
92/100
15/100
92/100
92/100
92/100
92/100
91/100
91/100
89/100
78/99
92/100
31/100
75/100
92/100
65/100
53/100
92/100
92/100
17/100
92/100
92/100
92/100
77/100
12/100
89/100
92/100
92/4
4/100
92/100
92/100
92/100
90/100
5/100
92/100
92/7
92/100
92/100
92/100
92/100
92/100
92/100
92/99
92/100
92/100
92/100
92/100

Figure 26: Up-to-date bacterial core gene phylogenomic tree reconstructed with 92 bacterial core genes. Tree has been formatted using UBCG_iTOL_maker and iTOL platform. Colors on the right represent groups described in Carro et al. (2018). GSI support (left) and bootstrap values (right) are given at nodes.

3.4.8- Phenotypic profiles

Thirty-one phenotypic tests reported previously to be useful for the differentiation of the species *M. saelicesensis* and *M. noduli* (Carro et al., 2016) were carried out with all test strains. The number of characteristics that phenotypically differentiated between the two species was significantly reduced to one test when the number of strains compared increased (Appendix VII). Specifically, the use of rhamnose as carbon source was positive for all strains in *M. noduli,* and negative for *M. saelicesensis* group, with the exception of isolate PSN13, which was positive. The results of the remaining test varied at strain level and did not relate to their species identification (Table 10).

Intra-species variability ranged from 0 to 33.3% for all the phenotypical traits essayed. Range of pH growth and lipase production were the most variable test in *M. saelicesensis* (Group I). Utilization of serine as carbon source, degradation of tyrosine and pH growth range were the most variable for the *M. noduli* group.

Phenotypic profiles using the Biolog system was also determined for all the strains in group I and II (Appendix VI). In this case, none of the 71 carbon sources or the 23 biochemical tests served to differentiate between the two species, given the variability observed among the duplicate tests. The isolate Lupac 06 was the most variable with 35.1% of discrepancies recorded (marked as conflictive results in Appendix VI). Overall intraspecies variability for *M. saelicesensis* and *M. noduli* groups was 25.5% and 26.6% respectively.

Table 10: Differential phenotypic characteristics between *M. saelicesensis* and *M. noduli* as reported by Carro and colleagues (Carro et al., 2016) +, Positive; -, Negative; w, Weak.

	***M. saelicesensis* (Group I)**						**intra-species variability *M. saelicesensis***	***M. noduli* (Group II)**						**intra-species variability *M. noduli***
	Lupac 09^T	**Lupac 06**	**GAR05**	**GAR06**	**PSN01**	**PSN13**		**GUI43^T**	**MED15**	**ONO23**	**ONO86**	**Lupac 07**	**LAH08**	
APIZYM														
Alkaline phosphatase	+	+	+	+	+	+	0%	+	w	+	-	+	w	16.67%
Lipase (C 14)	-	w	-	+	-	-	33.33%	-	-	-	-	-	-	0%
Acid phosphatase	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Naph- thol-AS-BI- Phosphohydrolase	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
α-galactosidase	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
α-glucosidase	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
β-glucosidase	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
α-mannosidase	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%
α-fucosidase	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%

Carbon sources														
Alanine	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Arginine	+	+	+	+	+	+	0%	-	+	+	+	+	+	16.67%
Gluconate	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Histidine	+	+	+	+	+	+	0%	-	+	+	+	+	+	16.67%
Lysine	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%
Melezitose	+	+	+	+	+	+	0%	-	+	+	+	+	+	16.67%
Proline	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%
Rhamnose	-	-	-	-	-	+	16.67%	+	+	+	+	+	+	0%
Salicin	+	+	+	+	+	+	0%	w	+	+	+	+	w	0%
Serine	-	-	-	-	-	-	0%	-	-	w	w	-	-	33.33%
Sorbose	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%
Starch	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Trehalose	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Valine	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%
Xylose	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Maximum NaCl tolerance	1	3	3	3	3	3	16.67%	3	3	3	3	3	3	0%
pH tolerance	6-Sep	6.5-9	7-9	6-9	6-9	6-9	33.33%	6-9	7-9	6-9	6-9	6-9	6.5-9	33.33%
Degradation of														
Starch	+	+	+	+	+	+	0%	+	+	+	+	+	+	0%
Tween 20	-	-	-	-	-	+	16.67%	-	-	-	-	-	-	0%
Tween 80	-	-	-	-	-	+	16.67%	-	-	-	-	-	-	0%
Tyrosine	-	-	+	-	-	-	16.67%	+	+	-	+	-	+	33.33%
Urea	-	-	-	-	-	-	0%	-	-	-	-	-	-	0%

Nineteen carbon sources were also assayed at different times (2007, 2016 and 2017) to check for reproducibility. Nine of the eleven strains tested expressed discrepant results over the different testing times. Three strains (Lupac 09^T, Lupac 06 and Lupac 07) showed the highest variation with 26% of the tests yielding conflicting results while MED15, LAH08 and PSN13 had the lowest variation (5.2%). The use of *D*-serine as carbon source was the least reproducible test with seven strains yielding conflicting results (Appendix VII).

Draft genomes of the test strains were screened for genes involved in the carbon metabolism of the corresponding 19 substrates assayed *in vitro*. The predicted phenotypes correlated 100% with the results obtained in the laboratory for 11 tests. However, in the case of *L*-alanine, *L*-arginine, *L*-histidine, *L*-lysine, *myo*-inositol, *L*-rhamnose, *D*-serine and *D*-trehalose, discrepant results were found between *in-vitro* results and *in-silico* predictions (Figure 27). For *L*-alanine, *myo*-inositol and *D*-trehalose, the genes were localized in the genome, but the experimental results varied (with positive or negative results), suggesting that even if the test are carried out in the same laboratory, under the same conditions, they were not 100% reproducible.

In the case of *L*-rhamnose, *in vitro* tests for strain GUI43^T were positive but the genes related to the metabolism of this compound were not located. This is probably explained by the fact that draft-genomes were used, and interpretation of genomic data should be done with precaution.

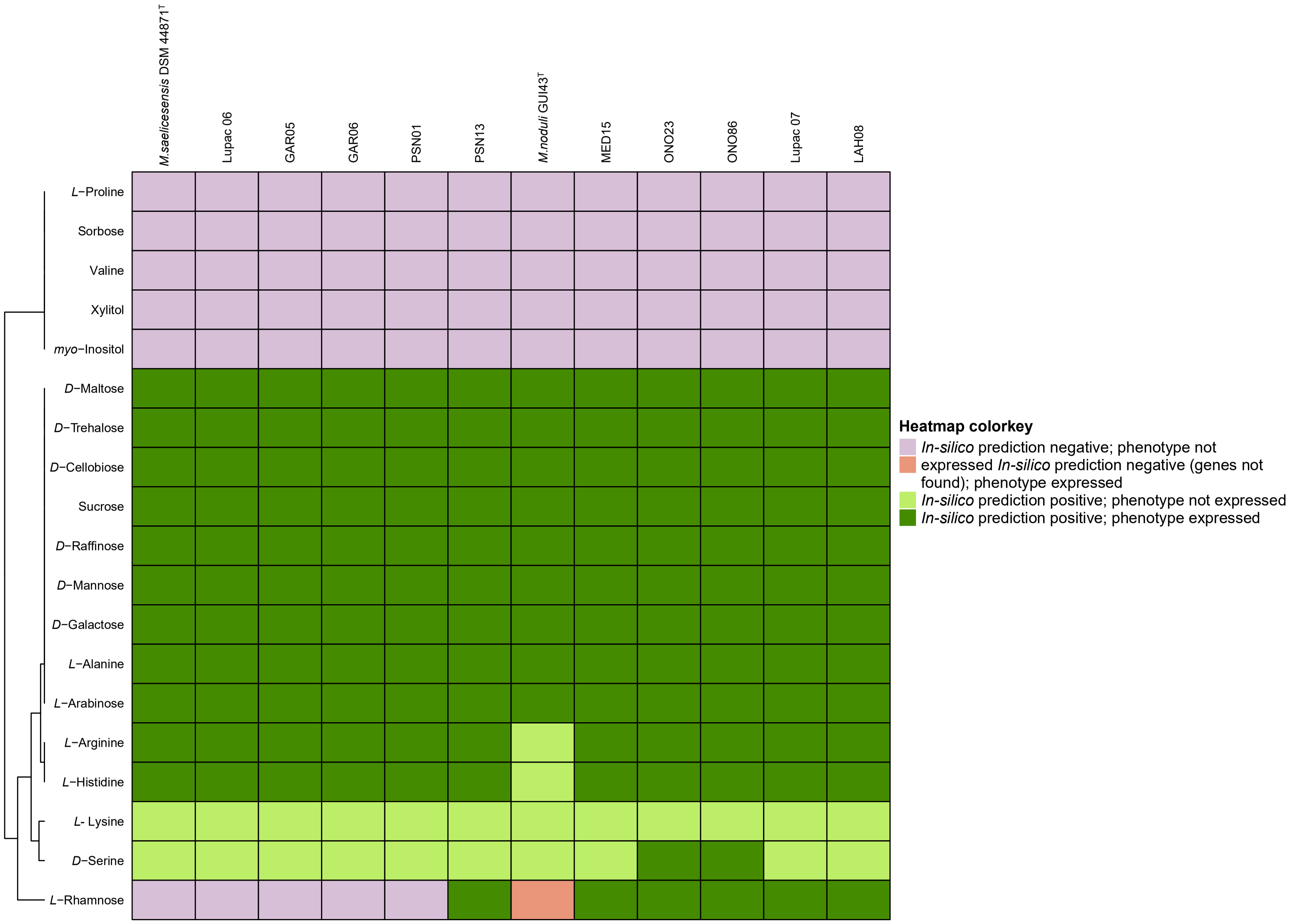


Figure 27: Predicted phenotypes vs. experimental phenotypic data based on 19 carbon source substrates. *In silico* prediction negative, phenotype not expressed (purple); *in silico* prediction negative (genes not found), phenotype expressed (red); *in silico* prediction positive, phenotype not expressed (light green) and *in silico* prediction positive, phenotype expressed in at least one duplicate (green).

3.5- DISCUSSION

The genus *Micromonospora* is highly relevant in biotechnological applications in areas such as medicine, agriculture and biofuels (Carro et al., 2018; Hirsch and Valdés, 2010; Trujillo et al., 2015). Within this framework, DNA-DNA hybridization (DDH) has been considered the key test to decide if a new strain represents a new species, despite its well spelled limitations. Given the drawbacks of DDH, it is not always straight forward to delineate the species limits, especially when DDH values are close to the threshold. Therefore, the development of whole genome sequencing seems more appropriate to deduce relatedness by comparing genome sequences rather than performing DDH experiments (Vandamme and Peeters, 2014). Genomic data was recently used as the backbone to revisit the classification of the genus *Micromonospora* using a set of 45 draft genomes providing a useful dataset for comparison (Carro et al., 2018).

While 16S rRNA is limited in resolving phylogenetic relationships at the species level (Carro et al., 2018; Hahnke et al., 2016; Katayama et al., 2007; Na et al., 2018), it has provided a good starting point for taxonomic studies. The sequence similarity values obtained for the study strains indicated that *M. saelicesensis* or *M. noduli* were the two most closely related species although in some cases, similarity values were identical between the study and both type strains (e.g. GAR06 and LAH08). In the present study 16S rRNA gene phylogeny was not sufficient to clearly determine the relationship of the strains identified as either *M. noduli* or *M. lupini*; the position of strain MED01 was the most uncertain as it was recovered between the previously mentioned species. Nevertheless, the remaining five isolates (NIE79, NIE111, PSH03, PSH25 and LAH09) were recovered distantly away from the *M. noduli* and *M. lupini* clusters and showing closer relationship with other species of *Micromonospora*.

The use of *gyr*B gene sequences to resolve phylogenetic relationships in the genus *Micromonospora* has been recommended by several authors (Carro et al., 2012; Garcia et al., 2010; Kasai et al., 2000) given its higher resolution when compared to 16S rRNA gene phylogeny. In this study, the *gyr*B gene tree topology showed a similar arrangement to the 16S tree with respect to the test strains, however several differences were observed. The branch lengths were slightly longer, but still very small when compared to the rest of the *Micromonospora* species included in the tree. The most relevant change was the position of strain Lupac 07, which, together with strains Lupac 06 and Lupac 09^T were originally classified as *M. saelicesensis*. The latter strains remained in the *M. saelicesensis* cluster but Lupac 07 moved to the *M. noduli* group. In addition, isolate MED01 was recovered as an independent strain, close to *M. noduli* group. Isolates NIE79, NIE111, PSH03 and PSH25 were again recovered as independent strains, but closer to *M. saelicesensis* and *M. noduli* groups. The strain LAH09 was recovered in a distant branch, being closer to *M. zamorensis* DMS 45600^T. As expected, topologies of both trees in relation to the type strains were very different confirming that phylogenies based on single genes are very limited and unstable, making identification of nearest phylogenetic neighbors difficult.

Multi-locus sequence approach (MLSA), based on a concatenated alignment of five core genes, served as an intermediate approach between single gene phylogeny and whole genome approaches. MLSA tree topology showed a similar arrangement to that of *gyr*B gene tree, including the position of strain Lupac 07, recovered in the *M. noduli* cluster. The topology of this tree also positioned the strains MED01, NIE79, NIE111, PSH03, PSH25 and LAH09 as independent isolates, away from *M. saelicesensis* and *M. noduli* clusters.

The tree topologies based on the phylogenomic analyses of the UBCG (92 genes) and the whole draft genomes were similar. In both trees, strain Lupac 07 was recovered in the *M. noduli* group, strongly suggesting that this strain should be reclassified as a member of this species. The remaining 9 strains were recovered in the same species groups throughout all analyses.

In this study, both phylogenomic analyses contained a total of 70 genomes, including six additional *Micromonospora* type strains. Overall, good agreement was found between the two phylogenies of this work and recently published data. In all cases, groups I, IV and V previously defined (Carro et al., 2018) were recovered in their entirety with *M. avicenniae* joining group IV. The main difference between the three phylogenies was the composition of Carro's groups II and III which were clearly influenced by the addition of *M. rosaria* DSM 803^T and *M. wenchangensis* CCTCC AA 2012002^T, producing a new group recovered in both phylogenies of the present work. Nevertheless, the groups I, IV and V remained very stable considering that eleven new genomes (*M. noduli* GUI43^T and ten test strains) were added and these were assigned to group IV where *M. saelicesensis* DSM 44871^T was originally assigned. These rearrangements reinforce the argument that classification and identification systems are data dependent and constant rearrangement should be expected as more data are added and alternative methods are applied (Carro et al., 2018).

The new analysis tool UBCG proved useful for the construction of phylogenomic analysis, showing good correlation with trees using whole-draft genome data even though it did not resolve well the position of the *Salinispora* representatives, however, this may be due to the small number of representatives in the data set. An advantage of this pipeline is the use of bootstrap and GSI values to support the phylogenetic branches. It is also expected that as more genome sequences are added to the database, the more resolutive it should become.

Genome relatedness indices (ANI, Ortho-ANI and dDDH) were calculated to complement the phylogenomic analyses for species demarcation. Overall, the three methods showed good agreement and the two species groups defined in the *gyr*B, MLSA, core-genome and whole-genome phylogenetic analyses supported the recognition of the ten strains in two species.

Furthermore, these studies served to highlight the close relationship between the species *M. saelicesensis* and *M. noduli.* ANI values proposed for species delineation have been set to 95-96% as this range has been found to be correlated with the experimental DDH threshold of 70% (Goris et al., 2007; Richter and Rosselló-Móra, 2009). An alternative means to measure relatedness between two genomes is the calculation of dDDH using the GBDP method which appears to show a better correlation than ANI to the data derived from DDH experiments (Auch et al., 2010; Meier-Kolthoff et al., 2013; Peeters et al., 2016).

In this work, the OGRI values were slightly above the recommended threshold for species delineation, if strictly applied, the study strains should be recognized as members of the same species. However, the consideration of other results in this work support the recognition of the strains as two separate species, *M. saelicesensis* and *M. noduli*. As previously expressed, thresholds are necessary for guidance but these should be applied in a flexible manner and considering other biological properties (Li et al., 2015). The present work is a good example for the interpretation and application of these values. Finally, OGRI values for isolates LAH09, MED01, PSH03, PSH25, NIE79 and NIE 111 strongly supports the proposal of six new species represented by these strains.

The use of phenotypic traits to identify and differentiate species in prokaryotic systematics is of limited value as previously discussed (Amaral et al., 2014; Sutcliffe et al., 2013; Vandamme and Peeters, 2014). Several strains identified as one species, expressed different phenotypes, highlighting the problem of using diagnostic tables based on single strains to list differential characteristics between species. Information about intra-species variation is crucial for the development of stable diagnostic characteristics and the convenience of using more than a single isolate have been previously discussed (Oren and Garrity, 2014; Sutcliffe et al., 2012).

Our results confirm that the use of phenotypic tests, even when performed under the same conditions are not reliable for species differentiation due to the high variability observed within several members of the same species (Kumar et al., 2015). Instead, phenotypic studies should be regarded as complementary information to understand the biology of a microorganism and they should be restricted to strain characterization. Understandably, the inclusion of additional strains for the description of a taxon is often regarded as a burden because a lot of extra work is needed, especially when looking for differential phenotypic tests with questionable taxonomic value (Oren and Garrity, 2014; Sutcliffe et al., 2012).

Genomic information can be used to determine the intrinsic variability between a set of strains based on the core and pangenome profiles (Coenye et al., 2005; Oren and Garrity, 2014; Sutcliffe et al., 2012). The calculation of these parameters has pointed out an important degree of variation between the species *M. saelicesensis* and *M. noduli* supporting their recognition as separate taxa. The complete elucidation of the gene functions within each group may provide an initial set of stable differential characteristics for each species, some of which may be phenotypically expressed.

3.6- CONCLUSIONS

1. Genome-based classifications should become more stable as additional data is generated, providing a new working frame for the systematics of prokaryotes.
2. The use of a diverse array of methods and the use of more than one isolate are of great importance for the characterization of intra-species variation.
3. OGRI values and especially dDDH values seem very appropriate for the delineation of prokaryotic species, but threshold numbers should be applied with a level of flexibility and considering other features inherent to a microorganism such as ecology, physiology, etc.
4. Phenotypic information is useful to complement strain characterization. However, these studies should aim to provide information on the biology of a microorganism and not only to fill out a table with results of questionable value.

3.7- REFERENCES

Amaral, G. R. S., Dias, G. M., Wellington-Oguri, M., Chimetto, L., Campeão, M. E., Thompson, F. L., et al. (2014). Genotype to phenotype: Identification of diagnostic *Vibrio* phenotypes using whole genome sequences. *Int. J. Syst. Evol. Microbiol.* 64, 357–365. doi:10.1099/ijs.0.057927-0.

Auch, A. F., von Jan, M., Klenk, H. P., and Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Benito, P. (2020). Comparative proteomic and transcriptomic profiling of *Micromonopora* strains associated with legumes.

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., et al. (2016). EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.* 44, W22–W28. doi:10.1093/nar/gkw255.

Carro, L. (2009). Avances en la Sistemática del Género *Micromonospora:* Estudio de Cepas aisladas de la Rizosfera y Nódulos de *Pisum sativum.*

Carro, L., Nouioui, I., Sangal, V., Meier-Kolthoff, J. P., Trujillo, M. E., Montero-Calasanz, M. del C., et al. (2018). Genome-based classification of *Micromonosporae* with a focus on their biotechnological and ecological potential. *Sci. Rep.* 8, 525. doi:10.1038/s41598-017-17392-0.

Carro, L., Pujic, P., Trujillo, M. E., and Normand, P. (2013). *Micromonospora* is a normal occupant of actinorhizal nodules. *J. Biosci.* 38, 685–693. doi:10.1007/s12038-013-9359-y.

Carro, L., Riesco, R., Spröer, C., and Trujillo, M. E. (2016). *Micromonospora ureilytica* sp. nov., *Micromonospora noduli* sp. nov. and *Micromonospora vinacea* sp. nov., isolated from *Pisum sativum* nodules. *Int. J. Syst. Evol. Microbiol.* 66, 3509–3514. Available at: https://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.001231.

Carro, L., Spröer, C., Alonso, P., and Trujillo, M. E. (2012). Diversity of *Micromonospora* strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. *Syst. Appl. Microbiol.* 35, 73–80. doi:10.1016/j.syapm.2011.11.003.

Cerda, M. E. (2008). Aislamiento de *Micromonospora* de Nódulos de Leguminosas Tropicales y Análisis de Su Interés Como Promotor del Crecimiento Vegetal.

Christensen, W. B. (1946). Urea Decomposition as a Means of Differentiating *Proteus* and *Paracolon* Cultures from Each Other and from *Salmonella and Shigella* Types. *J. Bacteriol.* 52, 461–466.

Chun, J., and Rainey, F. A. (2014). Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea. Int. J. Syst. Evol. Microbiol.* 64, 316–324. doi:10.1099/ijs.0.054171-0.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/nar/gkv1276.

Coenye, T., Gevers, D., Van De Peer, Y., Vandamme, P., and Swings, J. (2005). Towards a prokaryotic genomic taxonomy. *FEMS Microbiol. Rev.* 29, 147–167. doi:10.1016/j.femsre.2004.11.004.

Colwell, R. R. (1970). Polyphasic taxonomy of the genus *Vibrio:* numerical taxonomy of *Vibrio cholerae, Vibrio parahaemolyticus*, and related *Vibrio* species. *J. Bacteriol.* 104, 410–433. Available at: https://www.ncbi.nlm.nih.gov/pubmed/5473901.

de la Vega, P. A. (2010). Distribución, caracterización e importancia ecologica de *Micromonospora* en nódulos fijadores de nitrogeno de *Lupinus.*

de Menezes, A. B., McDonald, J. E., Allison, H. E., and McCarthy, A. J. (2012). Importance of *Micromonospora* spp. as colonizers of cellulose in freshwater lakes as demonstrated by quantitative reverse transcriptase PCR of 16s rRNA. *Appl. Environ. Microbiol.* 78, 3495–3499. doi:10.1128/AEM.07314-11.

Durot, M., Bourguignon, P. Y., and Schachter, V. (2009). Genome-scale models of bacterial metabolism: Reconstruction and applications. *FEMS Microbiol. Rev.* 33, 164–190. doi:10.1111/j.1574-6976.2008.00146.x.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi:10.1007/BF01734359.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution (N. Y).* 39, 783–791. doi:10.2307/2408678.

Garcia, L. C., Martinez-Molina, E., and Trujillo, M. E. (2010). *Micromonospora pisi* sp. nov., isolated from root nodules of *Pisum sativum. Int. J. Syst. Evol. Microbiol.* 60, 331–337. doi:10.1099/ijs.0.012708-0.

Genilloud, O. (2015a). *Micromonospora. Bergey's Man. Syst. Archaea Bact.* doi:doi:10.1002/9781118960608.gbm00148.

Genilloud, O. (2015b). *Micromonosporaceae. Bergey's Man. Syst. Archaea Bact.* doi:doi:10.1002/9781118960608.fbm00041.

Genilloud, O. (2015c). *Micromonosporales ord. nov. Bergey's Man. Syst. Archaea Bact.*, 1. doi:doi:10.1002/9781118960608.obm00015.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0.

Hahnke, R. L., Meier-Kolthoff, J. P., García-López, M., Mukherjee, S., Huntemann, M., Ivanova, N. N., et al. (2016). Genome-based taxonomic classification of *Bacteroidetes. Front. Microbiol.* 7. doi:10.3389/fmicb.2016.02003.

Hirsch, A. M., and Valdés, M. (2010). *Micromonospora:* An important microbe for biomedicine and potentially for biocontrol and biofuels. *Soil Biol. Biochem.* 42, 536–542. doi:10.1016/j.soilbio.2009.11.023.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.

Kasai, H., Tamura, T., and Harayama, S. (2000). Intrageneric relationships among *Micromonospora* species deduced from *gyr*B-based phylogeny and DNA relatedness. *Int. J. Syst. Evol. Microbiol.* 50, 127–134. doi:10.1099/00207713-50-1-127.

Katayama, T., Tanaka, M., Moriizumi, J., Nakamura, T., Brouchkov, A., Douglas, T. A., et al. (2007). Phylogenetic analysis of bacteria preserved in a permafrost ice wedge for 25,000 years. *Appl. Environ. Microbiol.* 73, 2360–2363. doi:10.1128/AEM.01715-06.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi:10.1007/BF01731581.

Kroppenstedt, R. M., Mayilraj, S., Wink, J. M., Kallow, W., Schumann, P., Secondini, C., et al. (2005). Eight new species of the genus *Micromonospora, Micromonospora citrea* sp. nov., *Micromonospora echinaurantiaca* sp. nov., *Micromonospora echinofusca* sp. nov. *Micromonospora fulviviridis* sp. nov., *Micromonospora inyonensis* sp. nov., *Micromonospora peucetia* s. *Syst. Appl. Microbiol.* 28, 328–339. doi:10.1016/j.syapm.2004.12.011.

Kumar, N., Lad, G., Giuntini, E., Kaye, M. E., Udomwong, P., Jannah Shamsani, N., et al. (2015). Bacterial genospecies that are not ecologically coherent: Population genomics of *Rhizobium leguminosarum. Open Biol.* 5, 140133–140133. doi:10.1098/rsob.140133.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Kutzner, H. J. (1981). The family *Streptomycetaceae. The Prokaryotes*. Available at: https://ci.nii.ac.jp/naid/10019908720/en/.

Lee, H. J., and Whang, K. S. (2017). *Micromonospora fulva* sp. nov., isolated from forest soil. *Int. J. Syst. Evol. Microbiol.* 67, 1746–1751. doi:10.1099/ijsem.0.001858.

Lee, I., Kim, Y. O., Park, S. C., and Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi:10.1099/ijsem.0.000760.

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. *Mol. Biol. Evol.* 32, 2798–2800. doi:10.1093/molbev/msv150.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi:10.1093/nar/gkw290.

Li, X., Huang, Y., and Whitman, W. B. (2015). The relationship of the whole genome sequence identity to DNA hybridization varies between genera of prokaryotes. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 107, 241–249. doi:10.1007/s10482-014-0322-1.

Martínez-Hidalgo, P., Olivares, J., Delgado, A., Bedmar, E., and Martínez-Molina, E. (2014). Endophytic *Micromonospora* from *Medicago sativa* are apparently not able to fix atmospheric nitrogen. *Soil Biol. Biochem.* 74, 201–203. doi:10.1016/j.soilbio.2014.03.011.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., Göker, M., and Access, O. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., Hahnke, R. L., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., et al. (2014a). Complete genome sequence of DSM 30083^T, the type strain (U5/41^T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand. Genomic Sci.* 9, 2. doi:10.1186/1944-3277-9-2.

Meier-Kolthoff, J. P., Klenk, H.-P., and Göker, M. (2014b). Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. *Int. J. Syst. Evol. Microbiol.* 64, 352–356. doi:https://doi.org/10.1099/ijs.0.056994-0.

Mendoza, S. N., Olivier, B. G., Molenaar, D., and Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* 20, 158. doi:10.1186/s13059-019-1769-1.

Na, S.-I., Kim, Y. O., Yoon, S.-H., Ha, S., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi:10.1007/s12275-018-8014-6.

Oren, A., and Garrity, G. M. (2014). Then and now: A systematic review of the systematics of prokaryotes in the last 80 years. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 106, 43–56. doi:10.1007/s10482-013-0084-1.

Peeters, C., Meier-Kolthoff, J. P., Verheyde, B., De Brandt, E., Cooper, V. S., and Vandamme, P. (2016). Phylogenomic study of *Burkholderia glathei*-like organisms, proposal of 13 novel *Burkholderia* species and emended descriptions of *Burkholderia sordidicola, Burkholderia zhejiangensis*, and *Burkholderia grimmiae. Front. Microbiol.* 7, 1–19. doi:10.3389/fmicb.2016.00877.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Sierra, G. (1957). A simple method for the detection of lipolytic activity of microorganisms and some observations on the influence of the contact between cells and fatty substrates. *Antonie Van Leeuwenhoek* 23, 15–22. doi:10.1007/BF02545855.

Sutcliffe, I. C., Trujillo, M. E., and Goodfellow, M. (2012). A call to arms for systematists: Revitalising the purpose and practises underpinning the description of novel microbial taxa. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 101, 13–20. doi:10.1007/s10482-011-9664-0.

Sutcliffe, I. C., Trujillo, M. E., Whitman, W. B., and Goodfellow, M. (2013). A call to action for the International Committee on Systematics of Prokaryotes. *Trends Microbiol.* 21, 51–52. doi:10.1016/j.tim.2012.11.004.

Suzuki, R., and Shimodaira, H. (2015). *pvclust:* hierarchical clustering with p-values via multiscale bootstrap resampling.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi:10.1080/10635150701472164.

Thawai, C., Kittiwongwattana, C., Thanaboripat, D., Laosinwattana, C., Koohakan, P., and Parinthawong, N. (2016). *Micromonospora soli* sp. nov., isolated from rice rhizosphere soil. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 109, 449–456. doi:10.1007/s10482-016-0651-3.

Thompson, C. C., Amaral, G. R., Campeão, M., Edwards, R. A., Polz, M. F., Dutilh, B. E., et al. (2015). Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch. Microbiol.* 197, 359–370. doi:10.1007/s00203-014-1071-2.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res* 22, 4673–4680.

Trujillo, M. E., Alonso-Vega, P., Rodríguez, R., Carro, L., Cerda, E., Alonso, P., et al. (2010). The genus *Micromonospora* is widespread in legume root nodules: The example of *Lupinus angustifolius. ISME J.* 4, 1265–1281. doi:10.1038/ismej.2010.55.

Trujillo, M. E., Fernández-Molinero, C., Velázquez, E., Kroppenstedt, R. M., Schumann, P., Mateos, P. F., et al. (2005). *Micromonospora mirobrigensis* sp. nov. *Int. J. Syst. Evol. Microbiol.* 55, 877–880. doi:10.1099/ijs.0.63361-0.

Trujillo, M. E., Kroppenstedt, R. M., Fernández-Molinero, C., Schumann, P., and Martínez-Molina, E. (2007). *Micromonospora lupini* sp. nov. and *Micromonospora saelicesensis* sp. nov., isolated from root nodules of *Lupinus angustifolius. Int. J. Syst. Evol. Microbiol.* 57, 2799–2804. doi:10.1099/ijs.0.65192-0.

Trujillo, M. E., Riesco, R., Benito, P., and Carro, L. (2015). Endophytic actinobacteria and the interaction of *Micromonospora* and nitrogen fixing plants. *Front. Microbiol.* 6, 1–15. doi:10.3389/fmicb.2015.01341.

Vaas, L. A. I., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H. P., et al. (2013). *Opm:* An R package for analysing OmniLog® phenotype microarray data. *Bioinformatics* 29, 1823–1824. doi:10.1093/bioinformatics/btt291.

Vaas, L. A. I., Sikorski, J., Michael, V., Göker, M., and Klenk, H. P. (2012). Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS One* 7. doi:10.1371/journal.pone.0034846.

Vandamme, P., and Peeters, C. (2014). Time to revisit polyphasic taxonomy. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 106, 57–65. doi:10.1007/s10482-014-0148-x.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol. Rev.* 60, 407–438. doi:10.1007/s12088-007-0022-x.

Veyisoglu, A., Carro, L., Cetin, D., Guven, K., Spröer, C., Pötter, G., et al. (2016). *Micromonospora profundi* sp. nov., isolated from deep marine sediment. *Int. J. Syst. Evol. Microbiol.* 66, 4735–4743. doi:10.1099/ijsem.0.001419.

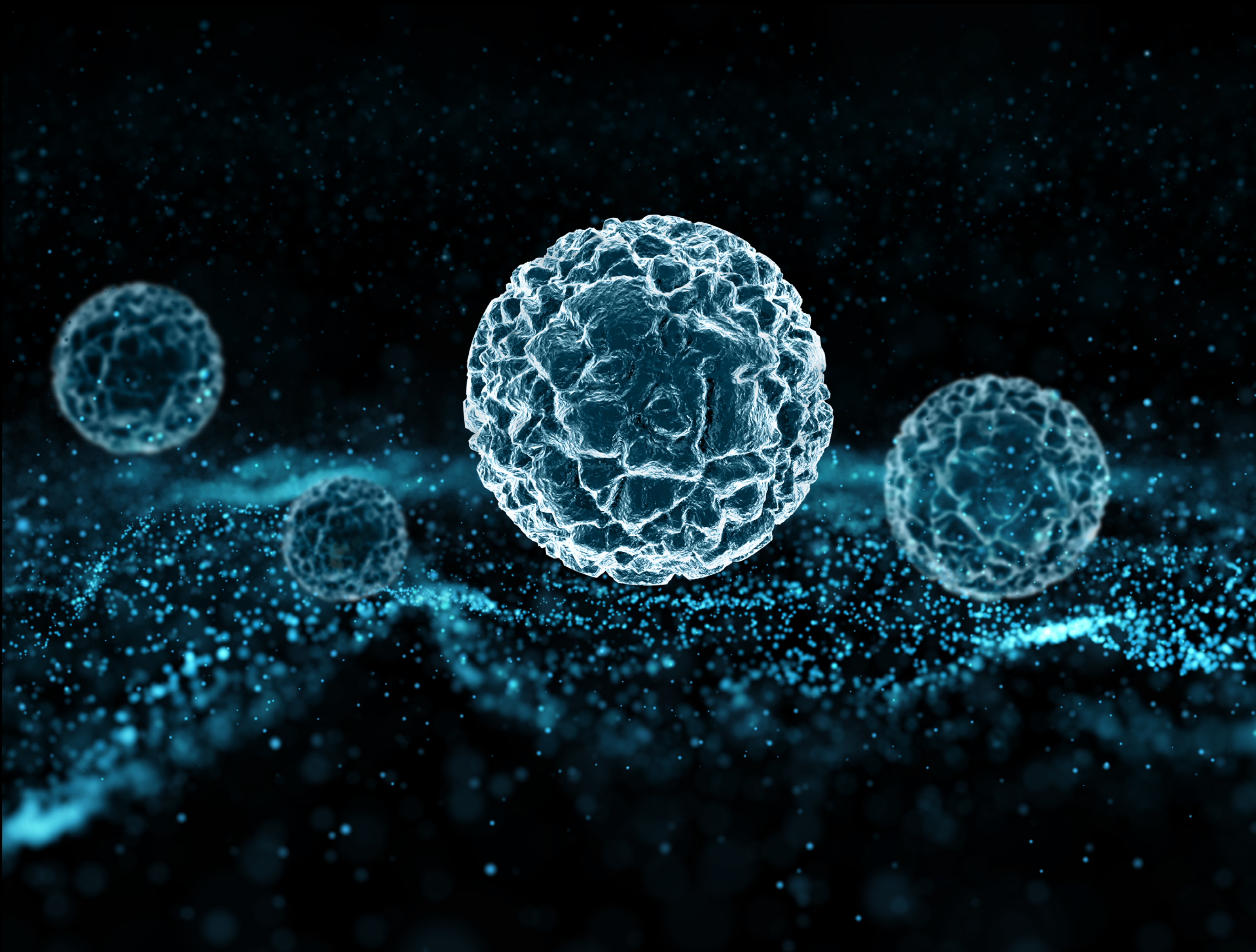
Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., et al. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int. J. Syst. Evol. Microbiol.* 37, 463–464. doi:https://doi.org/10.1099/00207713-37-4-463.

Williams, S. T., Goodfellow, M., Alderson, G., Wellington, E. M. H., Sneath, P. H. A., and Sackin, M. J. (1983). Numerical Classification of *Streptomyces* and Related Genera. *Microbiology* 129, 1743–1813. doi:10.1099/00221287-129-6-1743.

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017). Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* 67, 1613–1617. doi:10.1099/ijsem.0.001755.

Zhang, L., Xi, L., Ruan, J., and Huang, Y. (2012). *Micromonospora yangpuensis* sp. nov., isolated from a sponge. *Int. J. Syst. Evol. Microbiol.* 62, 272–278. doi:10.1099/ijs.0.029439-0.

FINAL REMARKS AND FUTURE WORK



FINAL REMARKS AND FUTURE WORK

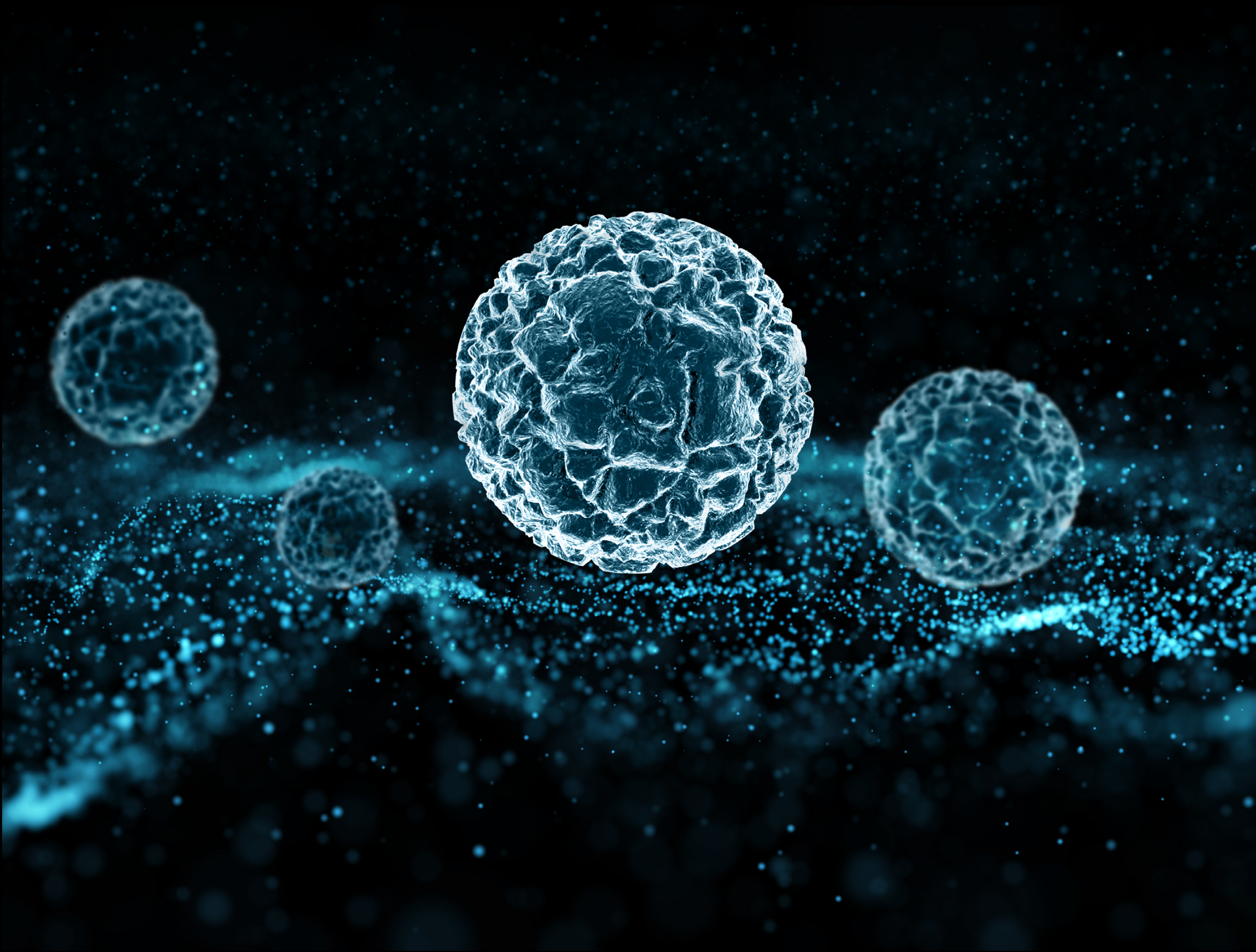
The results reported in this thesis revealed several plant-related features that were shared among many strains isolated from plant-tissues. The novel bioinformatic approach used in this work also pointed out that these features were shared not only among *Micromonospora* strains isolated from a plant-related environment, but also between strains isolated from soils. The possibility to select strains with the potential to adapt to the plant environment even when their origin of isolation is not related to the plant is of interest as it shows that strains from non-endophytic habitats may also serve as potential candidates for future agrobiotechological processes.

The next step is to adapt and automatize the pipeline with the aim of providing an easy to use and scalable method to catalogue any *Micromonospora* genome with a potential for interaction with a plant. New *Micromonospora* genomes need to be tested and if successful, this method could be very useful for the selection of new candidates for biotechnological applications in agriculture and the environment.

In this work, we have confirmed that *Micromonospora saelicesensis* and *Micromonospora noduli* are two very close but separate species. Genomic tools have proven to be a very useful tool to build a reliable framework to characterize and separate species. However, the organization of genomic data, the use command-based programs and the final visualization of results are still a challenge for non-bioinformaticians. In this thesis we have proposed two scripts, UBCG_iTOL_maker and GGDC Output Management Assistant (GOMA) that try to make the management of genomic data and presentation of the results easier for all end-users. UBCG_iTOL_maker in particular could be greatly improved by adding extra features to automatically add genomes in the internal database or modify the existing data. These features will be implemented in the near future, making the script more user-friendly.

We have sequenced seventeen strains in this work, six of which potentially are new species based on OGRI and 16S rRNA, *gyr*B and MLSA phylogenies. The strains LAH09, NIE79, MED01, NIE111, PSH03 and PSH25 will be further characterized and if appropriate, be described as new *Micromonospora* species.

APPENDIX



APPENDIX I

Distribution of COG categories in the *Micromonospora* genomes. A: RNA processing and modification; B: Chromatin structure and dynamics; C: Energy production and conversion; D: Cell cycle control, cell division, chromosome partitioning; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; J: Translation, ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Post-translational modification, protein turnover, and chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport, and catabolism; S: Function unknown; T: Signal transduction mechanisms; U: Intracellular trafficking, secretion, and vesicular transport; V: Defense mechanisms; Z: Cytoskeleton

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	S	T	U	V	Z
DSM45487	0,01%	0,01%	4,01%	0,58%	4,91%	1,43%	5,19%	2,32%	2,37%	2,62%	7,17%	3,27%	3,12%	0,03%	2,37%	4,42%	2,28%	31,71%	3,34%	0,61%	1,57%	0,01%
GAR05	0,02%	0,02%	4,50%	0,47%	4,76%	1,33%	6,21%	2,36%	2,26%	2,59%	8,73%	2,83%	3,24%	0,05%	2,29%	4,33%	1,67%	33,11%	4,09%	0,47%	1,92%	0,00%
GAR06	0,02%	0,02%	4,58%	0,42%	4,81%	1,34%	6,29%	2,41%	2,25%	2,67%	8,79%	2,65%	3,11%	0,05%	2,42%	4,32%	1,70%	33,35%	4,21%	0,45%	1,97%	0,00%
L5	0,02%	0,02%	4,37%	0,57%	5,42%	1,42%	5,47%	2,47%	2,71%	2,81%	7,62%	3,11%	3,37%	0,03%	2,52%	4,61%	2,50%	32,85%	3,46%	0,46%	1,61%	0,03%
LAH08	0,01%	0,01%	4,49%	0,49%	4,77%	1,35%	6,31%	2,30%	2,23%	2,54%	8,67%	2,78%	3,17%	0,06%	2,39%	4,13%	1,82%	33,12%	4,16%	0,48%	1,97%	0,00%
LAH09	0,02%	0,02%	4,16%	0,42%	4,88%	1,32%	5,95%	2,21%	2,23%	2,60%	8,10%	2,54%	3,44%	0,02%	2,15%	4,02%	2,01%	32,79%	3,68%	0,48%	1,70%	0,02%
Lupac 06	0,02%	0,02%	4,54%	0,44%	4,77%	1,34%	6,28%	2,36%	2,27%	2,65%	8,67%	2,79%	3,26%	0,06%	2,44%	4,36%	1,72%	33,24%	4,19%	0,47%	2,00%	0,00%
Lupac 07	0,02%	0,02%	4,59%	0,50%	4,85%	1,39%	6,41%	2,36%	2,24%	2,61%	8,70%	2,70%	3,26%	0,05%	2,44%	4,13%	1,85%	33,64%	4,16%	0,47%	2,01%	0,00%
Lupac 08	0,01%	0,01%	4,63%	0,46%	4,62%	1,27%	5,90%	2,31%	2,40%	2,53%	8,48%	2,60%	3,23%	0,06%	2,06%	3,85%	2,06%	32,43%	3,98%	0,49%	1,70%	0,01%
M. acroterricola	0,02%	0,02%	4,50%	0,44%	4,76%	1,44%	5,97%	2,26%	2,14%	2,71%	7,48%	2,68%	3,20%	0,02%	2,35%	4,16%	1,48%	30,95%	3,90%	0,44%	1,61%	0,00%
M. aurantiaca	0,02%	0,03%	4,15%	0,52%	5,22%	1,44%	5,29%	2,30%	2,64%	2,84%	7,46%	3,35%	3,23%	0,03%	2,56%	4,45%	2,48%	33,61%	3,44%	0,50%	1,55%	0,02%
M. auratinigra	0,02%	0,02%	4,16%	0,50%	4,93%	1,46%	5,50%	2,60%	2,43%	2,86%	7,70%	2,77%	3,54%	0,00%	2,39%	3,80%	2,12%	33,28%	4,03%	0,45%	1,47%	0,02%
M. avicenniae	0,02%	0,02%	4,80%	0,60%	5,21%	1,50%	6,06%	2,45%	2,29%	2,84%	8,58%	3,40%	3,27%	0,03%	2,79%	4,12%	1,76%	32,37%	3,56%	0,47%	1,44%	0,00%
M. carbonacea	0,01%	0,03%	4,09%	0,51%	4,90%	1,36%	5,46%	2,19%	2,54%	2,47%	6,97%	3,11%	3,30%	0,03%	2,12%	4,07%	2,62%	29,57%	3,40%	0,41%	1,48%	0,01%
M. chaiyaphumensis	0,02%	0,02%	4,17%	0,48%	5,08%	1,33%	5,61%	2,61%	2,40%	2,83%	8,35%	2,77%	3,39%	0,00%	2,61%	4,16%	2,09%	32,65%	4,01%	0,45%	1,63%	0,02%
M. chalcea	0,02%	0,03%	4,27%	0,59%	5,89%	1,49%	5,25%	2,41%	2,89%	2,75%	7,44%	3,40%	3,16%	0,03%	2,30%	4,78%	2,63%	32,62%	3,46%	0,43%	1,59%	0,03%
M. chersina	0,02%	0,02%	4,45%	0,45%	5,10%	1,35%	5,70%	2,70%	2,36%	2,86%	8,30%	2,79%	3,50%	0,00%	2,50%	4,25%	2,36%	32,80%	3,85%	0,45%	1,52%	0,00%
M. chokoriensis	0,02%	0,02%	4,24%	0,53%	4,55%	1,44%	5,83%	2,32%	2,45%	2,75%	7,96%	2,85%	3,25%	0,03%	2,37%	3,84%	2,05%	32,31%	3,79%	0,51%	2,00%	0,00%
M. citrea	0,02%	0,02%	4,03%	0,57%	4,81%	1,30%	5,09%	2,28%	2,12%	2,84%	8,19%	3,07%	3,05%	0,00%	2,26%	4,00%	2,17%	31,32%	3,89%	0,42%	1,91%	0,00%
M. coriariae	0,02%	0,02%	4,77%	0,50%	4,70%	1,46%	6,27%	2,19%	2,21%	2,71%	8,15%	3,28%	3,13%	0,00%	2,57%	4,23%	1,73%	32,01%	4,07%	0,45%	1,71%	0,00%

M. costi	0,02%	0,02%	4,84%	0,56%	5,26%	1,51%	6,30%	2,39%	2,37%	2,87%	9,27%	2,82%	3,69%	0,03%	2,65%	4,71%	1,88%	33,48%	3,74%	0,47%	1,80%	0,02%
M. coxensis	0,02%	0,02%	4,35%	0,55%	4,96%	1,45%	5,73%	2,55%	2,55%	2,95%	7,69%	2,88%	3,36%	0,02%	2,58%	4,13%	2,20%	32,41%	3,68%	0,43%	1,72%	0,02%
M. cremea	0,01%	0,01%	4,27%	0,41%	4,54%	1,26%	5,53%	2,15%	1,88%	2,44%	7,84%	4,21%	2,87%	0,00%	2,29%	3,77%	1,45%	29,63%	3,58%	0,33%	1,67%	0,01%
M. eburnea	0,02%	0,02%	4,25%	0,55%	5,08%	1,45%	5,10%	2,39%	2,42%	2,91%	8,24%	3,08%	3,26%	0,03%	2,44%	4,20%	3,04%	30,55%	3,60%	0,44%	1,83%	0,00%
M. echinaurantiaca	0,02%	0,02%	4,32%	0,53%	4,91%	1,32%	5,48%	2,27%	2,32%	2,77%	8,71%	3,14%	3,21%	0,00%	2,58%	4,04%	2,03%	32,21%	3,74%	0,42%	1,51%	0,00%
M. echinofusca	0,02%	0,02%	4,11%	0,59%	5,23%	1,49%	5,17%	2,38%	2,49%	2,95%	6,77%	3,55%	3,32%	0,02%	2,39%	4,23%	3,30%	31,15%	3,66%	0,45%	2,20%	0,02%
M. echinospora	0,02%	0,02%	4,13%	0,53%	5,42%	1,42%	6,08%	2,36%	2,69%	2,82%	6,94%	3,16%	3,21%	0,02%	2,47%	4,32%	3,33%	29,28%	3,24%	0,42%	1,57%	0,02%
M. endolithica	0,02%	0,02%	4,08%	0,55%	4,74%	1,43%	5,20%	2,24%	2,41%	2,79%	7,89%	3,37%	3,11%	0,03%	2,45%	4,60%	2,15%	32,48%	4,17%	0,48%	1,75%	0,00%
M. globispora	0,02%	0,02%	4,96%	0,59%	4,75%	1,27%	6,17%	2,37%	2,55%	2,63%	7,45%	4,22%	3,26%	0,00%	2,49%	4,02%	2,12%	30,65%	3,17%	0,42%	1,13%	0,00%
M. haikouensis	0,02%	0,03%	4,16%	0,58%	4,99%	1,46%	5,64%	2,34%	2,43%	2,79%	6,87%	3,41%	3,20%	0,03%	2,40%	4,34%	3,06%	30,55%	3,14%	0,42%	1,46%	0,02%
M. halophytica	0,02%	0,02%	4,21%	0,57%	5,38%	1,58%	4,44%	2,64%	2,85%	2,99%	6,71%	3,31%	3,19%	0,02%	2,48%	3,95%	2,66%	31,09%	3,66%	0,51%	1,42%	0,02%
M. humi	0,03%	0,02%	4,25%	0,47%	5,22%	1,39%	5,90%	2,33%	2,62%	2,81%	7,91%	2,72%	3,22%	0,02%	2,26%	4,12%	2,44%	32,15%	3,98%	0,47%	1,71%	0,02%
M. inaquosa	0,01%	0,03%	4,54%	0,45%	4,85%	1,32%	5,87%	2,02%	2,18%	2,38%	7,79%	3,78%	3,13%	0,03%	2,32%	4,03%	2,04%	30,99%	3,96%	0,49%	1,73%	0,00%
M. inositola	0,02%	0,02%	4,47%	0,57%	4,88%	1,43%	5,54%	2,22%	2,47%	2,67%	7,58%	4,04%	3,41%	0,02%	2,34%	3,71%	1,70%	31,09%	3,57%	0,47%	1,38%	0,00%
M. inyonensis	0,02%	0,02%	3,58%	0,62%	5,05%	1,42%	3,77%	2,37%	2,43%	2,89%	5,77%	5,29%	2,78%	0,00%	2,24%	3,84%	2,43%	26,58%	3,01%	0,46%	1,15%	0,00%
M. krabiensis	0,02%	0,02%	4,32%	0,49%	4,88%	1,51%	5,74%	2,20%	2,19%	2,69%	8,45%	2,91%	3,37%	0,02%	2,43%	4,02%	1,69%	32,07%	3,72%	0,45%	1,59%	0,00%
M. marina	0,02%	0,02%	4,15%	0,61%	5,28%	1,61%	4,24%	2,38%	2,80%	3,19%	6,53%	3,30%	3,50%	0,00%	2,44%	4,38%	3,08%	31,55%	3,34%	0,49%	1,62%	0,04%
M. matsumotoense	0,01%	0,03%	4,11%	0,55%	4,90%	1,42%	5,69%	2,37%	2,58%	2,62%	7,46%	3,20%	3,31%	0,04%	2,09%	4,23%	2,85%	30,38%	3,65%	0,46%	1,58%	0,01%
M. mirobrigensis	0,02%	0,02%	4,41%	0,62%	5,19%	1,63%	5,67%	2,51%	2,71%	2,93%	8,00%	3,09%	3,69%	0,00%	2,39%	3,99%	2,15%	31,34%	3,99%	0,49%	1,42%	0,03%
M. narathiwatensis	0,02%	0,02%	4,27%	0,61%	5,44%	1,45%	5,36%	2,39%	2,58%	3,04%	7,29%	3,21%	3,65%	0,00%	2,49%	3,98%	2,63%	31,57%	3,65%	0,55%	1,74%	0,00%
M. nigra	0,02%	0,04%	3,88%	0,61%	5,03%	1,58%	4,94%	2,64%	2,71%	2,98%	5,89%	3,67%	3,20%	0,02%	2,28%	3,97%	2,86%	29,61%	3,40%	0,54%	1,44%	0,00%
M. noduli	0,02%	0,02%	4,65%	0,52%	4,90%	1,38%	6,43%	2,36%	2,23%	2,59%	8,81%	2,77%	3,27%	0,05%	2,46%	4,24%	1,88%	33,97%	4,10%	0,46%	1,91%	0,00%
M. olivasterospora	0,02%	0,02%	4,17%	0,56%	5,33%	1,38%	4,31%	2,44%	2,57%	2,69%	5,94%	7,71%	2,99%	0,00%	2,32%	4,01%	2,13%	29,39%	3,13%	0,50%	1,21%	0,00%
M. pallida	0,01%	0,01%	3,74%	0,52%	4,98%	1,25%	4,73%	2,30%	2,52%	2,59%	6,26%	2,94%	3,03%	0,00%	2,56%	3,94%	2,85%	28,82%	3,33%	0,39%	1,38%	0,00%
M. palomenae	0,02%	0,02%	4,31%	0,54%	4,80%	1,44%	5,13%	2,47%	2,61%	2,99%	7,53%	3,42%	3,57%	0,02%	2,33%	3,94%	1,88%	31,87%	3,72%	0,44%	1,39%	0,00%
M. pattaloongensis	0,02%	0,02%	4,49%	0,57%	5,48%	1,76%	5,65%	2,67%	2,45%	3,14%	6,80%	3,60%	3,68%	0,59%	2,81%	4,07%	1,98%	30,51%	4,51%	0,57%	1,30%	0,00%
M. peucetia	0,02%	0,02%	3,81%	0,53%	5,44%	1,33%	4,85%	2,18%	2,80%	2,82%	7,51%	3,36%	3,02%	0,00%	2,43%	4,51%	3,14%	30,04%	3,52%	0,43%	2,03%	0,00%

	A	B	C	D	E	F	G	H	I	J	K	M	N	O	P	Q	R	S	T	U	V	Z
M. pisi	0,01%	0,01%	4,21%	0,56%	4,56%	1,42%	5,71%	2,31%	2,10%	2,36%	8,47%	3,72%	3,14%	0,06%	2,18%	4,12%	1,83%	31,87%	3,13%	0,35%	1,69%	0,00%
M. purpureochromogenes	0,02%	0,02%	4,45%	0,44%	4,50%	1,42%	4,94%	2,49%	2,49%	2,81%	7,43%	3,53%	3,50%	0,00%	2,62%	3,63%	1,88%	32,77%	3,68%	0,49%	1,24%	0,00%
M. rhizosphaerae	0,01%	0,01%	5,03%	0,47%	4,94%	1,34%	5,68%	2,31%	2,12%	2,56%	7,79%	4,03%	3,22%	0,01%	2,40%	3,68%	1,57%	29,68%	3,19%	0,44%	1,41%	0,00%
M. rifamycinica	0,02%	0,03%	4,46%	0,56%	4,97%	1,48%	5,47%	2,48%	2,73%	2,81%	7,39%	3,06%	3,55%	0,02%	2,44%	4,29%	3,06%	31,00%	3,67%	0,52%	1,77%	0,03%
M. rosaria	0,02%	0,02%	4,31%	0,67%	5,08%	1,44%	5,85%	2,64%	2,47%	2,61%	7,29%	3,82%	2,94%	0,03%	2,38%	4,44%	2,94%	30,39%	4,00%	0,44%	1,58%	0,03%
M. saelicesensis	0,02%	0,02%	4,59%	0,46%	4,79%	1,34%	6,28%	2,36%	2,22%	2,63%	8,56%	2,71%	3,24%	0,05%	2,37%	4,24%	1,67%	33,11%	4,07%	0,46%	1,93%	0,00%
M. sagamiensis	0,02%	0,02%	4,14%	0,57%	5,32%	1,40%	5,69%	2,78%	2,76%	3,02%	6,65%	3,91%	3,45%	0,02%	2,54%	4,28%	3,32%	31,37%	3,81%	0,49%	1,80%	0,00%
M. sediminicola	0,02%	0,02%	3,90%	0,56%	5,11%	1,35%	5,31%	2,17%	2,22%	2,78%	7,43%	2,90%	3,21%	0,02%	2,19%	4,02%	2,39%	32,11%	3,91%	0,43%	1,47%	0,02%
M. siamensis	0,02%	0,02%	4,47%	0,61%	5,19%	1,52%	5,03%	2,43%	2,64%	3,11%	7,82%	3,13%	3,39%	0,00%	2,69%	3,98%	2,06%	32,45%	4,24%	0,51%	1,68%	0,03%
M. tulbaghiae	0,02%	0,03%	4,28%	0,57%	5,54%	1,50%	5,66%	2,26%	2,70%	2,96%	8,09%	2,90%	3,33%	0,03%	2,50%	4,81%	2,25%	32,66%	3,79%	0,45%	1,70%	0,03%
M. viridifaciens	0,02%	0,02%	4,03%	0,51%	5,85%	1,45%	4,45%	2,21%	2,54%	2,88%	7,64%	3,90%	3,08%	0,00%	2,68%	3,97%	2,13%	31,08%	3,42%	0,44%	1,49%	0,00%
M. wenchangensis	0,02%	0,03%	4,65%	0,56%	4,99%	1,46%	5,54%	2,58%	2,68%	2,64%	7,64%	3,03%	3,40%	0,03%	2,26%	4,54%	3,31%	30,61%	3,51%	0,44%	1,49%	0,03%
M. yangpuensis	0,02%	0,02%	3,93%	0,55%	4,77%	1,44%	5,10%	2,52%	2,42%	2,74%	6,53%	3,16%	3,24%	0,03%	2,13%	4,57%	2,30%	29,84%	3,51%	0,54%	1,60%	0,02%
M. zamorensis	0,02%	0,02%	4,58%	0,55%	4,89%	1,46%	6,12%	2,29%	2,42%	2,65%	8,26%	2,74%	3,32%	0,05%	2,42%	4,04%	1,99%	32,72%	3,76%	0,48%	1,85%	0,00%
MED01	0,01%	0,01%	4,10%	0,44%	4,55%	1,29%	5,59%	2,18%	2,02%	2,28%	8,16%	2,80%	3,03%	0,01%	2,08%	4,00%	1,58%	31,46%	3,84%	0,41%	1,75%	0,01%
MED15	0,02%	0,02%	4,63%	0,54%	4,81%	1,38%	6,47%	2,31%	2,24%	2,56%	8,71%	2,81%	3,18%	0,05%	2,42%	4,14%	1,86%	33,79%	4,11%	0,45%	1,97%	0,00%
NIE111	0,02%	0,02%	4,19%	0,40%	4,94%	1,32%	6,02%	2,28%	2,12%	2,48%	8,56%	2,77%	3,40%	0,02%	2,16%	3,99%	1,72%	32,66%	4,14%	0,48%	1,82%	0,02%
NIE79	0,01%	0,01%	4,14%	0,38%	4,58%	1,29%	5,95%	2,26%	2,04%	2,42%	8,56%	2,59%	3,64%	0,01%	2,12%	4,07%	1,69%	32,19%	4,01%	0,47%	1,87%	0,01%
ONO23	0,02%	0,02%	4,48%	0,51%	4,78%	1,33%	6,60%	2,23%	2,25%	2,54%	8,72%	2,78%	3,18%	0,05%	2,41%	4,18%	1,84%	33,38%	4,04%	0,45%	1,90%	0,00%
ONO86	0,02%	0,02%	4,45%	0,47%	4,68%	1,35%	6,35%	2,18%	2,17%	2,47%	8,53%	2,90%	3,14%	0,05%	2,32%	4,08%	1,80%	32,53%	3,89%	0,47%	1,95%	0,00%
PSH03	0,02%	0,02%	4,22%	0,38%	4,74%	1,34%	5,96%	2,29%	2,12%	2,55%	8,24%	2,74%	3,37%	0,02%	2,11%	4,14%	1,96%	32,25%	3,95%	0,44%	1,82%	0,02%
PSH25	0,01%	0,01%	4,10%	0,45%	4,56%	1,21%	6,05%	2,14%	2,03%	2,41%	8,23%	2,67%	3,17%	0,01%	2,02%	4,01%	1,72%	30,64%	3,71%	0,44%	1,45%	0,01%
PSN01	0,02%	0,00%	4,40%	0,46%	4,67%	1,35%	6,25%	2,25%	2,16%	2,59%	8,47%	2,56%	3,22%	0,05%	2,31%	4,23%	1,68%	32,96%	3,94%	0,46%	1,90%	0,00%
PSN13	0,01%	0,01%	4,50%	0,46%	4,72%	1,32%	6,52%	2,31%	2,13%	2,50%	8,84%	2,86%	3,37%	0,03%	2,38%	4,11%	1,63%	33,09%	4,11%	0,46%	1,87%	0,00%
S. arenicola	0,02%	0,02%	4,96%	0,99%	5,74%	1,66%	4,45%	2,99%	3,28%	3,28%	6,96%	4,37%	3,48%	0,04%	2,63%	4,11%	3,70%	30,44%	3,05%	0,51%	1,21%	0,38%
S. pacifica	0,02%	0,02%	4,34%	0,76%	5,18%	1,67%	3,98%	2,81%	2,70%	3,11%	6,26%	4,24%	2,93%	0,00%	2,24%	3,57%	2,43%	29,38%	2,82%	0,69%	1,52%	0,20%
S. tropica	0,02%	0,02%	5,00%	0,80%	5,44%	1,70%	4,48%	3,10%	3,24%	3,62%	6,76%	3,79%	3,12%	0,00%	2,45%	4,14%	2,91%	30,37%	2,80%	0,52%	1,38%	0,15%

APPENDIX II

Significantly over-represented and under-represented KEGG annotations identified in the clustering analysis, distributed by KEGG pathway maps and clusters. As some of the KEGG annotations are repeated in different pathways, the number of unique KEGG annotations have been summarized at the bottom of each cluster table. Fold change column is colored in green if the fold change is above 2 and in red if the fold change is under 0.5.

Cluster I

CatA	CatB	CatC	KO	Short name	Long name	EC	v test	Mean in category	Overall mean	sd in category	Overall sd	p value	q value	Fold
A09100 Metabolism	09101 Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K01689	ENO, eno	enolase	[EC:4.2.1.11]	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	-5,513573322	0,033333333	0,418918919	0,179505494	0,493382061	3,52E-08	0,000134284	0,08
		00030 Pentose phosphate pathway [PATH:ko00030]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase	[EC:1.1.1.49 1.1.1.363]	7,125804862	1,7	0,810810811	0,585946528	0,880351215	1,03E-12	3,95E-09	2,10
			K00851	E2.7.1.12, gntK, idnK	gluconokinase	[EC:2.7.1.12]	4,944147809	1,9	1,364864865	0,472581563	0,7636032	7,65E-07	0,002920682	1,39
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	-5,513573322	0,033333333	0,418918919	0,179505494	0,493382061	3,52E-08	0,000134284	0,08
		00051 Fructose and mannose metabolism [PATH:ko00051]	K00064	E1.1.1.122	D-threo-aldose 1-dehydrogenase	[EC:1.1.1.122]	6,628126793	2,166666667	1,054054054	1,035481638	1,184264989	3,40E-11	1,30E-07	2,06
			K18335	K18335	2-keto-3-deoxy-L-fuconate dehydrogenase	[EC:1.1.1.-]	5,304685126	0,966666667	0,554054054	0,314466038	0,548755149	1,13E-07	0,000431042	1,74
		00052 Galactose metabolism [PATH:ko00052]	K01187	malZ	alpha-glucosidase	[EC:3.2.1.20]	6,43045666	1,033333333	0,432432432	0,657436097	0,65926005	1,27E-10	4,86E-07	2,39
			K01193	INV, sacA	beta-fructofuranosidase	[EC:3.2.1.26]	4,552034656	0,766666667	0,445945946	0,422952585	0,497069572	5,31E-06	0,020290187	1,72
			K07407	E3.2.1.22B, galA, rafA	alpha-galactosidase	[EC:3.2.1.22]	5,725074831	3,866666667	2,594594595	0,763034876	1,567567568	1,03E-08	3,95E-05	1,49
		00053 Ascorbate and aldarate metabolism [PATH:ko00053]	K00064	E1.1.1.122	D-threo-aldose 1-dehydrogenase	[EC:1.1.1.122]	6,628126793	2,166666667	1,054054054	1,035481638	1,184264989	3,40E-11	1,30E-07	2,06
			K13875	K13875, araC	L-arabonate dehydrase	[EC:4.2.1.25]	6,865576216	0,833333333	0,364864865	0,372677996	0,481392247	6,62E-12	2,53E-08	2,28
			K18981	udh	uronate dehydrogenase	[EC:1.1.1.203]	4,763975549	1,066666667	0,662162162	0,249443826	0,599031358	1,90E-06	0,007249055	1,61
		00500 Starch and sucrose metabolism [PATH:ko00500]	K01187	malZ	alpha-glucosidase	[EC:3.2.1.20]	6,43045666	1,033333333	0,432432432	0,657436097	0,65926005	1,27E-10	4,86E-07	2,39
			K01193	INV, sacA	beta-fructofuranosidase	[EC:3.2.1.26]	4,552034656	0,766666667	0,445945946	0,422952585	0,497069572	5,31E-06	0,020290187	1,72
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	-5,513573322	0,033333333	0,418918919	0,179505494	0,493382061	3,52E-08	0,000134284	0,08
			K05350	bglB	beta-glucosidase	[EC:3.2.1.21]	5,017187097	5,9	5,013513514	0,746100976	1,246543797	5,24E-07	0,002002435	1,18
		00520 Amino sugar and nucleotide sugar metabolism [PATH:ko00520]	K00820	glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing)	[EC:2.6.1.16]	4,948390841	1,6	1,283783784	0,489897949	0,450833171	7,48E-07	0,00285774	1,25
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	-5,513573322	0,033333333	0,418918919	0,179505494	0,493382061	3,52E-08	0,000134284	0,08
			K15921	xynD	arabinoxylan arabinofuranohydrolase	[EC:3.2.1.55]	6,432040841	1,3	0,581081081	0,73711148	0,788545647	1,26E-10	4,81E-07	2,24
		00562 Inositol phosphate metabolism [PATH:ko00562]	K16044	iolW	scyllo-inositol 2-dehydrogenase (NADP+)	[EC:1.1.1.371]	5,212865632	1	0,621621622	0,25819889	0,512089063	1,86E-07	0,000710126	1,61
		00640 Propanoate metabolism [PATH:ko00640]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00650 Butanoate metabolism [PATH:ko00650]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
	09102 Energy metabolism	00680 Methane metabolism [PATH:ko00680]	K01689	ENO, eno	enolase	[EC:4.2.1.11]	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
		00910 Nitrogen metabolism [PATH:ko00910]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	-4,560291954	0,1	0,418918919	0,3	0,493382061	5,11E-06	0,019508426	0,24
	09103 Lipid metabolism	00061 Fatty acid biosynthesis [PATH:ko00061]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
			K18660	ACSF3	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	-5,393087304	0,166666667	0,851351351	0,372677996	0,895671761	6,93E-08	0,000264494	0,20
		00071 Fatty acid degradation [PATH:ko00071]	K00496	alkB1_2, alkM	alkane 1-monooxygenase	[EC:1.14.15.3]	-4,370203322	0,066666667	0,364864865	0,249443826	0,481392247	1,24E-05	0,047405591	0,18
			K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
		00561 Glycerolipid metabolism [PATH:ko00561]	K07407	E3.2.1.22B, galA, rafA	alpha-galactosidase	[EC:3.2.1.22]	5,725074831	3,866666667	2,594594595	0,763034876	1,567567568	1,03E-08	3,95E-05	1,49
		00600 Sphingolipid metabolism [PATH:ko00600]	K07407	E3.2.1.22B, galA, rafA	alpha-galactosidase	[EC:3.2.1.22]	5,725074831	3,866666667	2,594594595	0,763034876	1,567567568	1,03E-08	3,95E-05	1,49
	09104 Nucleotide metabolism	00230 Purine metabolism [PATH:ko00230]	K01515	nudF	ADP-ribose pyrophosphatase	[EC:3.6.1.13]	-5,019324833	0,033333333	0,378378378	0,179505494	0,484982661	5,19E-07	0,001980281	0,09
			K13480	ygeU, xdhC	xanthine dehydrogenase iron-sulfur-binding subunit	NA	5,281899666	1,066666667	0,527027027	0,77172246	0,7207911	1,28E-07	0,000488264	2,02
			K13483	yagT	xanthine dehydrogenase YagT iron-sulfur-binding subunit	NA	5,343253313	1,1	0,540540541	0,789514619	0,738683804	9,13E-08	0,000348648	2,03
		00220 Arginine biosynthesis [PATH:ko00220]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	-4,560291954	0,1	0,418918919	0,3	0,493382061	5,11E-06	0,019508426	0,24
		00250 Alanine, aspartate and glutamate metabolism [PATH:ko00250]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	-4,560291954	0,1	0,418918919	0,3	0,493382061	5,11E-06	0,019508426	0,24
			K00820	glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing)	[EC:2.6.1.16]	4,948390841	1,6	1,283783784	0,489897949	0,450833171	7,48E-07	0,00285774	1,25

	09105 Amino acid metabolism		K01953	asnB, ASNS	asparagine synthase (glutamine-hydrolysing)	[EC:6.3.5.4]	5,563134646	0,866666667	0,472972973	0,339934634	0,499269005	2,65E-08	0,000101193	1,83
		00270 Cysteine and methionine metabolism [PATH:ko00270]	K01243	mtnN, mtn, pfs	adenosylhomocysteine nucleosidase	[EC:3.2.2.9]	7,355725157	0,933333333	0,418918919	0,249443826	0,493382061	1,90E-13	7,25E-10	2,23
			K01251	E3.3.1.1, ahcY	adenosylhomocysteinase	[EC:3.3.1.1]	4,632720273	0,633333333	0,310810811	0,546707316	0,491156266	3,61E-06	0,013782466	2,04
		00280 Valine, leucine and isoleucine degradation [PATH:ko00280]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
			K18660	ACSF3	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	-5,393087304	0,166666667	0,851351351	0,372677996	0,895671761	6,93E-08	0,000264494	0,20
			K18661	matB	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	-5,200328691	0,2	0,945945946	0,4	1,011980823	1,99E-07	0,000759738	0,21
			K18662	matB	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	-5,533007996	0,166666667	0,864864865	0,372677996	0,890252383	3,15E-08	0,000120216	0,19
		00290 Valine, leucine and isoleucine biosynthesis [PATH:ko00290]	K01687	ilvD	dihydroxy-acid dehydratase	[EC:4.2.1.9]	5,967854821	1,833333333	1,378378378	0,45338235	0,537831047	2,40E-09	9,18E-06	1,33
		00310 Lysine degradation [PATH:ko00310]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00380 Tryptophan metabolism [PATH:ko00380]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00400 Phenylalanine, tyrosine and tryptophan biosynthesis [PATH:ko00400]	K00891	E2.7.1.71, aroK, aroL	shikimate kinase	[EC:2.7.1.71]	4,560291954	0,9	0,581081081	0,3	0,493382061	5,11E-06	0,019508426	1,55
	09106 Metabolism of other amino acids	00410 beta-Alanine metabolism [PATH:ko00410]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00460 Cyanoamino acid metabolism [PATH:ko00460]	K05350	bglB	beta-glucosidase	[EC:3.2.1.21]	5,017187097	5,9	5,013513514	0,746100976	1,246543797	5,24E-07	0,002002435	1,18
		00473 D-Alanine metabolism [PATH:ko00473]	K01775	alr	alanine racemase	[EC:5.1.1.1]	5,662246898	3,466666667	2,743243243	0,805536398	0,901362494	1,49E-08	5,71E-05	1,26
		00480 Glutathione metabolism [PATH:ko00480]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase	[EC:1.1.1.49 1.1.1.363]	7,125804862	1,7	0,810810811	0,585946528	0,880351215	1,03E-12	3,95E-09	2,10
			K06048	gshA, ybdK	glutamate---cysteine ligase / carboxylate-amine ligase	[EC:6.3.2.2 6.3.-.-]	6,103364917	2,9	2,121621622	0,65064071	0,899740243	1,04E-09	3,97E-06	1,37
	09107 Glycan biosynthesis and metabolism	00603 Glycosphingolipid biosynthesis - globo and isoglobo series [PATH:ko00603]	K07407	E3.2.1.22B, galA, rafA	alpha-galactosidase	[EC:3.2.1.22]	5,725074831	3,866666667	2,594594595	0,763034876	1,567567568	1,03E-08	3,95E-05	1,49
	09108 Metabolism of cofactors and vitamins	00730 Thiamine metabolism [PATH:ko00730]	K00941	thiD	hydroxymethylpyrimidine/phosphomethylpyrimidine kinase	[EC:2.7.1.49 2.7.4.7]	5,19232388	1	0,648648649	0	0,477392479	2,08E-07	0,00079315	1,54
		00740 Riboflavin metabolism [PATH:ko00740]	K01497	ribA, RIB1	GTP cyclohydrolase II	[EC:3.5.4.25]	-4,847196857	0,333333333	0,689189189	0,471404521	0,517939673	1,25E-06	0,004782082	0,48
		00750 Vitamin B6 metabolism [PATH:ko00750]	K05275	E1.1.1.65	pyridoxine 4-dehydrogenase	[EC:1.1.1.65]	6,352120423	1,266666667	0,648648649	0,512076383	0,686401357	2,12E-10	8,11E-07	1,95
		00770 Pantothenate and CoA biosynthesis [PATH:ko00770]	K01687	ilvD	dihydroxy-acid dehydratase	[EC:4.2.1.9]	5,967854821	1,833333333	1,378378378	0,45338235	0,537831047	2,40E-09	9,18E-06	1,33
			K03525	coaX	type III pantothenate kinase	[EC:2.7.1.33]	-4,799317532	1,133333333	1,472972973	0,339934634	0,499269005	1,59E-06	0,006080123	0,77
		00790 Folate biosynthesis [PATH:ko00790]	K01497	ribA, RIB1	GTP cyclohydrolase II	[EC:3.5.4.25]	-4,847196857	0,333333333	0,689189189	0,471404521	0,517939673	1,25E-06	0,004782082	0,48
			K07141	mocA	molybdenum cofactor cytidylyltransferase	[EC:2.7.7.76]	6,605847948	4,366666667	2,972972973	1,048279013	1,488450791	3,95E-11	1,51E-07	1,47
	09109 Metabolism of terpenoids and polyketides	00281 Geraniol degradation [PATH:ko00281]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00523 Polyketide sugar unit biosynthesis [PATH:ko00523]	K13315	eryBII, tylCII, tylC1,	NDP-hexose C3-ketoreductase / dTDP-4-oxo-2-deoxy-alpha-D-pentos-2-ene 2,3-reductase	[EC:1.1.1.-]	-5,544565841	0,1	0,513513514	0,3	0,526160063	2,95E-08	0,00011254	0,19
			K13327	spnN, oleW	dTDP-3,4-didehydro-2,6-dideoxy-alpha-D-glucose 3-reductase	[EC:1.1.1.384]	-4,642276823	0,066666667	0,540540541	0,249443826	0,720157438	3,45E-06	0,013159921	0,12
		00903 Limonene and pinene degradation [PATH:ko00903]	K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		01053 Biosynthesis of siderophore group nonribosomal peptides [PATH:ko01053]	K04780	dhbF	nonribosomal peptide synthetase DhbF	NA	-4,685253672	2,466666667	4,878378378	1,802467445	3,631516334	2,80E-06	0,010678436	0,51
	09110 Biosynthesis of other secondary metabolites	00940 Phenylpropanoid biosynthesis [PATH:ko00940]	K05350	bglB	beta-glucosidase	[EC:3.2.1.21]	5,017187097	5,9	5,013513514	0,746100976	1,246543797	5,24E-07	0,002002435	1,18
		00960 Tropane, piperidine and pyridine alkaloid biosynthesis [PATH:ko00960]	K08081	TR1	tropinone reductase I	[EC:1.1.1.206]	5,865466269	0,933333333	0,405405405	0,679869268	0,634991358	4,48E-09	1,71E-05	2,30
	09111 Xenobiotics biodegradation and metabolism	00362 Benzoate degradation [PATH:ko00362]	K01607	pcaC	4-carboxymuconolactone decarboxylase	[EC:4.1.1.44]	5,670295538	1,3	0,716216216	0,525991128	0,726343495	1,43E-08	5,44E-05	1,82
			K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
		00930 Caprolactam degradation [PATH:ko00930]	K00496	alkB1_2, alkM	alkane 1-monooxygenase	[EC:1.14.15.3]	-4,370203322	0,066666667	0,364864865	0,249443826	0,481392247	1,24E-05	0,047405591	0,18
			K01782	fadJ	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase	[EC:1.1.1.35 4.2.1.17	5,59032167	0,7	0,310810811	0,525991128	0,491156266	2,27E-08	8,66E-05	2,25
A09120 Genetic Information Processing	09122 Translation	00970 Aminoacyl-tRNA biosynthesis [PATH:ko00970]	K01870	IARS, ileS	isoleucyl-tRNA synthetase	[EC:6.1.1.5]	4,912316744	1,6	1,27027027	0,489897949	0,473551769	9,00E-07	0,003437348	1,26
	09123 Folding, sorting and degradation	03018 RNA degradation [PATH:ko03018]	K01689	ENO, eno	enolase	[EC:4.2.1.11]	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
			K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	5,343506108	9,733333333	7,77027027	1,547758235	2,59181298	9,12E-08	0,000348162	1,25
			K01996	livF	branched-chain amino acid transport system ATP-binding protein	NA	6,212382198	10,56666667	7,864864865	1,282792094	3,068252097	5,22E-10	1,99E-06	1,34
			K02011	afuB, fbpB	iron(III) transport system permease protein	NA	-4,586726207	0,2	0,635135135	0,476095229	0,669294189	4,50E-06	0,017195067	0,31

A09130 Environmental Information Processing	09131 Membrane transport	02010 ABC transporters [PATH:ko02010]	K05816	ugpC	sn-glycerol 3-phosphate transport system ATP-binding protein	[EC:7.6.2.10]	5,005007787	0,566666667	0,256756757	0,495535625	0,436844051	5,59E-07	0,002133285	2,21
			K10008	gluA	glutamate transport system ATP-binding protein	[EC:7.4.2.1]	6,432909509	4,5	3,27027027	0,991631652	1,348645941	1,25E-10	4,78E-07	1,38
			K10111	malK, mtlK, thuK	multiple sugar transport system ATP-binding protein	[EC:3.6.3.-]	5,3182487	0,566666667	0,243243243	0,495535625	0,429040753	1,05E-07	0,00040012	2,33
			K10112	msmX, msmK,	multiple sugar transport system ATP-binding protein	NA	5,675711802	1,066666667	0,486486486	0,727247474	0,72117103	1,38E-08	5,27E-05	2,19
			K10117	msmE	raffinose/stachyose/melibiose transport system substrate-binding protein	NA	6,2739803	4,566666667	2,945945946	0,955103252	1,822471532	3,52E-10	1,34E-06	1,55
			K10118	msmF	raffinose/stachyose/melibiose transport system permease protein	NA	5,61843792	5,4	4,027027027	1,113552873	1,72401929	1,93E-08	7,36E-05	1,34
			K10119	msmG	raffinose/stachyose/melibiose transport system permease protein	NA	5,400613175	7,133333333	5,486486486	1,203698006	2,151324188	6,64E-08	0,000253633	1,30
			K10232	aglE, ggtB	alpha-glucoside transport system substrate-binding protein	NA	7,002786036	1,733333333	0,864864865	0,573488351	0,8749413	2,51E-12	9,58E-09	2,00
			K10234	aglG, ggtD	alpha-glucoside transport system permease protein	NA	5,983263365	2,866666667	2,27027027	0,561743318	0,703222261	2,19E-09	8,35E-06	1,26
			K10238	thuG, sugB	trehalose/maltose transport system permease protein	NA	5,484026816	1,9	1,351351351	0,597215762	0,705814317	4,16E-08	0,000158776	1,41
			K10439	rbsB	ribose transport system substrate-binding protein	NA	4,375171767	4,9	3,905405405	0,86986589	1,603787063	1,21E-05	0,04633856	1,25
			K10440	rbsC	ribose transport system permease protein	NA	4,588838156	4,233333333	3,418918919	0,80346472	1,252098312	4,46E-06	0,01702204	1,24
			K10441	rbsA	ribose transport system ATP-binding protein	[EC:7.5.2.7]	5,684687642	4,566666667	3,243243243	1,308519095	1,642434011	1,31E-08	5,00E-05	1,41
			K10546	ABC.GGU.S, chvE	putative multiple sugar transport system substrate-binding protein	NA	6,584509939	1,966666667	1,189189189	0,406885187	0,833028919	4,56E-11	1,74E-07	1,65
			K10548	ABC.GGU.A, gguA	putative multiple sugar transport system ATP-binding protein	[EC:7.5.2.-]	5,604179502	4,066666667	2,77027027	1,123486636	1,632005045	2,09E-08	7,99E-05	1,47
			K10559	rhaS	rhamnose transport system substrate-binding protein	NA	4,532771498	0,5	0,22972973	0,5	0,420658984	5,82E-06	0,022232224	2,18
			K10561	rhaQ	rhamnose transport system permease protein	NA	4,863411594	0,5	0,216216216	0,5	0,411663411	1,15E-06	0,004406346	2,31
			K11963	urtE	urea transport system ATP-binding protein	NA	4,659777841	0,666666667	0,351351351	0,471404521	0,477392479	3,17E-06	0,012089077	1,90
			K15772	ganQ	arabinogalactan oligomer / maltooligosaccharide transport system permease protein	NA	4,847196857	1,666666667	1,310810811	0,53748385	0,517939673	1,25E-06	0,004782082	1,27
			K17241	aguE	alpha-1,4-digalacturonate transport system substrate-binding protein	NA	6,133274208	0,8	0,378378378	0,4	0,484982661	8,61E-10	3,29E-06	2,11
			K17242	aguF	alpha-1,4-digalacturonate transport system permease protein	NA	4,659777841	0,666666667	0,351351351	0,471404521	0,477392479	3,17E-06	0,012089077	1,90
			K18230	tylC, oleB, carA, srmB	macrolide transport system ATP-binding/permease protein	NA	4,383920897	2,2	1,635135135	0,702376917	0,909028652	1,17E-05	0,044515071	1,35
			K19350	lsa	lincosamide and streptogramin A transport system ATP-binding/permease protein	NA	5,991641332	0,666666667	0,283783784	0,471404521	0,450833171	2,08E-09	7,93E-06	2,35
	09132 Signal transduction	02020 Two-component system [PATH:ko02020]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	7,210590737	1,833333333	0,945945946	0,45338235	0,86823667	5,57E-13	2,13E-09	1,94
			K07641	creC	two-component system, OmpR family, sensor histidine kinase CreC	[EC:2.7.13.3]	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB	[EC:2.7.13.3]	5,186809286	0,633333333	0,297297297	0,48189441	0,457068501	2,14E-07	0,00081699	2,13
			K07654	mtrB	two-component system, OmpR family, sensor histidine kinase MtrB	[EC:2.7.13.3]	7,121500093	2,8	1,932432432	0,476095229	0,859463709	1,07E-12	4,08E-09	1,45
			K07711	glrK, qseE	two-component system, NtrC family, sensor histidine kinase GlrK	[EC:2.7.13.3]	4,717525274	0,766666667	0,418918919	0,422952585	0,520050849	2,39E-06	0,009117127	1,83
			K11103	dctA	aerobic C4-dicarboxylate transport protein	NA	4,678340917	0,966666667	0,608108108	0,314466038	0,540709433	2,89E-06	0,011044755	1,59
			K11638	K11638, citT	two-component system, CitB family, response regulator CitT	NA	4,556387748	0,933333333	0,581081081	0,359010987	0,545417191	5,20E-06	0,019874393	1,61
			K14980	chvG	two-component system, OmpR family, sensor histidine kinase ChvG	[EC:2.7.13.3]	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
			K18348	vanT	serine/alanine racemase	[EC:5.1.1.18 5.1.1.1]	6,567670848	1,933333333	1,445945946	0,249443826	0,523550558	5,11E-11	1,95E-07	1,34
		04066 HIF-1 signaling pathway [PATH:ko04066]	K01689	ENO, eno	enolase	[EC:4.2.1.11]	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
A09140 Cellular Processes	09141 Transport and catabolism	04146 Peroxisome [PATH:ko04146]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
	09142 Cell motility	02030 Bacterial chemotaxis [PATH:ko02030]	K10439	rbsB	ribose transport system substrate-binding protein	NA	4,375171767	4,9	3,905405405	0,86986589	1,603787063	1,21E-05	0,04633856	1,25
	09143 Cell growth and death	04216 Ferroptosis [PATH:ko04216]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
	09145 Cellular community - prokaryotes	02024 Quorum sensing [PATH:ko02024]	K01497	ribA, RIB1	GTP cyclohydrolase II	[EC:3.5.4.25]	-4,847196857	0,333333333	0,689189189	0,471404521	0,517939673	1,25E-06	0,004782082	0,48
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
			K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	5,343506108	9,733333333	7,77027027	1,547758235	2,59181298	9,12E-08	0,000348162	1,25
			K01996	livF	branched-chain amino acid transport system ATP-binding protein	NA	6,212382198	10,56666667	7,864864865	1,282792094	3,068252097	5,22E-10	1,99E-06	1,34

			K07711	glrK, qseE	two-component system, NtrC family, sensor histidine kinase GlrK	[EC:2.7.13.3]	4,717525274	0,766666667	0,418918919	0,422952585	0,520050849	2,39E-06	0,009117127	1,83
		02026 Biofilm formation - Escherichia coli [PATH:ko02026]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	7,210590737	1,833333333	0,945945946	0,45338235	0,86823667	5,57E-13	2,13E-09	1,94
A09150 Organismal Systems	09152 Endocrine system	03320 PPAR signaling pathway [PATH:ko03320]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
		04920 Adipocytokine signaling pathway [PATH:ko04920]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
	09159 Environmental adaptation	04714 Thermogenesis [PATH:ko04714]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
A09160 Human Diseases	09161 Cancer: overview	05230 Central carbon metabolism in cancer [PATH:ko05230]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49 1.1.1.363]	NA	7,125804862	1,7	0,810810811	0,585946528	0,880351215	1,03E-12	3,95E-09	2,10
	09167 Endocrine and metabolic disease	04931 Insulin resistance [PATH:ko04931]	K00820	glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16]	NA	4,948390841	1,6	1,283783784	0,489897949	0,450833171	7,48E-07	0,00285774	1,25
	09174 Infectious disease: parasitic	05146 Amoebiasis [PATH:ko05146]	K13963	SERPINB	serpin B	NA	5,767633139	0,633333333	0,27027027	0,48189441	0,444099371	8,04E-09	3,07E-05	2,34
	09175 Drug resistance: antimicrobial	01502 Vancomycin resistance [PATH:ko01502]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	5,662246898	3,466666667	2,743243243	0,805536398	0,901362494	1,49E-08	5,71E-05	1,26
			K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	6,567670848	1,933333333	1,445945946	0,249443826	0,523550558	5,11E-11	1,95E-07	1,34
	09181 Protein families: metabolism	01001 Protein kinases [BR:ko01001]	K00936	pdtaS	two-component system, sensor histidine kinase PdtaS [EC:2.7.13.3]	NA	6,081666438	0,633333333	0,256756757	0,48189441	0,436844051	1,19E-09	4,54E-06	2,47
			K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ [EC:2.7.13.3]	NA	7,210590737	1,833333333	0,945945946	0,45338235	0,86823667	5,57E-13	2,13E-09	1,94
			K07641	creC	two-component system, OmpR family, sensor histidine kinase CreC [EC:2.7.13.3]	NA	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB [EC:2.7.13.3]	NA	5,186809286	0,633333333	0,297297297	0,48189441	0,457068501	2,14E-07	0,00081699	2,13
			K07654	mtrB	two-component system, OmpR family, sensor histidine kinase MtrB [EC:2.7.13.3]	NA	7,121500093	2,8	1,932432432	0,476095229	0,859463709	1,07E-12	4,08E-09	1,45
			K07711	glrK, qseE	two-component system, NtrC family, sensor histidine kinase GlrK [EC:2.7.13.3]	NA	4,717525274	0,766666667	0,418918919	0,422952585	0,520050849	2,39E-06	0,009117127	1,83
			K14980	chvG	two-component system, OmpR family, sensor histidine kinase ChvG [EC:2.7.13.3]	NA	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
		01002 Peptidases [BR:ko01002]	K00820	glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16]	NA	4,948390841	1,6	1,283783784	0,489897949	0,450833171	7,48E-07	0,00285774	1,25
			K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	6,621320139	1,633333333	0,72972973	0,795124029	0,962784494	3,56E-11	1,36E-07	2,24
			K01953	asnB, ASNS	asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]	NA	5,563134646	0,866666667	0,472972973	0,339934634	0,499269005	2,65E-08	0,000101193	1,83
			K05520	pfpI	protease I [EC:3.5.1.124]	NA	5,670111576	0,8	0,405405405	0,4	0,490970328	1,43E-08	5,45E-05	1,97
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	6,79778813	1,633333333	0,716216216	0,795124029	0,951815765	1,06E-11	4,06E-08	2,28
			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	6,79778813	1,633333333	0,716216216	0,795124029	0,951815765	1,06E-11	4,06E-08	2,28
		01004 Lipid biosynthesis proteins [BR:ko01004]	K00666	ACSF2	fatty-acyl-CoA synthase [EC:6.2.1.-]	NA	-4,910139081	2,933333333	4,135135135	0,727247474	1,726771118	9,10E-07	0,003475741	0,71
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
			K12429	fadD36	fatty acid CoA ligase FadD36	NA	-5,533007996	0,166666667	0,864864865	0,372677996	0,890252383	3,15E-08	0,000120216	0,19
		01007 Amino acid related enzymes [BR:ko01007]	K01870	IARS, ileS	isoleucyl-tRNA synthetase [EC:6.1.1.5]	NA	4,912316744	1,6	1,27027027	0,489897949	0,473551769	9,00E-07	0,003437348	1,26
		01008 Polyketide biosynthesis proteins [BR:ko01008]	K04780	dhbF	nonribosomal peptide synthetase DhbF	NA	-4,685253672	2,466666667	4,878378378	1,802467445	3,631516334	2,80E-06	0,010678436	0,51
		01009 Protein phosphatases and associated proteins [BR:ko01009]	K01251	E3.3.1.1, ahcY	adenosylhomocysteinase [EC:3.3.1.1]	NA	4,632720273	0,633333333	0,310810811	0,546707316	0,491156266	3,61E-06	0,013782466	2,04
		01011 Peptidoglycan biosynthesis and degradation proteins [BR:ko01011]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	5,662246898	3,466666667	2,743243243	0,805536398	0,901362494	1,49E-08	5,71E-05	1,26
	09182 Protein families: genetic information processing	03000 Transcription factors [BR:ko03000]	K02525	kdgR	LacI family transcriptional regulator, kdg operon repressor	NA	5,159379025	1,066666667	0,621621622	0,442216639	0,608558392	2,48E-07	0,000946235	1,72
			K02529	lacI, galR	LacI family transcriptional regulator	NA	6,135255332	22,06666667	16,94594595	3,151013946	5,888357482	8,50E-10	3,25E-06	1,30
			K05499	cytR	LacI family transcriptional regulator, repressor for deo operon, udp, cdd, tsx, nupC, and nupG	NA	6,048459671	3,333333333	2,108108108	1,105541597	1,429113953	1,46E-09	5,58E-06	1,58
			K10947	padR	PadR family transcriptional regulator, regulatory protein PadR	NA	5,921640065	1,5	0,837837838	0,619139187	0,788892947	3,19E-09	1,22E-05	1,79
		03016 Transfer RNA biogenesis [BR:ko03016]	K01870	IARS, ileS	isoleucyl-tRNA synthetase [EC:6.1.1.5]	NA	4,912316744	1,6	1,27027027	0,489897949	0,473551769	9,00E-07	0,003437348	1,26
		03019 Messenger RNA biogenesis [BR:ko03019]	K01689	ENO, eno	enolase [EC:4.2.1.11]	NA	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
		03021 Transcription machinery [BR:ko03021]	K03088	rpoE	RNA polymerase sigma-70 factor, ECF subfamily	NA	6,4904	28,3	21,78378378	3,328162656	7,083040785	8,56E-11	3,27E-07	1,30
			K19707	rsbQ	sigma-B regulation protein RsbQ	NA	5,470016087	0,633333333	0,283783784	0,48189441	0,450833171	4,50E-08	0,000171853	2,23
		03110 Chaperones and folding catalysts [BR:ko03110]	K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	6,621320139	1,633333333	0,72972973	0,795124029	0,962784494	3,56E-11	1,36E-07	2,24
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	6,79778813	1,633333333	0,716216216	0,795124029	0,951815765	1,06E-11	4,06E-08	2,28

A09180 Brite Hierarchies			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	6,79778813	1,633333333	0,716216216	0,795124029	0,951815765	1,06E-11	4,06E-08	2,28
	09183 Protein families: signaling and cellular processes	01504 Antimicrobial resistance genes [BR:ko01504]	K18230	tylC, oleB, carA, srmB	macrolide transport system ATP-binding/permease protein	NA	4,383920897	2,2	1,635135135	0,702376917	0,909028652	1,17E-05	0,044515071	1,35
			K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	6,567670848	1,933333333	1,445945946	0,249443826	0,523550558	5,11E-11	1,95E-07	1,34
			K19350	lsa	lincosamide and streptogramin A transport system ATP-binding/permease protein	NA	5,991641332	0,666666667	0,283783784	0,471404521	0,450833171	2,08E-09	7,93E-06	2,35
		02000 Transporters [BR:ko02000]	K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	5,343506108	9,733333333	7,77027027	1,547758235	2,59181298	9,12E-08	0,000348162	1,25
			K01996	livF	branched-chain amino acid transport system ATP-binding protein	NA	6,212382198	10,56666667	7,864864865	1,282792094	3,068252097	5,22E-10	1,99E-06	1,34
			K02004	ABC.CD.P	putative ABC transport system permease protein	NA	5,011762562	2,433333333	1,527027027	1,256538455	1,275793055	5,39E-07	0,002059729	1,59
			K02011	afuB, fbpB	iron(III) transport system permease protein	NA	-4,586726207	0,2	0,635135135	0,476095229	0,669294189	4,50E-06	0,017195067	0,31
			K02025	ABC.MS.P	multiple sugar transport system permease protein	NA	6,294276599	21,2	16,54054054	1,557776193	5,222583986	3,09E-10	1,18E-06	1,28
			K02026	ABC.MS.P1	multiple sugar transport system permease protein	NA	5,728148767	25,83333333	21,32432432	2,114762923	5,553445748	1,02E-08	3,88E-05	1,21
			K02027	ABC.MS.S	multiple sugar transport system substrate-binding protein	NA	6,469058021	22,43333333	17,2027027	2,02786149	5,704383639	9,86E-11	3,77E-07	1,30
			K02028	ABC.PA.A	polar amino acid transport system ATP-binding protein [EC:7.4.2.1]	NA	5,785082698	5,3	4,081081081	1,004987562	1,486486486	7,25E-09	2,77E-05	1,30
			K02056	ABC.SS.A	simple sugar transport system ATP-binding protein [EC:7.5.2.-]	NA	5,454950329	7,9	6,081081081	1,106044002	2,352438394	4,90E-08	0,00018708	1,30
			K02824	pyrP, uraA	uracil permease	NA	-5,64398312	0,1	0,5	0,3	0,5	1,66E-08	6,35E-05	0,20
			K03284	corA	magnesium transporter	NA	5,002553204	2,8	2,243243243	0,6	0,785180489	5,66E-07	0,002160636	1,25
			K03322	mntH	manganese transport protein	NA	5,280685202	0,933333333	0,540540541	0,249443826	0,524769942	1,29E-07	0,000491512	1,73
			K05816	ugpC	sn-glycerol 3-phosphate transport system ATP-binding protein [EC:7.6.2.10]	NA	5,005007787	0,566666667	0,256756757	0,495535625	0,436844051	5,59E-07	0,002133285	2,21
			K09016	rutG	putative pyrimidine permease RutG	NA	-5,64398312	0,1	0,5	0,3	0,5	1,66E-08	6,35E-05	0,20
			K10008	gluA	glutamate transport system ATP-binding protein [EC:7.4.2.1]	NA	6,432909509	4,5	3,27027027	0,991631652	1,348645941	1,25E-10	4,78E-07	1,38
			K10111	malK, mtlK, thuK	multiple sugar transport system ATP-binding protein [EC:3.6.3.-]	NA	5,3182487	0,566666667	0,243243243	0,495535625	0,429040753	1,05E-07	0,00040012	2,33
			K10112	msmX, msmK,	multiple sugar transport system ATP-binding protein	NA	5,675711802	1,066666667	0,486486486	0,727247474	0,72117103	1,38E-08	5,27E-05	2,19
			K10117	msmE	raffinose/stachyose/melibiose transport system substrate-binding protein	NA	6,2739803	4,566666667	2,945945946	0,955103252	1,822471532	3,52E-10	1,34E-06	1,55
			K10118	msmF	raffinose/stachyose/melibiose transport system permease protein	NA	5,61843792	5,4	4,027027027	1,113552873	1,72401929	1,93E-08	7,36E-05	1,34
			K10119	msmG	raffinose/stachyose/melibiose transport system permease protein	NA	5,400613175	7,133333333	5,486486486	1,203698006	2,151324188	6,64E-08	0,000253633	1,30
			K10232	aglE, ggtB	alpha-glucoside transport system substrate-binding protein	NA	7,002786036	1,733333333	0,864864865	0,573488351	0,8749413	2,51E-12	9,58E-09	2,00
			K10234	aglG, ggtD	alpha-glucoside transport system permease protein	NA	5,983263365	2,866666667	2,27027027	0,561743318	0,703222261	2,19E-09	8,35E-06	1,26
			K10238	thuG, sugB	trehalose/maltose transport system permease protein	NA	5,484026816	1,9	1,351351351	0,597215762	0,705814317	4,16E-08	0,000158776	1,41
			K10439	rbsB	ribose transport system substrate-binding protein	NA	4,375171767	4,9	3,905405405	0,86986589	1,603787063	1,21E-05	0,04633856	1,25
			K10440	rbsC	ribose transport system permease protein	NA	4,588838156	4,233333333	3,418918919	0,80346472	1,252098312	4,46E-06	0,01702204	1,24
			K10441	rbsA	ribose transport system ATP-binding protein [EC:7.5.2.7]	NA	5,684687642	4,566666667	3,243243243	1,308519095	1,642434011	1,31E-08	5,00E-05	1,41
			K10546	ABC.GGU.S, chvE	putative multiple sugar transport system substrate-binding protein	NA	6,584509939	1,966666667	1,189189189	0,406885187	0,833028919	4,56E-11	1,74E-07	1,65
			K10548	ABC.GGU.A, gguA	putative multiple sugar transport system ATP-binding protein [EC:7.5.2.-]	NA	5,604179502	4,066666667	2,77027027	1,123486636	1,632005045	2,09E-08	7,99E-05	1,47
			K10559	rhaS	rhamnose transport system substrate-binding protein	NA	4,532771498	0,5	0,22972973	0,5	0,420658984	5,82E-06	0,022232224	2,18
			K10561	rhaQ	rhamnose transport system permease protein	NA	4,863411594	0,5	0,216216216	0,5	0,411663411	1,15E-06	0,004406346	2,31
			K11103	dctA	aerobic C4-dicarboxylate transport protein	NA	4,678340917	0,966666667	0,608108108	0,314466038	0,540709433	2,89E-06	0,011044755	1,59
			K11963	urtE	urea transport system ATP-binding protein	NA	4,659777841	0,666666667	0,351351351	0,471404521	0,477392479	3,17E-06	0,012089077	1,90
			K15772	ganQ	arabinogalactan oligomer / maltooligosaccharide transport system permease protein	NA	4,847196857	1,666666667	1,310810811	0,53748385	0,517939673	1,25E-06	0,004782082	1,27
			K16301	efeB	deferrochelatase/peroxidase EfeB [EC:1.11.1.-]	NA	4,480206365	1,333333333	0,918918919	0,596284794	0,652578187	7,46E-06	0,028478628	1,45
			K17241	aguE	alpha-1,4-digalacturonate transport system substrate-binding protein	NA	6,133274208	0,8	0,378378378	0,4	0,484982661	8,61E-10	3,29E-06	2,11
			K17242	aguF	alpha-1,4-digalacturonate transport system permease protein	NA	4,659777841	0,666666667	0,351351351	0,471404521	0,477392479	3,17E-06	0,012089077	1,90

			K18230	tylC, oleB, carA, srmB	macrolide transport system ATP-binding/permease protein	NA	4,383920897	2,2	1,635135135	0,702376917	0,909028652	1,17E-05	0,044515071	1,35
			K19350	lsa	lincosamide and streptogramin A transport system ATP-binding/permease protein	NA	5,991641332	0,666666667	0,283783784	0,471404521	0,450833171	2,08E-09	7,93E-06	2,35
		02022 Two-component system [BR:ko02022]	K00936	pdtaS	two-component system, sensor histidine kinase PdtaS [EC:2.7.13.3]	NA	6,081666438	0,633333333	0,256756757	0,48189441	0,436844051	1,19E-09	4,54E-06	2,47
			K02475	K02475	two-component system, CitB family, response regulator	NA	4,556387748	0,933333333	0,581081081	0,359010987	0,545417191	5,20E-06	0,019874393	1,61
			K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ [EC:2.7.13.3]	NA	7,210590737	1,833333333	0,945945946	0,45338235	0,86823667	5,57E-13	2,13E-09	1,94
			K07641	creC	two-component system, OmpR family, sensor histidine kinase CreC [EC:2.7.13.3]	NA	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB [EC:2.7.13.3]	NA	5,186809286	0,633333333	0,297297297	0,48189441	0,457068501	2,14E-07	0,00081699	2,13
			K07654	mtrB	two-component system, OmpR family, sensor histidine kinase MtrB [EC:2.7.13.3]	NA	7,121500093	2,8	1,932432432	0,476095229	0,859463709	1,07E-12	4,08E-09	1,45
			K07711	glrK, qseE	two-component system, NtrC family, sensor histidine kinase GlrK [EC:2.7.13.3]	NA	4,717525274	0,766666667	0,418918919	0,422952585	0,520050849	2,39E-06	0,009117127	1,83
			K11638	K11638, citT	two-component system, CitB family, response regulator CitT	NA	4,556387748	0,933333333	0,581081081	0,359010987	0,545417191	5,20E-06	0,019874393	1,61
			K14980	chvG	two-component system, OmpR family, sensor histidine kinase ChvG [EC:2.7.13.3]	NA	5,191129548	0,766666667	0,405405405	0,422952585	0,490970328	2,09E-07	0,000798255	1,89
		02048 Prokaryotic defense system [BR:ko02048]	K07451	mcrA	5-methylcytosine-specific restriction enzyme A [EC:3.1.21.-]	NA	6,764435042	3,733333333	1,554054054	2,159217965	2,272882988	1,34E-11	5,11E-08	2,40
		04147 Exosome [BR:ko04147]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49 1.1.1.363]	NA	7,125804862	1,7	0,810810811	0,585946528	0,880351215	1,03E-12	3,95E-09	2,10
			K01251	E3.3.1.1, ahcY	adenosylhomocysteinase [EC:3.3.1.1]	NA	4,632720273	0,633333333	0,310810811	0,546707316	0,491156266	3,61E-06	0,013782466	2,04
			K01689	ENO, eno	enolase [EC:4.2.1.11]	NA	6,164100203	1,733333333	1,324324324	0,442216639	0,46812184	7,09E-10	2,71E-06	1,31
			K01810	GPI, pgi	glucose-6-phosphate isomerase [EC:5.3.1.9]	NA	-5,513573322	0,033333333	0,418918919	0,179505494	0,493382061	3,52E-08	0,000134284	0,08
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	-5,8116511	3,866666667	5,716216216	1,175679473	2,245236822	6,19E-09	2,36E-05	0,68
A09190 Not Included in Pathway or Brite	09191 Unclassified: metabolism	99980 Enzymes with EC numbers	K00257	mbtN, fadE14	acyl-ACP dehydrogenase	[EC:1.3.99.-]	6,059559011	3,9	3,054054054	0,3	0,984911738	1,36E-09	5,21E-06	1,28
			K00574	cfa	cyclopropane-fatty-acyl-phospholipid synthase	[EC:2.1.1.79]	4,659777841	0,666666667	0,351351351	0,471404521	0,477392479	3,17E-06	0,012089077	1,90
			K01811	xylS, yicI	alpha-D-xyloside xylohydrolase	[EC:3.2.1.17 7]	5,95401176	2,033333333	1,22972973	0,752034278	0,952199407	2,62E-09	9,99E-06	1,65
			K03518	coxS	aerobic carbon-monoxide dehydrogenase small subunit	[EC:1.2.5.3]	5,343253313	1,1	0,540540541	0,789514619	0,738683804	9,13E-08	0,000348648	2,03
			K03534	rhaM	L-rhamnose mutarotase	[EC:5.1.3.32]	4,657172554	1,1	0,662162162	0,3	0,663264389	3,21E-06	0,012243017	1,66
			K06221	dkgA	2,5-diketo-D-gluconate reductase A	[EC:1.1.1.34 6]	7,091826308	1,966666667	1,324324324	0,314466038	0,639004887	1,32E-12	5,05E-09	1,49
			K07302	iorA	isoquinoline 1-oxidoreductase subunit alpha	[EC:1.3.99.1 6]	5,343253313	1,1	0,540540541	0,789514619	0,738683804	9,13E-08	0,000348648	2,03
			K18581	ugl	unsaturated chondroitin disaccharide hydrolase	[EC:3.2.1.18 0]	5,470016087	0,633333333	0,283783784	0,48189441	0,450833171	4,50E-08	0,000171853	2,23
		99999 Others	K06999	K06999	phospholipase/carboxylesterase	NA	5,230831132	0,966666667	0,581081081	0,179505494	0,520050849	1,69E-07	0,000644455	1,66
	09193 Unclassified: signaling and cellular processes	99977 Transport	K03316	TC.CPA1	monovalent cation:H+ antiporter, CPA1 family	NA	4,415789608	2,6	2	0,553774924	0,958602587	1,01E-05	0,038435155	1,30
			K03458	TC.NCS2	nucleobase:cation symporter-2, NCS2 family	NA	-5,64398312	0,1	0,5	0,3	0,5	1,66E-08	6,35E-05	0,20
			K07011	wbbL	rhamnosyltransferase	NA	5,971368893	1,533333333	0,797297297	0,805536398	0,869602732	2,35E-09	8,99E-06	1,92
		99994 Others	K07214	fes	enterochelin esterase and related enzymes	NA	4,867916394	0,666666667	0,324324324	0,53748385	0,496150264	1,13E-06	0,004307107	2,06
			K07217	K07217	Mn-containing catalase	NA	6,859360239	1,433333333	0,662162162	0,61553951	0,793163816	6,92E-12	2,64E-08	2,16
			K07222	K07222	putative flavoprotein involved in K+ transport	NA	4,620969051	4,066666667	3,148648649	1,030641656	1,401567004	3,82E-06	0,014586738	1,29
	09194 Poorly characterized	99996 General function prediction only	K07001	K07001	NTE family protein	NA	4,418113521	1,533333333	1,013513514	0,669991708	0,830064199	9,96E-06	0,038024303	1,51
		99997 Function unknown	K06860	K06860	putative heme uptake system protein	NA	-5,513922927	0,3	0,797297297	0,458257569	0,636284182	3,51E-08	0,000134017	0,38
			K06975	K06975	uncharacterized protein	NA	5,442987795	2,766666667	1,891891892	1,085766498	1,133847403	5,24E-08	0,000200094	1,46
			K07126	K07126	uncharacterized protein	NA	4,466678461	0,533333333	0,256756757	0,498887652	0,436844051	7,94E-06	0,030339432	2,08
			K09992	K09992	uncharacterized protein	NA	6,926832381	10,4	6,040540541	2,847220867	4,440109535	4,30E-12	1,64E-08	1,72
	Total KEGG (unique) 0.05 q value	125 1.31E-05 p vaue												

Cluster II

CatA	CatB	CatC	KO	Short name	Long name	EC	v test	Mean in category	Overall mean	sd in category	Overall sd	p value	q value	Fold
A09100 Metabolism	09101 Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
			K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
		00030 Pentose phosphate pathway [PATH:ko00030]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase	[EC:1.1.1.49 1.1.1.363]	-5,809785964	0,125	0,810810811	0,330718914	0,880351215	6,26E-09	2,39E-05	0,15
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
		00051 Fructose and mannose metabolism [PATH:ko00051]	K00064	E1.1.1.122	D-threo-aldose 1-dehydrogenase	[EC:1.1.1.122]	-4,669879747	0,3125	1,054054054	0,463512405	1,184264989	3,01E-06	0,011509553	0,30
		00052 Galactose metabolism [PATH:ko00052]	K01187	malZ	alpha-glucosidase	[EC:3.2.1.20]	-4,891851601	0	0,432432432	0	0,65926005	9,99E-07	0,003814866	0,00
		00053 Ascorbate and aldarate metabolism [PATH:ko00053]	K00064	E1.1.1.122	D-threo-aldose 1-dehydrogenase	[EC:1.1.1.122]	-4,669879747	0,3125	1,054054054	0,463512405	1,184264989	3,01E-06	0,011509553	0,30
			K13875	K13875, araC	L-arabonate dehydrase	[EC:4.2.1.25]	-5,652554094	0	0,364864865	0	0,481392247	1,58E-08	6,04E-05	0,00
		00500 Starch and sucrose metabolism [PATH:ko00500]	K01187	malZ	alpha-glucosidase	[EC:3.2.1.20]	-4,891851601	0	0,432432432	0	0,65926005	9,99E-07	0,003814866	0,00
			K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
		00520 Amino sugar and nucleotide sugar metabolism [PATH:ko00520]	K01810	GPI, pgi	glucose-6-phosphate isomerase	[EC:5.3.1.9]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
			K15921	xynD	arabinoxylan arabinofuranohydrolase	[EC:3.2.1.55]	-4,609024815	0,09375	0,581081081	0,291480595	0,788545647	4,05E-06	0,015450225	0,16
		00620 Pyruvate metabolism [PATH:ko00620]	K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
		00630 Glyoxylate and dicarboxylate metabolism [PATH:ko00630]	K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
		00640 Propanoate metabolism [PATH:ko00640]	K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
			K01908	prpE	propionyl-CoA synthetase	[EC:6.2.1.17]	4,791592552	0,5	0,22972973	0,5	0,420658984	1,65E-06	0,006319019	2,18
	09102 Energy metabolism	00680 Methane metabolism [PATH:ko00680]	K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
		00720 Carbon fixation pathways in prokaryotes [PATH:ko00720]	K01895	ACSS, acs	acetyl-CoA synthetase	[EC:6.2.1.1]	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
		00910 Nitrogen metabolism [PATH:ko00910]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
	09103 Lipid metabolism	00061 Fatty acid biosynthesis [PATH:ko00061]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
			K18660	ACSF3	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	4,880571283	1,4375	0,851351351	0,826797285	0,895671761	1,06E-06	0,004039699	1,69
		00071 Fatty acid degradation [PATH:ko00071]	K00496	alkB1_2, alkM	alkane 1-monooxygenase	[EC:1.14.15.3]	4,514192505	0,65625	0,364864865	0,47495888	0,481392247	6,36E-06	0,024272978	1,80
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09105 Amino acid metabolism	00220 Arginine biosynthesis [PATH:ko00220]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
			K01755	argH, ASL	argininosuccinate lyase	[EC:4.3.2.1]	4,473063202	0,84375	0,459459459	0,711924109	0,640717275	7,71E-06	0,029447154	1,84
		00250 Alanine, aspartate and glutamate metabolism [PATH:ko00250]	K00262	E1.4.1.4, gdhA	glutamate dehydrogenase (NADP+)	[EC:1.4.1.4]	4,532158291	0,71875	0,418918919	0,449609205	0,493382061	5,84E-06	0,022296877	1,72
			K01755	argH, ASL	argininosuccinate lyase	[EC:4.3.2.1]	4,473063202	0,84375	0,459459459	0,711924109	0,640717275	7,71E-06	0,029447154	1,84
		00270 Cysteine and methionine metabolism [PATH:ko00270]	K01243	mtnN, mtn, pfs	adenosylhomocysteine nucleosidase	[EC:3.2.2.9]	-5,387523377	0,0625	0,418918919	0,242061459	0,493382061	7,14E-08	0,000272811	0,15
		00280 Valine, leucine and isoleucine degradation [PATH:ko00280]	K18660	ACSF3	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	4,880571283	1,4375	0,851351351	0,826797285	0,895671761	1,06E-06	0,004039699	1,69
			K18661	matB	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	4,774008114	1,59375	0,945945946	1,026504232	1,011980823	1,81E-06	0,006896927	1,68
			K18662	matB	malonyl-CoA/methylmalonyl-CoA synthetase	[EC:6.2.1.-]	4,797076237	1,4375	0,864864865	0,826797285	0,890252383	1,61E-06	0,006148527	1,66
		00290 Valine, leucine and isoleucine biosynthesis [PATH:ko00290]	K01687	ilvD	dihydroxy-acid dehydratase	[EC:4.2.1.9]	-5,246772296	1	1,378378378	0,25	0,537831047	1,55E-07	0,000591132	0,73
	09106 Metabolism of other amino acids	00473 D-Alanine metabolism [PATH:ko00473]	K01775	alr	alanine racemase	[EC:5.1.1.1]	-5,115306601	2,125	2,743243243	0,544862368	0,901362494	3,13E-07	0,001196232	0,77
		00480 Glutathione metabolism [PATH:ko00480]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase	[EC:1.1.1.49 1.1.1.363]	-5,809785964	0,125	0,810810811	0,330718914	0,880351215	6,26E-09	2,39E-05	0,15
	09108 Metabolism of cofactors and vitamins	00770 Pantothenate and CoA biosynthesis [PATH:ko00770]	K01687	ilvD	dihydroxy-acid dehydratase	[EC:4.2.1.9]	-5,246772296	1	1,378378378	0,25	0,537831047	1,55E-07	0,000591132	0,73
		00790 Folate biosynthesis [PATH:ko00790]	K07141	mocA	molybdenum cofactor cytidylyltransferase	[EC:2.7.7.76]	-4,875038722	2	2,972972973	0,790569415	1,488450791	1,09E-06	0,00415458	0,67
	09109 Metabolism of terpenoids and polyketides	00523 Polyketide sugar unit biosynthesis [PATH:ko00523]	K13315	eryBII, tylCII, tylC1,	NDP-hexose C3-ketoreductase / dTDP-4-oxo-2-deoxy-alpha-D-pentos-2-ene 2,3-reductase	[EC:1.1.1.-]	6,009604727	0,9375	0,513513514	0,347985273	0,526160063	1,86E-09	7,10E-06	1,83

	09110 Biosynthesis of other secondary metabolites	00960 Tropane, piperidine and pyridine alkaloid biosynthesis [PATH:ko00960]	K08081	TR1	tropinone reductase I	[EC:1.1.1.206]	-4,761387135	0	0,405405405	0	0,634991358	1,92E-06	0,00734267	0,00
	09111 Xenobiotics biodegradation and metabolism	00930 Caprolactam degradation [PATH:ko00930]	K00496	alkB1_2, alkM	alkane 1-monooxygenase	[EC:1.14.15.3]	4,514192505	0,65625	0,364864865	0,47495888	0,481392247	6,36E-06	0,024272978	1,80
A09130 Environmental Information Processing	09131 Membrane transport	02010 ABC transporters [PATH:ko02010]	K02011	afuB, fbpB	iron(III) transport system permease protein	NA	4,41383262	1,03125	0,635135135	0,6366109	0,669294189	1,02E-05	0,038784423	1,62
			K02012	afuA, fbpA	iron(III) transport system substrate-binding protein	NA	4,527960859	0,625	0,337837838	0,484122918	0,472972973	5,96E-06	0,022744287	1,85
			K10112	msmX, msmK,	multiple sugar transport system ATP-binding protein	NA	-4,384554937	0,0625	0,486486486	0,242061459	0,72117103	1,16E-05	0,044385621	0,13
			K10232	aglE, ggtB	alpha-glucoside transport system substrate-binding protein	NA	-4,708243361	0,3125	0,864864865	0,463512405	0,8749413	2,50E-06	0,00954218	0,36
			K11081	phnS	2-aminoethylphosphonate transport system substrate-binding protein	NA	4,527960859	0,625	0,337837838	0,484122918	0,472972973	5,96E-06	0,022744287	1,85
	09132 Signal transduction	02020 Two-component system [PATH:ko02020]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB	[EC:2.7.13.3]	-4,850890179	0	0,297297297	0	0,457068501	1,23E-06	0,004693879	0,00
			K18348	vanT	serine/alanine racemase	[EC:5.1.1.18 5.1.1.1]	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
A09140 Cellular Processes	09141 Transport and catabolism	04146 Peroxisome [PATH:ko04146]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09143 Cell growth and death	04216 Ferroptosis [PATH:ko04216]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09145 Cellular community - prokaryotes	02024 Quorum sensing [PATH:ko02024]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
		02026 Biofilm formation - Escherichia coli [PATH:ko02026]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
A09150 Organismal Systems	09152 Endocrine system	03320 PPAR signaling pathway [PATH:ko03320]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
		04920 Adipocytokine signaling pathway [PATH:ko04920]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09159 Environmental adaptation	04714 Thermogenesis [PATH:ko04714]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
A09160 Human Diseases	09161 Cancer: overview	05230 Central carbon metabolism in cancer [PATH:ko05230]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49 1.1.1.363]	NA	-5,809785964	0,125	0,810810811	0,330718914	0,880351215	6,26E-09	2,39E-05	0,15
	09175 Drug resistance: antimicrobial	01502 Vancomycin resistance [PATH:ko01502]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	-5,115306601	2,125	2,743243243	0,544862368	0,901362494	3,13E-07	0,001196232	0,77
			K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
A09180 Brite Hierarchies	09181 Protein families: metabolism	01001 Protein kinases [BR:ko01001]	K00936	pdtaS	two-component system, sensor histidine kinase PdtaS [EC:2.7.13.3]	NA	-4,383360904	0	0,256756757	0	0,436844051	1,17E-05	0,044629703	0,00
			K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ [EC:2.7.13.3]	NA	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB [EC:2.7.13.3]	NA	-4,850890179	0	0,297297297	0	0,457068501	1,23E-06	0,004693879	0,00
		01002 Peptidases [BR:ko01002]	K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	-4,442227118	0,15625	0,72972973	0,506789836	0,962784494	8,90E-06	0,034001515	0,21
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
		01004 Lipid biosynthesis proteins [BR:ko01004]	K00666	ACSF2	fatty-acyl-CoA synthase [EC:6.2.1.-]	NA	4,680065722	5,21875	4,135135135	1,653299258	1,726771118	2,87E-06	0,010952242	1,26
			K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	NA	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
			K01908	prpE	propionyl-CoA synthetase [EC:6.2.1.17]	NA	4,791592552	0,5	0,22972973	0,5	0,420658984	1,65E-06	0,006319019	2,18
			K12429	fadD36	fatty acid CoA ligase FadD36	NA	4,797076237	1,4375	0,864864865	0,826797285	0,890252383	1,61E-06	0,006148527	1,66
		01011 Peptidoglycan biosynthesis and degradation proteins [BR:ko01011]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	-5,115306601	2,125	2,743243243	0,544862368	0,901362494	3,13E-07	0,001196232	0,77
	09182 Protein families: genetic information processing	03000 Transcription factors [BR:ko03000]	K02525	kdgR	LacI family transcriptional regulator, kdg operon repressor	NA	-4,937148556	0,21875	0,621621622	0,483113276	0,608558392	7,93E-07	0,003027439	0,35
		03021 Transcription machinery [BR:ko03021]	K19707	rsbQ	sigma-B regulation protein RsbQ	NA	-4,694436691	0	0,283783784	0	0,450833171	2,67E-06	0,010209806	0,00
		03110 Chaperones and folding catalysts [BR:ko03110]	K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	-4,442227118	0,15625	0,72972973	0,506789836	0,962784494	8,90E-06	0,034001515	0,21
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
		01504 Antimicrobial resistance genes [BR:ko01504]	K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
			K02011	afuB, fbpB	iron(III) transport system permease protein	NA	4,41383262	1,03125	0,635135135	0,6366109	0,669294189	1,02E-05	0,038784423	1,62

	09110 Biosynthesis of other secondary metabolites	00960 Tropane, piperidine and pyridine alkaloid biosynthesis [PATH:ko00960]	K08081	TR1	tropinone reductase I	[EC:1.1.1.206]	-4,761387135	0	0,405405405	0	0,634991358	1,92E-06	0,00734267	0,00
	09111 Xenobiotics biodegradation and metabolism	00930 Caprolactam degradation [PATH:ko00930]	K00496	alkB1_2, alkM	alkane 1-monooxygenase	[EC:1.14.15.3]	4,514192505	0,65625	0,364864865	0,47495888	0,481392247	6,36E-06	0,024272978	1,80
A09130 Environmental Information Processing	09131 Membrane transport	02010 ABC transporters [PATH:ko02010]	K02011	afuB, fbpB	iron(III) transport system permease protein	NA	4,41383262	1,03125	0,635135135	0,6366109	0,669294189	1,02E-05	0,038784423	1,62
			K02012	afuA, fbpA	iron(III) transport system substrate-binding protein	NA	4,527960859	0,625	0,337837838	0,484122918	0,472972973	5,96E-06	0,022744287	1,85
			K10112	msmX, msmK,	multiple sugar transport system ATP-binding protein	NA	-4,384554937	0,0625	0,486486486	0,242061459	0,72117103	1,16E-05	0,044385621	0,13
			K10232	aglE, ggtB	alpha-glucoside transport system substrate-binding protein	NA	-4,708243361	0,3125	0,864864865	0,463512405	0,8749413	2,50E-06	0,00954218	0,36
			K11081	phnS	2-aminoethylphosphonate transport system substrate-binding protein	NA	4,527960859	0,625	0,337837838	0,484122918	0,472972973	5,96E-06	0,022744287	1,85
	09132 Signal transduction	02020 Two-component system [PATH:ko02020]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB	[EC:2.7.13.3]	-4,850890179	0	0,297297297	0	0,457068501	1,23E-06	0,004693879	0,00
			K18348	vanT	serine/alanine racemase	[EC:5.1.1.18 5.1.1.1]	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
A09140 Cellular Processes	09141 Transport and catabolism	04146 Peroxisome [PATH:ko04146]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09143 Cell growth and death	04216 Ferroptosis [PATH:ko04216]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09145 Cellular community - prokaryotes	02024 Quorum sensing [PATH:ko02024]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase	[EC:6.2.1.3]	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
		02026 Biofilm formation - Escherichia coli [PATH:ko02026]	K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ	[EC:2.7.13.3]	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
A09150 Organismal Systems	09152 Endocrine system	03320 PPAR signaling pathway [PATH:ko03320]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
		04920 Adipocytokine signaling pathway [PATH:ko04920]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
	09159 Environmental adaptation	04714 Thermogenesis [PATH:ko04714]	K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
A09160 Human Diseases	09161 Cancer: overview	05230 Central carbon metabolism in cancer [PATH:ko05230]	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49 1.1.1.363]	NA	-5,809785964	0,125	0,810810811	0,330718914	0,880351215	6,26E-09	2,39E-05	0,15
	09175 Drug resistance: antimicrobial	01502 Vancomycin resistance [PATH:ko01502]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	-5,115306601	2,125	2,743243243	0,544862368	0,901362494	3,13E-07	0,001196232	0,77
			K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
A09180 Brite Hierarchies	09181 Protein families: metabolism	01001 Protein kinases [BR:ko01001]	K00936	pdtaS	two-component system, sensor histidine kinase PdtaS [EC:2.7.13.3]	NA	-4,383360904	0	0,256756757	0	0,436844051	1,17E-05	0,044629703	0,00
			K07638	envZ	two-component system, OmpR family, osmolarity sensor histidine kinase EnvZ [EC:2.7.13.3]	NA	-4,367354431	0,4375	0,945945946	0,496078371	0,86823667	1,26E-05	0,048027958	0,46
			K07653	mprB	two-component system, OmpR family, sensor histidine kinase MprB [EC:2.7.13.3]	NA	-4,850890179	0	0,297297297	0	0,457068501	1,23E-06	0,004693879	0,00
		01002 Peptidases [BR:ko01002]	K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	-4,442227118	0,15625	0,72972973	0,506789836	0,962784494	8,90E-06	0,034001515	0,21
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
		01004 Lipid biosynthesis proteins [BR:ko01004]	K00666	ACSF2	fatty-acyl-CoA synthase [EC:6.2.1.-]	NA	4,680065722	5,21875	4,135135135	1,653299258	1,726771118	2,87E-06	0,010952242	1,26
			K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	NA	4,537772768	0,65625	0,310810811	0,68962739	0,56772842	5,69E-06	0,021711577	2,11
			K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	NA	4,990842489	7,21875	5,716216216	1,866386465	2,245236822	6,01E-07	0,002295849	1,26
			K01908	prpE	propionyl-CoA synthetase [EC:6.2.1.17]	NA	4,791592552	0,5	0,22972973	0,5	0,420658984	1,65E-06	0,006319019	2,18
			K12429	fadD36	fatty acid CoA ligase FadD36	NA	4,797076237	1,4375	0,864864865	0,826797285	0,890252383	1,61E-06	0,006148527	1,66
		01011 Peptidoglycan biosynthesis and degradation proteins [BR:ko01011]	K01775	alr	alanine racemase [EC:5.1.1.1]	NA	-5,115306601	2,125	2,743243243	0,544862368	0,901362494	3,13E-07	0,001196232	0,77
	09182 Protein families: genetic information processing	03000 Transcription factors [BR:ko03000]	K02525	kdgR	LacI family transcriptional regulator, kdg operon repressor	NA	-4,937148556	0,21875	0,621621622	0,483113276	0,608558392	7,93E-07	0,003027439	0,35
		03021 Transcription machinery [BR:ko03021]	K19707	rsbQ	sigma-B regulation protein RsbQ	NA	-4,694436691	0	0,283783784	0	0,450833171	2,67E-06	0,010209806	0,00
		03110 Chaperones and folding catalysts [BR:ko03110]	K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	NA	-4,442227118	0,15625	0,72972973	0,506789836	0,962784494	8,90E-06	0,034001515	0,21
			K08652	C5AP, scpA, scpB	C5a peptidase [EC:3.4.21.110]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
			K14647	vpr	minor extracellular serine protease Vpr [EC:3.4.21.-]	NA	-4,632391093	0,125	0,716216216	0,414578099	0,951815765	3,61E-06	0,013804405	0,17
		01504 Antimicrobial resistance genes [BR:ko01504]	K18348	vanT	serine/alanine racemase [EC:5.1.1.18 5.1.1.1]	NA	-5,462070459	1,0625	1,445945946	0,242061459	0,523550558	4,71E-08	0,000179727	0,73
			K02011	afuB, fbpB	iron(III) transport system permease protein	NA	4,41383262	1,03125	0,635135135	0,6366109	0,669294189	1,02E-05	0,038784423	1,62

Cluster III

CatA	CatB	CatC	KO	Short name	Long name	EC	v test	Mean in category	Overall mean	sd in category	Overall sd	p value	q value	Fold
A09100 Metabolism	09101 Carbohydrate metabolism	00030 Pentose phosphate pathway [PATH:ko00030]	K00851	7.1.12, gntK, i	gluconokinase	[EC:2.7.1.12]	-4,6675351	0,416666667	1,3648649	0,493006649	0,7636032	3,05E-06	0,0116416	0,31
		00040 Pentose and glucuronate interconversions [PATH:ko00040]	K01804	araA	L-arabinose isomerase	[EC:5.3.1.4]	-4,8851846	0,333333333	0,9054054	0,471404521	0,4401756	1,03E-06	0,0039462	0,37
		00650 Butanoate metabolism [PATH:ko00650]	K01715	crt	enoyl-CoA hydratase	[EC:4.2.1.17]	4,4411447	0,583333333	0,1351351	0,640095479	0,3793424	8,95E-06	0,034173	4,32
	09111 Xenobiotics biodegradation and metabolism	00361 Chlorocyclohexane and chlorobenzene degradation [PATH:ko00361]	K07104	catE	catechol 2,3-dioxygenase	[EC:1.13.11.2	-5,227898	0,583333333	0,9324324	0,493006649	0,2510024	1,71E-07	0,0006548	0,63
		00362 Benzoate degradation [PATH:ko00362]	K07104	catE	catechol 2,3-dioxygenase	[EC:1.13.11.2	-5,227898	0,583333333	0,9324324	0,493006649	0,2510024	1,71E-07	0,0006548	0,63
		00622 Xylene degradation [PATH:ko00622]	K07104	catE	catechol 2,3-dioxygenase	[EC:1.13.11.2	-5,227898	0,583333333	0,9324324	0,493006649	0,2510024	1,71E-07	0,0006548	0,63
		00643 Styrene degradation [PATH:ko00643]	K07104	catE	catechol 2,3-dioxygenase	[EC:1.13.11.2	-5,227898	0,583333333	0,9324324	0,493006649	0,2510024	1,71E-07	0,0006548	0,63
		00984 Steroid degradation [PATH:ko00984]	K05898	kstD	3-oxosteroid 1-dehydrogenase	[EC:1.3.99.4]	4,4905788	0,666666667	0,1351351	0,849836586	0,444921	7,10E-06	0,0271263	4,93
A09130 Environmental Information Processing	09131 Membrane transport	02010 ABC transporters [PATH:ko02010]	K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	-4,6219582	4,583333333	7,7702703	2,430992024	2,591813	3,80E-06	0,0145173	0,59
			K10118	msmF	raffinose/stachyose/melibiose transport system permease protein	NA	-4,7828865	1,833333333	4,027027	0,897527468	1,7240193	1,73E-06	0,0065991	0,46
			K10119	msmG	raffinose/stachyose/melibiose transport system permease protein	NA	-5,0724794	2,583333333	5,4864865	1,656217243	2,1513242	3,93E-07	0,0014996	0,47
			K10238	thuG, sugB	trehalose/maltose transport system permease protein	NA	-4,5339275	0,5	1,3513514	0,5	0,7058143	5,79E-06	0,0221108	0,37
			K17209	iatP	inositol transport system permease protein	NA	-5,1684332	0,5	0,9459459	0,5	0,3243243	2,36E-07	0,0009015	0,53
A09140 Cellular Processes	09143 Cell growth and death	04112 Cell cycle - Caulobacter [PATH:ko04112]	K01338	lon	ATP-dependent Lon protease	[EC:3.4.21.53	-5,3098333	0,416666667	0,8783784	0,493006649	0,3268483	1,10E-07	0,000419	0,47
	09145 Cellular community - prokaryotes	02024 Quorum sensing [PATH:ko02024]	K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	-4,6219582	4,583333333	7,7702703	2,430992024	2,591813	3,80E-06	0,0145173	0,59
A09180 Brite Hierarchies	09181 Protein families: metabolism	01002 Peptidases [BR:ko01002]	K01338	lon	ATP-dependent Lon protease [EC:3.4.21.53]	NA	-5,3098333	0,416666667	0,8783784	0,493006649	0,3268483	1,10E-07	0,000419	0,47
	09182 Protein families: genetic information processing	03000 Transcription factors [BR:ko03000]	K02529	lacI, galR	LacI family transcriptional regulator	NA	-5,4979015	8,333333333	16,945946	3,771236166	5,8883575	3,84E-08	0,0001468	0,49
	09183 Protein families: signaling and cellular processes	02000 Transporters [BR:ko02000]	K01995	livG	branched-chain amino acid transport system ATP-binding protein	NA	-4,6219582	4,583333333	7,7702703	2,430992024	2,591813	3,80E-06	0,0145173	0,59
			K02025	ABC.MS.P	multiple sugar transport system permease protein	NA	-6,2068775	7,916666667	16,540541	2,430992024	5,222584	5,40E-10	2,06E-06	0,48
			K02026	ABC.MS.P1	multiple sugar transport system permease protein	NA	-6,2547826	12,08333333	21,324324	3,302986864	5,5534457	3,98E-10	1,52E-06	0,57
			K02027	ABC.MS.S	multiple sugar transport system substrate-binding protein	NA	-6,1738752	7,833333333	17,202703	2,823512391	5,7043836	6,66E-10	2,54E-06	0,46
			K02057	ABC.SS.P	simple sugar transport system permease protein	NA	-4,5168987	4,083333333	5,8648649	1,605113357	1,4825501	6,28E-06	0,023965	0,70
			K10118	msmF	raffinose/stachyose/melibiose transport system permease protein	NA	-4,7828865	1,833333333	4,027027	0,897527468	1,7240193	1,73E-06	0,0065991	0,46
			K10119	msmG	raffinose/stachyose/melibiose transport system permease protein	NA	-5,0724794	2,583333333	5,4864865	1,656217243	2,1513242	3,93E-07	0,0014996	0,47
			K10238	thuG, sugB	trehalose/maltose transport system permease protein	NA	-4,5339275	0,5	1,3513514	0,5	0,7058143	5,79E-06	0,0221108	0,37
			K17209	iatP	inositol transport system permease protein	NA	-5,1684332	0,5	0,9459459	0,5	0,3243243	2,36E-07	0,0009015	0,53
A09190 Not Included in Pathway or Brite	09191 Unclassified: metabolism	99980 Enzymes with EC numbers	K06221	dkgA	2,5-diketo-D-gluconate reductase A	[EC:1.1.1.346	-4,3587801	0,583333333	1,3243243	0,493006649	0,6390049	1,31E-05	0,0499485	0,44
	09193 Unclassified: signaling and cellular processes	99977 Transport	K03316	TC.CPA1	monovalent cation:H+ antiporter, CPA1 family	NA	-4,9014844	0,75	2	0,829156198	0,9586026	9,51E-07	0,0036325	0,38

Total KEGG (unique)	18
0.05 q value	1.31E-05 p vaue

APPENDIX III

Genome sequence accession numbers of strains used in the Chapter III.

Strain	**Accession number**
Actinoplanes derwentensis DSM 43941	NZ_LT629758
Actinoplanes globisporus DSM 43857^T	NZ_ARBJ00000000
Actinoplanes missouriensis NBRC 102363^T	NC_017093
Actinoplanes utahensis NRRL 12052	NZ_JRTT00000000
Catelliglobosispora koreensis DSM 44566^T	NZ_ARBL00000000
GAR05	PXXW00000000
GAR06	PYAH00000000
Hamadaea tsunoensis DSM 44101^T	NZ_AUAX00000000
LAH08	PYAA00000000
Longispora albida DSM 44784^T	NZ_ARBS00000000
Lupac 06	PYAJ00000000
Lupac 07	PYAB00000000
MED15	PYAC00000000
Micromonospora aurantiaca ATCC 27029^T	NC_014391
Micromonospora aurantiaca DSM 45487	NZ_FMHX00000000
Micromonospora aurantiaca L5	NC_014815
Micromonospora auratinigra DSM 44815^T	NZ_LT594323
Micromonospora avicenniae DSM 45758^T	NZ_FTNF00000000
Micromonospora carbonacea DSM 43168^T	NZ_FMCT00000000
Micromonospora chaiyaphumensis DSM 45246^T	NZ_FMCS00000000
Micromonospora chalcea DSM 43026^T	MAGP00000000
Micromonospora chersina DSM 44151^T	NZ_FMIB00000000
Micromonospora chokoriensis DSM 45160^T	NZ_LT607409
Micromonospora citrea DSM 43903^T	NZ_FMHZ00000000
Micromonospora coriariae DSM 44875^T	NZ_LT607412
Micromonospora coxensis DSM 45161^T	NZ_LT607753
Micromonospora cremea DSM 45599^T	NZ_FSQT00000000
Micromonospora eburnea DSM 44814^T	NZ_FMHY0000000
Micromonospora echinaurantiaca DSM 43904^T	NZ_LT607750
Micromonospora echinofusca DSM 43913^T	NZ_LT607733
Micromonospora echinospora DSM 1040	jgi_1043257
Micromonospora echinospora DSM 43816^T	NZ_LT607413
Micromonospora endolithica DSM 44398^T	jgi_1043245
Micromonospora haikouensis DSM 45626^T	NZ_FMCW00000000

Micromonospora halophytica DSM 43171^{T}	NZ_FMDN00000000
Micromonospora humi DSM 45647^{T}	NZ_FMDM00000000
Micromonospora inositola DSM 43819^{T}	NZ_LT607754
Micromonospora inyonensis DSM 46123^{T}	NZ_FMHU00000000
Micromonospora krabiensis DSM 45344^{T}	NZ_LT598496
Micromonospora lupini Lupac 08	NZ_CAIE00000000
Micromonospora marina DSM 45555^{T}	NZ_FMCV00000000
Micromonospora matsumotoense DSM 44100^{T}	NZ_FMCU00000000
Micromonospora mirobrigensis DSM 44830^{T}	NZ_FMCX00000000
Micromonospora narathiwatensis DSM 45248^{T}	NZ_LT594324
Micromonospora nigra DSM 43818^{T}	NZ_FMHT00000000
Micromonospora noduli GUI43^{T}	PYAK00000000
Micromonospora olivasterospora DSM 43868^{T}	jgi_1043242
Micromonospora pallida DSM 43817^{T}	NZ_FMHW00000000
Micromonospora pattaloongensis DSM 45245^{T}	NZ_FNPH00000000
Micromonospora peucetia DSM 43363^{T}	NZ_FMIC00000000
Micromonospora pisi DSM 45175^{T}	jgi_1067754
Micromonospora purpureochromogenes DSM 43821^{T}	NZ_LT607410
Micromonospora rhizosphaerae DSM 45431^{T}	NZ_FMHV00000000
Micromonospora rifamycinica DSM 44983T	NZ_LT607752
Micromonospora rosaria DSM 803^{T}	NZ_LRQV00000000
Micromonospora saelicesensis DSM 44871^{T}	NZ_FMCR00000000
Micromonospora sagamiensis DSM 43912^{T}	jgi_1043248
Micromonospora sediminicola DSM 45794^{T}	NZ_FLRH00000000
Micromonospora siamensis DSM 45097^{T}	NZ_LT607751
Micromonospora tulbaghiae DSM 45142^{T}	NZ_FMCQ00000000
Micromonospora viridifaciens DSM 43909^{T}	NZ_LT607411
Micromonospora wenchangensis CCTCC AA 2012002^{T}	NZ_MZMV00000000
Micromonospora yangpuensis DSM 45577^{T}	NZ_FMIA00000000
Micromonospora zamorensis DSM 45600^{T}	NZ_LT607755
ONO23	PYAD00000000
ONO86	PYAE00000000
PSN01	PYAI00000000
PSN13	PYAG00000000
Salinispora arenicola CNH643^{T}	AIIF00000000
Salinispora pacifica CNR114^{T}	NZ_AZWO00000000

APPENDIX IV

Pairwise similarity values based on a concatenation of 16S rRNA, *gyr*B, *rpo*B, *atp*D and *rec*A genes. A total of 4340 nucleotide positions were used to calculate the matrix.

The table is too big to see in one page. The whole table can be seen in https://drive.google.com/file/d/1nGH-jMsyvRUwtco_xkfstbJ0HGasxI4B9/view?usp=sharing

APPENDIX V

ANI, OrthoANI, dDDH and G+C differences. A: Upper half of the matrix: Pairwise ANI values; Bottom half: Pairwise OrthoANI values. B: Upper half: G+C mol% differences; Bottom half: Pairwise dDDH values.

A

		Group I						Group II						Single strain groups					
		DSM 44871^{T}	Lupac 06	GAR05	GAR06	PSN01	PSN13	MED15	ONO23	ONO86	GUI43^{T}	Lupac 07	LAH08	MED01	NIE79	NIE111	PSH03	PSH25	LAH09
Group I	DSM 44871^{T}	-	99.13	99.11	99.09	99.12	97.82	96.59	96.62	96.59	96.64	96.59	96.23	94.9	94.88	95.28	95.69	94.34	90.27
	Lupac 06	99.18	-	99.19	99.1	99.16	97.87	96.61	96.63	96.6	96.63	96.61	96.63	94.89	94.87	95.31	95.71	94.39	90.25
	GAR05	99.15	99.25	-	99.18	99.16	97.8	96.61	96.61	96.54	96.68	96.55	96.63	94.94	94.83	95.28	95.71	94.35	90.78
	GAR06	99.18	99.18	99.25	-	99.13	97.78	96.52	96.54	96.52	96.61	96.52	96.54	94.9	94.81	95.25	95.66	94.33	90.2
	PSN01	99.19	99.21	99.21	99.21	-	97.76	96.65	96.65	96.58	96.61	96.63	96.6	94.86	94.91	95.31	95.75	94.4	90.36
	PSN13	97.96	98.01	97.97	97.93	97.9	-	96.59	96.58	96.56	96.55	96.58	96.58	94.83	94.75	95.21	95.61	94.34	90.1
Group II	MED15	96.78	96.82	96.84	96.71	96.82	96.78	-	99.16	99.16	99.09	99.17	99.16	94.84	94.75	95.27	95.64	94.28	90.09
	ONO23	96.86	96.85	96.8	96.76	96.9	96.83	99.2	-	99.14	99.05	99.07	99.19	94.8	94.72	95.25	95.63	94.24	90.07
	ONO86	96.23	96.8	96.81	96.81	96.7	96.81	99.19	99.22	-	99.09	99.17	99.13	94.77	94.7	95.28	95.56	94.25	90.08
	GUI43^{T}	96.82	96.81	96.81	96.78	96.81	96.78	99.13	99.12	99.13	-	99.09	99.09	94.87	94.8	95.27	95.69	94.28	90.06
	Lupac 07	96.8	96.8	96.76	96.76	96.84	96.81	99.22	99.15	99.22	99.14	-	99.17	94.86	94.75	95.28	95.64	94.31	90.13
	LAH08	96.82	96.44	96.8	96.78	96.8	96.76	99.21	99.21	99.21	99.14	99.21	-	94.85	94.76	95.3	95.64	94.33	90
Single strain groups	MED01	95.2	95.22	95.19	95.2	95.16	95.08	95.11	95.12	95.07	95.14	95.14	95.07	-	94.11	94.6	94.87	93.47	89.69
	NIE79	95.54	95.57	95.53	95.51	95.59	95.52	95.59	95.56	95.63	95.1	95.53	95.54	94.48	-	94.57	94.88	93.99	90
	NIE111	95.94	96	95.99	96	95.75	95.9	95.91	95.93	95.8	95.55	95.91	95.88	94.91	94.88	-	95.72	93.87	89.87
	PSH03	95.13	95.14	95.14	95.09	95.19	95.01	95.08	95.11	95.04	95.9	95.06	95.08	95.19	95.15	95.54	-	94.26	90.16
	PSH25	94.66	94.69	94.69	94.65	94.63	94.6	94.59	94.64	94.6	94.57	94.62	94.63	93.87	94.34	94.13	94.52	-	90.42
	LAH09	90.66	90.61	90.65	90.58	90.69	90.53	90.52	90.56	90.54	90.53	90.55	90.5	90.15	90.47	90.37	90.53	90.93	-

-	ANI
OAT	-

B

		Group I						Group II						Single strain groups					
		DSM 44871^T	Lupac 06	GAR05	GAR06	PSN01	PSN13	MED15	ONO23	ONO86	GUI43^T	Lupac 07	LAH08	MED01	NIE79	NIE111	PSH03	PSH25	LAH09
Group I	DSM 44871^T	-	0.02	0.03	0.02	0.1	0.08	0.08	0.12	0.24	0.25	0.07	0.14	0.34	0.06	0.07	0.14	0.06	0.49
	Lupac 06	92.7	-	0	0.01	0.12	0.1	0.11	0.15	0.27	0.28	0.09	0.1	0.36	0.09	0.1	0.16	0.04	0.46
	GAR05	92.5	93.2	-	0.01	0.12	0.11	0.11	0.15	0.27	0.28	0.1	0.11	0.37	0.09	0.1	0.17	0.03	0.46
	GAR06	92.9	93	93.7	-	0.11	0.1	0.1	0.14	0.26	0.27	0.09	0.1	0.36	0.08	0.09	0.16	0.04	0.47
	PSN01	92.8	93	93	93.1	-	0.02	0.01	0.03	0.15	0.16	0.03	0.02	0.24	0.03	0.02	0.04	0.16	0.58
	PSN13	91.3	81.5	81.2	81	81	-	0	0.04	0.16	0.17	0.01	0	0.26	0.02	0.01	0.06	0.14	0.57
Group II	MED15	71.1	71.2	71.4	71.1	71.4	71.5	-	0.04	0.16	0.17	0.02	0	0.26	0.02	0.01	0.06	0.14	0.57
	ONO23	71.3	71.3	71.3	71.3	71.4	71.6	93.1	-	0.12	0.13	0.05	0.04	0.22	0.06	0.05	0.02	0.18	0.61
	ONO86	71.5	71.6	71.5	71.3	71.2	71.8	93.1	93.3	-	0.01	0.17	0.16	0.1	0.18	0.17	0.1	0.3	0.73
	GUI43^T	71.2	71.2	71.5	71.3	71.6	71.3	92.6	92.3	92.8	-	0.18	0.17	0.11	0.17	0.16	0.09	0.29	0.72
	Lupac 07	71.3	71.3	71	71.2	71.4	71.4	93.2	92.5	93.5	92.7	-	0.01	0.27	0.01	0	0.07	0.13	0.55
	LAH08	71.3	71.3	71.5	71.2	71.2	71.5	93.3	93.8	93.1	92.5	93.1	-	0.26	0.02	0.01	0.06	0.14	0.57
Single strain groups	MED01	60.6	60.7	60.8	60.5	61	60.4	60.4	60.4	60.4	60.3	60.3	60.4	-	0.28	0.27	0.2	0.4	0.83
	NIE79	60.4	60.4	60.6	60.4	60.9	60.3	60.2	60.2	60.3	60.3	60.1	60.2	56.6	-	0.01	0.08	0.12	0.55
	NIE111	62.7	62.8	62.8	62.8	63.2	62.6	63.1	63.3	63.4	63.1	63.2	63.2	58.7	59.2	-	0.07	0.13	0.56
	PSH03	65.4	65.5	65.7	65.5	65.8	65.1	65.6	65.6	65.8	65.8	65.5	65.5	60.7	61	66.1	-	0.2	0.63
	PSH25	57.9	58	58.1	58	58.2	57.7	57.4	57.3	57.5	57.4	57.5	57.6	54	56.4	55.4	57.8	-	0.42
	LAH09	41	41	41.1	41	41.4	40.9	40.8	40.7	41	40.8	40.8	40.8	39.6	40.6	40.1	41	42.1	-

-	G+C
dDDH	-

APPENDIX VI

BIOLOG phenotypic profiles of *M. saelicesensis* and *M. noduli* strains. +, positive; –, negative; c, conflicting.

	***M. saelicesensis* (Group I)**						***M. noduli* (Group II)**					
	Lupac 09^{T}	**Lupac 06**	**GAR05**	**GAR06**	**PSN01**	**PSN13**	**GUI43^{T}**	**MED15**	**ONO23**	**ONO86**	**Lupac 07**	**LAH08**
Dextrin	+	+	+	+	+	+	+	c	+	+	+	+
D-Maltose	+	+	+	+	+	+	+	+	+	+	+	+
D-Trehalose	c	-	+	+	-	+	+	+	c	+	+	c
D-Cellobiose	+	+	+	+	+	+	+	+	+	+	+	+
β-Gentiobiose	+	+	+	+	+	+	+	+	+	+	+	+
Sucrose	+	+	+	+	+	+	+	+	+	+	+	+
Turanose	+	+	+	+	+	+	+	+	+	+	+	+
Stachyose	+	c	-	c	c	-	-	-	+	c	c	+
pH 6	+	+	+	+	+	+	+	+	+	c	+	+
pH 5	-	-	-	-	-	-	-	-	-	-	-	-
D-Raffinose	+	c	+	+	+	+	+	+	+	+	+	+
α-D-Lactose	-	c	c	-	-	-	-	+	+	+	-	+
D-Melibiose	+	+	+	+	+	+	+	+	+	+	+	+
β-Methyl-D-Glucoside	+	c	+	+	+	+	+	+	+	+	+	+
D-Salicin	c	c	+	c	+	+	c	+	+	+	+	+
N-Acetyl-D-Glucosamine	c	+	c	-	c	-	-	+	+	c	-	+
N-Acetyl-β-D-Mannosamine	c	c	-	-	-	-	+	c	+	+	c	+
N-Acetyl-D-Galactosamine	+	c	c	c	c	-	+	+	+	c	-	+
N-Acetyl-Neuraminic Acid	-	-	c	-	-	+	-	-	-	-	-	-
1% NaCl	+	+	c	+	+	c	+	+	+	c	+	+
4% NaCl	-	-	-	-	-	-	-	-	-	-	-	-
8% NaCl	-	-	c	-	-	-	-	-	-	-	-	-
D-Glucose	+	+	+	+	+	+	+	+	+	+	+	+
D-Mannose	+	+	+	+	+	+	+	+	+	+	+	+
D-Fructose	+	+	+	+	+	+	+	+	+	+	+	+
D-Galactose	+	+	+	+	+	+	+	+	+	+	+	+
3-O-Methyl-D-Glucose	c	+	-	+	+	c	c	+	+	+	-	+
D-Fucose	-	-	-	-	-	-	-	c	c	-	-	c
L-Fucose	+	+	+	+	+	+	+	+	+	+	+	+
L-Rhamnose	c	c	-	-	-	+	+	+	+	-	+	+
Inosine	+	c	c	c	+	c	-	c	-	+	c	-
1% Sodium Lactate	+	+	+	+	c	+	+	+	+	-	+	+
Fusidic Acid	-	-	-	-	-	-	-	-	-	-	-	-
D-Serine #2	-	-	-	-	-	-	-	-	-	-	-	-
D-Sorbitol	c	c	-	+	c	+	+	c	+	c	-	+
D-Mannitol	-	c	-	-	-	+	-	-	-	-	-	c

D-Arabitol	-	c	-	-	-	-	-	-	-	-	-	-
myo-Inositol	c	c	-	-	-	-	-	-	-	c	-	c
Glycerol	-	c	c	+	c	-	-	c	-	-	-	-
D-Glucose-6-Phosphate	+	c	+	+	+	+	+	+	+	+	+	+
D-Fructose-6-Phosphate	+	c	+	+	+	+	+	+	c	+	+	+
D-Aspartic Acid	-	-	-	-	+	-	-	-	-	-	-	-
D-Serine #1	-	-	-	-	-	-	-	-	-	-	-	-
Troleandomycin	-	-	-	-	-	-	-	-	-	-	-	-
Rifamycin SV	+	+	+	+	+	+	+	+	+	+	+	+
Minocycline	-	-	-	-	-	-	-	-	-	-	-	-
Gelatin	+	+	+	c	+	+	+	+	+	+	+	+
Gly-Pro	c	c	-	+	c	c	-	-	c	-	c	-
L-Alanine	+	c	-	c	-	+	+	-	+	+	+	c
L-Arginine	-	c	+	-	+	+	-	+	-	+	+	-
L-Aspartic Acid	+	+	+	+	+	+	+	+	-	+	+	+
L-Glutamic Acid	+	+	+	+	+	+	+	+	+	+	+	+
L-Histidine	-	c	-	-	-	-	-	-	-	-	-	-
L-Pyroglutamic Acid	c	c	c	-	-	-	-	-	-	c	-	c
L-Serine	-	-	-	-	c	-	-	-	-	-	c	-
Lincomycin	-	+	-	+	-	-	c	-	+	-	-	c
Guanidine Hydrochloride	-	-	-	-	-	-	-	-	-	-	-	-
Niaproof	-	-	-	-	-	-	-	-	-	-	-	-
Pectin	c	-	+	+	+	+	+	-	-	+	+	-
D-Galacturonic Acid	-	-	+	+	+	c	+	c	-	+	c	c
L-Galactonic Acid-γ-Lactone	-	-	-	-	-	-	-	c	-	-	-	-
D-Gluconic Acid	c	-	c	+	-	+	c	+	c	+	+	c
D-Glucuronic Acid	-	-	c	c	c	+	-	c	-	c	c	-
Glucuronamide	-	-	-	-	-	-	-	-	-	-	-	-
Mucic Acid	-	c	c	c	-	-	-	-	-	-	c	-
Quinic Acid	-	-	-	-	-	-	c	c	-	-	c	-
D-Saccharic Acid	c	c	-	-	-	-	-	-	-	c	c	-
Vancomycin	-	-	-	-	-	-	-	-	-	-	-	-
Tetrazolium Violet	-	-	-	-	-	-	-	-	-	-	-	-
Tetrazolium Blue	-	-	-	-	-	-	-	-	-	-	-	-
p-Hydroxy-Phenylacetic Acid	-	-	-	-	-	-	-	-	-	-	-	-
Methyl Pyruvate	+	+	+	c	+	+	+	c	+	+	+	+
D-Lactic Acid Methyl Ester	c	c	c	+	+	-	c	+	c	+	+	+
L-Lactic Acid	-	c	c	c	+	c	-	-	-	-	-	c
Citric Acid	-	c	-	-	-	-	-	-	-	-	-	-

α-Keto-Glutaric Acid	-	c	+	c	+	c	-	+	-	-	c	-
D-Malic Acid	-	c	-	c	-	-	-	+	c	-	-	-
L-Malic Acid	+	+	+	+	+	+	+	+	+	+	+	+
Bromo-Succinic Acid	+	+	+	+	+	+	-	-	-	+	+	-
Nalidixic Acid	+	+	+	+	+	+	+	+	+	+	c	+
Lithium Chloride	+	-	-	-	-	-	+	-	+	-	-	-
Potassium Tellurite	c	-	-	c	-	-	c	-	+	-	-	+
Tween 40	+	+	+	+	+	+	+	+	+	c	+	+
γ-Amino-n-Butyric Acid	-	c	-	-	-	-	-	-	-	-	-	-
α-Hydroxy-Butyric Acid	-	c	c	-	-	-	-	-	-	-	-	-
β-Hydroxy-Butyric Acid	+	c	+	+	+	+	+	+	-	-	+	+
α-Keto-Butyric Acid	-	c	c	-	-	-	-	-	-	-	-	-
Acetoacetic Acid	+	-	+	+	+	+	+	+	+	+	+	+
Propionic Acid	+	+	+	+	+	+	+	-	c	+	+	-
Acetic Acid	+	+	+	+	+	+	+	+	+	+	+	-
Sodium Formate	c	c	-	-	-	-	-	-	-	c	-	-
Aztreonam	+	+	+	+	+	+	+	-	+	+	+	c
Butyric Acid	-	-	-	-	-	-	-	-	-	-	c	-
Sodium Bromate	c	-	-	-	-	-	c	-	+	-	-	-

+	**Positive**
-	**Negative**
c	**Conflictive results between the two replica**

APENDICE VII

Carbon source substrates tested at different times using the same laboratory conditions. +, positive; –, negative; w, weak.

	M. saelicesensis (Group I)											*M noduli* (Group II)										
	Lupac 06			GAR05		GAR06		PSN01		PSN13		MED15		ONO23		ONO86		Lupac 07			LAH08	
	2007	2016	2017	2016	2017	2016	2017	2016	2017	2016	2017	2016	2017	2016	2017	2016	2017	2007	2016	2017	2016	2017
D-Maltose	+	+	+	+	+	+	+	+	+	+	+	+	+	w	+	+	+	+	+	+	+	+
D-Trehalose	-	+	+	+	+	+	+	+	+	+	+	+	+	w	+	+	+	-	+	+	+	w
D-Cellobiose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sucrose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Raffinose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Mannose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Galactose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
L-Rhamnose	-	-	-	-	-	-	-	-	w	+	w	-	+	+	+	+	+	+	+	+	+	+
D-Serine	-	w	w	w	-	-	-	-	w	-	w	-	-	w	w	w	w	-	w	-	-	w
L-Alanine	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
L-Arginine	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
L-Histidine	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
L-Arabinose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Lysine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
L-Proline	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sorbose	-	-	-	-	w	-	w	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Valine	-	-	-	-	-	w	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
xylitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
myo-Inositol	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-

+	Positive
-	Negative
w	Weak

Descifrando genomas: Genómica comparativa de cepas de *Micromonospora* asociadas a leguminosas

Resumen en español

Resumen en español de la memoria presentada por **Raúl Riesco Jarrín**
para optar al Grado de Doctor por la Universidad de Salamanca.
Directora: Martha E. Trujillo Toledo

Universidad de Salamanca
Departamento de Microbiología y Genética

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS
VNiVERSiDAD D SALAMANCA
1218 ~ 2018

Índice de la tesis

Resumen

Micromonospora es un género de bacterias Gram positivas con una amplia distribución en diversos ecosistemas (Genilloud, 2015). Este género es conocido por su gran capacidad para producir metabolitos secundarios, lo que lo convierte en un recurso biotecnológico valioso en medicina y agricultura (Carro et al., 2018; Martínez-Hidalgo et al., 2015; Trujillo et al., 2011). En los últimos años, se ha observado que *Micromonospora* forma una estrecha relación con tejidos vegetales. En particular, *Micromonospora* ha podido ser aislada de forma sistemática como endófitas en nódulos de fijación de nitrógeno en leguminosas y plantas actinorrízcas (Trujillo et al., 2015).

Para intentar clarificar la diversidad del género *Micromonospora* en leguminosas, se han realizado múltiples estudios moleculares (BOX-PCR, ARDRA, RFLP, RAPDS) (Carro, 2009; Cerda, 2008; de la Vega, 2010; Trujillo et al., 2010). Estos estudios han revelado una alta diversidad genética en aislados de *Micromonospora*, dando lugar a la descripción de un número importante de nuevas especies (Carro et al., 2012, 2013; Garcia et al., 2010; Trujillo et al., 2007). Entre ellas, la más frecuentemente aislada es la especie *Micromonospora saelicesensis*, seguida de cerca por *Micromonospora noduli*, una especie muy cercana filogenéticamente con la que comparte múltiples rasgos genéticos y fisiológicos.

La función ecológica de *Micromonospora* en relación con su planta hospedadora es aún desconocida. Sin embargo, se ha podido confirmar la presencia de estas bacterias en el interior del tejido vegetal (Trujillo et al., 2010), e incluso monitorizar el proceso de colonización del nódulo (Benito et al., 2017). *Micromonospora* parece interactuar de forma estrecha no solamente con la leguminosa, sino también con bacterias fijadoras de nitrógeno (rhizobia), estableciendo una relación tripartita (Benito et al., 2017).

En estudios previos, se ha secuenciado y anotado el genoma de la cepa *Micromonospora lupini*, Lupac 08 aislada de nódulos de la planta *Lupinus angustifolius* (Alonso-Vega et al., 2012; Trujillo et al., 2014). Un primer análisis del genoma de esta bacteria sugirió que este microorganismo posee un alto número de genes codificantes para la producción de enzimas hidrolíticas. Se han identificado alrededor de 865 genes (9.7% del genoma) relacionados con el metabolismo de carbohidratos, de los cuales, aproximadamente 192 están relacionados con la producción de este tipo de enzimas, principalmente celulasas.

La actividad de estas enzimas hidrolíticas ha sido comprobada mediante ensayos *in-vitro*, demostrando gran actividad. No obstante, la presencia de estas enzimas no parece perjudicar la simbiosis con la planta. El genoma de *M. lupini* Lupac 08 también contiene varios genes codificantes para funciones comúnmente relacionadas con las bacterias promotoras de crecimiento vegetal, como la producción de sideróforos, antibióticos y fitohormonas (Trujillo et al., 2014).

Hasta 2015, se habían secuenciado y descrito solamente dos cepas de *Micromonospora* relacionadas con nódulos fijadores de nitrógeno, lo que supone un porcentaje ínfimo de la colección de más de 2000 cepas del género aisladas en nuestro laboratorio. Por lo tanto, gran parte del potencial ecológico, metabólico y genético de esta bacteria quedaba aún sin estudiar. Esta tesis busca ahondar en la relación beneficiosa de *Micromonospora* con la planta, haciendo un estudio de genómica comparativa a nivel de género para intentar dilucidar los sistemas biológicos que condicionan esta relación planta-microorganismo. Para ello, se han propuesto los siguientes objetivos específicos:

1. Creación de nuevas herramientas bioinformáticas que complementen o mejoren las capacidades de programas frecuentemente usados por la comunidad científica en la caracterización genómica bacteriana. Creación de un nuevo proceso bioinformático para la caracterización de funciones que intervienen en la relación planta-microorganismo en *Micromonospora.*
2. Caracterización mediante genómica comparativa de los principales rasgos genómicos que intervienen en la relación *Micromonospora*-planta.
3. Caracterización genómica de cepas aisladas de tejidos vegetales, cercanas a *Micromonospora saelicesensis* y *Micromonospora noduli*. las principales especies de *Micromonospora* encontradas en nódulos fijadores de nitrógeno de leguminosas.

La presente tesis se ha dividido por tanto en tres partes, una por cada objetivo propuesto. En el primer capítulo, algunos de los programas y aplicaciones más usados en la caracterización genómica de microorganismos fueron brevemente analizados, remarcando sus puntos fuertes y débiles. Se han propuesto dos soluciones bioinformáticas diseñadas para mejorar la experiencia del usuario final en el uso de estos programas de caracterización genómica: UBCG_iTOL_maker y GGDC Output Management Assistant

(GOMA). Por último, se ha propuesto un proceso basado en R, *Micromonospora* Plant Associated Gene tool (MicroPLAGE), cuyo objetivo es separar diferencias genómicas significativas que potencialmente intervienen en la relación entre *Micromonospora* y su planta huésped. Todas estas soluciones bioinformáticas han sido implementadas como scripts basados en R, que pueden ser fácilmente introducidos en procesos bioinformáticos en el futuro, cambiando unas pocas variables.

El segundo capítulo se centró en el segundo objetivo: la caracterización de rasgos genómicos que intervienen en la relación planta-microorganismo en *Micromonospora*. Para lograr este objetivo se secuenciaron los genomas de diecisiete cepas de *Micromonospora*, asiladas de diferentes leguminosas (*Cicer* sp., *Medicago* sp., *Lupinus* sp., *Ononis* sp., *Pisum* sp. and *Trifolium* sp.) y tejidos vegetales (nódulos y hojas). Gracias a la inclusión de estos genomas, hemos construido una base de datos de setenta y cuatro genomas, con un número similar de genomas asociados a cepas aisladas de suelo y a cepas endófitas. Mediante el uso de genómica comparativa, y basándonos en la base de datos generada en 2018 (Levy et al., 2018) y en el proteoma de plantas huésped de *Micromonospora*, se han podido determinar varios rasgos genómicos que potencialmente pueden estar relacionados con la interacción *Micromonospora*-planta.

El tercer capítulo se centró en el último objetivo. La especie *M. saelicesensis* ha sido aislada repetidamente de nódulos de leguminas, siendo la especie de *Micromonospora* más frecuentemente aislada en este tejido vegetal (Carro, 2009; Carro et al., 2012; Cerda, 2008; de la Vega, 2010; Trujillo et al., 2010, 2015). En 2016, la especie *M. noduli* fue descrita como una especie muy próxima a *M. saelicesensis* (Carro et al., 2016). Ambas especies compartían muchos rasgos, tanto genómicos como fisiológicos, y surgió la duda de si deberían ser consideradas como una sola especie. Usando diferentes aproximaciones genómicas, en el capítulo III hemos estudiado la relación taxonómica entre las especies *Micromonospora saelicesensis* y *Micromonospora noduli* para concluir que efectivamente son dos especies distintas pero muy cercanas entre sí.

Conclusiones

Capítulo I:

Se han propuesto tres procesos bioinformáticos en esta tesis. Estos protocolos están basados en scripts de R, que intentan ser lo suficientemente flexibles como para implementarse en cualquier otro proceso bioinformático, con independencia del sistema operativo empleado.

Micromonospora Plant Associated Gene tool (MicroPLAGE) ha sido concebido como un método innovador para la búsqueda de características genómicas relacionadas con la interacción planta-microrganismo a nivel de género. Este script funciona bajo la premisa de una adaptación genómica al ambiente vegetal por parte de la bacteria que nos permita extraer diferencias significativas entre los distintos genomas. *Micromonospora* es un género bacteriano que se puede encontrar en varios ambientes, y por lo tanto resulta el candidato perfecto para el uso de este script. En la segunda parte de esta tesis, MicroPLAGE será la herramienta central en el análisis de genómica comparativa.

UBCG_iTOL_maker y GGDC Output Management Assistant (GOMA) han sido creados para optimizar la entrada de datos y la presentación final de los resultados de los programas UBCG y GGDC, de forma intuitiva y visual. Puesto que estos scripts dependen de los programas base, en muchas ocasiones no pueden evitar heredar algunos de sus puntos débiles. Como ejemplo, el script GOMA toma como datos de entrada los resultados del servicio online GGDC, y por lo tanto requiere que el usuario suba manualmente todos los genomas al servidor para hacer todas las interacciones. Puesto que en la actualidad el servidor GGDC no tiene manera de automatizar estas subidas de datos, no se puede hacer nada para evitar este punto débil. Sin embargo, la capacidad de estos scripts para ser incorporados en procesos bioinformáticos de forma sencilla conlleva que todos estos problemas pueden ser solucionados si el programa GGDC es liberado para su uso de forma local. UBCG_iTOL_maker y GOMA serán usados en la tercera parte de esta tesis para mejorar la presentación y visualización de los resultados.

Capítulo II

- No se encontró una correlación significativa entre la distribución de los hábitats de aislamiento y la longitud de los genomas en nuestra base de datos. Sin embargo, sí que se ha encontrado una diferencia significativa en el tamaño genómico entre los miembros del género *Salinispora* y los miembros del género *Micromonospora*, siendo estos últimos significativamente más grandes.
- Se ha obtenido una base de datos final de 69046 genes con potencial relación planta-microorganismo. Las cepas de origen endofítico contribuyeron con una media de 1036 ± 57 genes potencialmente relacionados con la planta a la base de datos.
- El nuevo proceso bioinformático ha revelado varias cepas con potencial relación cercana con la planta. Se han de llevar a cabo estudios posteriores para probar si estas bacterias pueden colonizar la planta.
- Las cepas de *Micromonospora* relacionadas con planta tienen reforzados en sus genomas aquellos sistemas relacionados con el metabolismo y transporte de fuentes de carbono.
- En este estudio se ha generado una lista de las principales diferencias funcionales entre las cepas de *Micromonospora* potencialmente relacionadas con la planta y el resto de las cepas inlcuidas en la base de datos. Estas diferencias se debieron principalmente al refuerzo de sistemas funcionales, y no a la presencia o ausencia de genes específicos.

Capítulo III

- Las clasificaciones filogenéticas basadas en datos genómicos serán más estables a medida que se vayan incorporando nuevos datos, lo que supone la creación de un nuevo marco de trabajo fiable en la sistemática de procariotas.
- Lo trabajos basados en varias cepas de estudio y el empleo de distintos métodos de análisis son fundamentales para la correcta caracterización de la variabilidad intra-especie.

- Los valores OGRI, han demostrado ser muy apropiados para la correcta separación de especies bacterianas, especialmente la hibridación digital (dDDH). Sin embargo, los límites numéricos de referencia han de aplicarse con cierta flexibilidad, considerando también aspectos inherentes al microorganismo como la ecología, fisiología, etc.
- La información fenotípica es útil para complementar la caracterización de las cepas de estudio. Sin embargo, estos estudios deberían estar enfocados en proporcionar información sobre la biología del microorganismo, y no solo en generar una tabla de resultados de valor cuestionable.

Referencias

Alonso-Vega, P., Normand, P., Bacigalupe, R., Pujic, P., Lajus, A., Vallenet, D., et al. (2012). Genome Sequence of *Micromonospora lupini* Lupac 08, Isolated from Root Nodules of *Lupinus angustifolius*. *J. Bacteriol.* 194, 4135–4135. doi:10.1128/JB.00628-12.

Benito, P., Alonso-Vega, P., Aguado, C., Luján, R., Anzai, Y., Hirsch, A. M., et al. (2017). Monitoring the colonization and infection of legume nodules by *Micromonospora* in co-inoculation experiments with rhizobia. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-11428-1.

Carro, L. (2009). Avances en la Sistemática del Género *Micromonospora*: Estudio de Cepas aisladas de la Rizosfera y Nódulos de *Pisum sativum*.

Carro, L., Nouioui, I., Sangal, V., Meier-Kolthoff, J. P., Trujillo, M. E., Montero-Calasanz, M. del C., et al. (2018). Genome-based classification of *micromonosporae* with a focus on their biotechnological and ecological potential. *Sci. Rep.* 8, 525. doi:10.1038/s41598-017-17392-0.

Carro, L., Pukall, R., Spröer, C., Kroppenstedt, R. M., and Trujillo, M. E. (2013). *Micromonospora* halotolerans sp. nov., isolated from the rhizosphere of a *Pisum sativum* plant. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 103, 1245–1254. doi:10.1007/s10482-013-9903-7.

Carro, L., Riesco, R., Spröer, C., and Trujillo, M. E. (2016). Micromonospora ureilytica sp. nov., Micromonospora noduli sp. nov. and *Micromonospora* vinacea sp. nov., isolated from *Pisum sativum* nodules. *Int. J. Syst. Evol. Microbiol.* 66, 3509–3514. Available at: https://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.001231.

Carro, L., Spröer, C., Alonso, P., and Trujillo, M. E. (2012). Diversity of Micromonospora strains isolated from nitrogen fixing nodules and rhizosphere of *Pisum sativum* analyzed by multilocus sequence analysis. *Syst. Appl. Microbiol.* 35, 73–80. doi:10.1016/j.syapm.2011.11.003.

Cerda, M. E. (2008). Aislamiento de *Micromonospora* de Nódulos de Leguminosas

Tropicales y Análisis de Su Interés Como Promotor del Crecimiento Vegetal.

de la Vega, P. A. (2010). Distribución, caracterización e importancia ecologica de *Micromonospora* en nódulos fijadores de nitrogeno de *Lupinus*.

Garcia, L. C., Martinez-Molina, E., and Trujillo, M. E. (2010). *Micromonospora pisi* sp. nov., isolated from root nodules of *Pisum sativum*. *Int. J. Syst. Evol. Microbiol.* 60, 331–337. doi:10.1099/ijs.0.012708-0.

Genilloud, O. (2015). *Micromonospora*. *Bergey's Man. Syst. Archaea Bact.* doi:doi:10.1002/9781118960608.gbm00148.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018). Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50, 138–150. doi:10.1038/s41588-017-0012-9.

Martínez-Hidalgo, P., Galindo-Villardón, P., Trujillo, M. E., Igual, J. M., and Martínez-Molina, E. (2015). *Micromonospora* from nitrogen fixing nodules of alfalfa (*Medicago sativa* L.). A new promising Plant Probiotic Bacteria. *Sci. Rep.* 4, 6389. doi:10.1038/srep06389.

Trujillo, M. E., Alonso-Vega, P., Rodríguez, R., Carro, L., Cerda, E., Alonso, P., et al. (2010). The genus *Micromonospora* is widespread in legume root nodules: The example of *Lupinus angustifolius*. *ISME J.* 4, 1265–1281. doi:10.1038/ismej.2010.55.

Trujillo, M. E., Bacigalupe, R., Pujic, P., Igarashi, Y., Benito, P., Riesco, R., et al. (2014). Genome Features of the Endophytic Actinobacterium *Micromonospora lupini* Strain Lupac 08: On the Process of Adaptation to an Endophytic Life Style? *PLoS One* 9, e108522. doi:10.1371/journal.pone.0108522.

Trujillo, M. E., Igarashi, Y., Saiki, I., Yanase, S., Miyanaga, S., Enomoto, M., et al. (2011). Lupinacidin C, an Inhibitor of Tumor Cell Invasion from *Micromonospora lupini* . *J. Nat. Prod.* 74, 862–865. doi:10.1021/np100779t.

Trujillo, M. E., Kroppenstedt, R. M., Fernández-Molinero, C., Schumann, P., and Martínez-Molina, E. (2007). *Micromonospora lupini* sp. nov. and *Micromonospora saelicesensis* sp. nov., isolated from root nodules of *Lupinus angustifolius*. *Int. J. Syst.*

Evol. Microbiol. 57, 2799–2804. doi:10.1099/ijs.0.65192-0.

Trujillo, M. E., Riesco, R., Benito, P., and Carro, L. (2015). Endophytic actinobacteria and the interaction of *Micromonospora* and nitrogen fixing plants. *Front. Microbiol.* 6, 1–15. doi:10.3389/fmicb.2015.01341.