Screening Soil Metagenomic Libraries Searching for Degrading Enzymes Useful for the Production of Biofuels and Value Added Chemicals

by

Nolberto Figueroa Matías

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER IN SCIENCE

In

CHEMICAL ENGINEERING

UNIVERSITY OF PUERTO RICO

MAYAGÜEZ CAMPUS

2012

Approved by:

_____	_____
Patricia Ortiz Bermúdez, PhD President, Graduate Committee	Date
_____	_____
Carlos Ríos Velázquez, PhD Member, Graduate Committee	Date
_____	_____
Magda Latorre Esteves, PhD Member, Graduate Committee	Date
_____	_____
Dr. Carlos Quiñones Padovani Representative of Graduate Studies	Date
_____	_____
Aldo Acevedo Rullán, PhD Chairperson of the Department	Date

Abstract

Value-added chemicals and fuels produced via the biotransformation of lignocellulosic biomass resources like crop biomass waste are necessary to reduce our dependence on the non-renewable conversion of raw materials. Moreover, these alternative energy and chemical sources are crucial in order to achieve sustainability, and decrease the adverse impact of these activities on the environment. For instance, an alternative, renewable source of energy is the bio-fuel ethanol. Even though sugar ethanol production has been demonstrated to work in countries like Brazil, achieving a cost-effective, commercial-scale cellulosic bio-fuel industry requires new technologies to lower the price of bio-transforming these materials into bio-fuels. In addition, efforts to develop these technologies should focus on raw materials that are not committed to other primary needs, such as food crops. Therefore, there is a high interest in the use of low-value materials, such as waste crop biomass, switch grass, wood chips and other sources of lignocellulose and hemicelluloses to convert them into useful products such as bio-fuels. Challenges with the utilization on certain raw materials, like coffee crop waste, reside on the presence of toxic compounds like caffeine that represent a limiting step in the extraction of their energy potential confined in their lignocellulose. Biotechnological research is key in accelerating the discovery of new genes with the “information” that enables microorganisms to produce yet unknown catalytically active enzymes that can modify these raw materials when using them as carbon source and transforming them into high-value products. Out of the total microbial diversity that exists in different ecosystems in our planet, only 1% of those microorganisms are able to grow under known laboratory techniques. This means that 99% of those microorganisms with possible new genes encoding for different abilities remain undiscovered. This is where the importance of metagenomics relies. After screening a group of Metagenomic Libraries constructed from samples of Rainy and Dry Forest soil of Puerto Rico and from Hyper-saline Microbial Mats from Puerto Rico searching for genes that encodes for activities that enable a host to degrade lignin, cellulose, tween and caffeine as carbon source; we found a list of interesting genes. Some of them are related to aminoacid biosynthesis and degradation while others have more specific activities. These genes can be used for the production of biofuels or value added chemicals. We found a putative glycosidase with possible activity over saponins and possible lipase activity, we found genes related to the diaminopimelate pathway that possibly are involved in aminoacid metabolism, a LysR Hdfr gene that possibly relates to aminoacid metabolism, flagellum development and pigment production, and a putative Ketol acid reductoisomerase with possible activity over 2-acetolactate, related to aminoacid metabolism. These results prove that the metagenomic approach is useful to isolate

genes from different environments and from cultivable and uncultivable microorganism. Finally, the data support the use of metagenomics to find activities that can be applied to processes for the generation and production of value added chemicals and biofuels.

Keywords: Value added chemicals, metagenomic, carbon sources

Resumen

Químicos con valor añadido y combustibles producidos por medio de la transformación de fuentes de biomasa con contenido lignoceluloso como lo son desperdicios de cosechas son necesarios para reducir nuestra dependencia de la conversión no renovable de materia prima. Más aun, estas fuentes químicas y energéticas alternas son cruciales para poder alcanzar sostenibilidad, y disminuir los impactos adversos de estas actividades sobre el medio ambiente. Por ejemplo, una fuente de energía alterna y renovable es el biocombustible etanol. A pesar que se ha demostrado que funciona la producción de etanol basada en azúcar en países como Brasil, alcanzar la comercialización de una industria costo efectiva para la producción de combustibles basados en celulosa; requiere nuevas tecnologías para bajar el costo de biotransformar estos materiales en bio combustibles. Adicionalmente, esfuerzos dirigidos para desarrollar estas tecnologías deben ser enfocados en utilizar materia prima que no esté comprometida con otras necesidades primarias, como lo están las cosechas de alimentos. Por consiguiente, hay un alto interés en el uso de materiales de bajo valor, como lo son la biomasa de desperdicios de cosechas, hierbas, astillas de madera y otras fuentes de lignina, celulosa y hemicelulosa para convertirlas en productos útiles como bio combustibles. Dificultades con la utilización de ciertas materias primas, como los desechos de la cosecha del café, residen en la presencia de compuestos tóxicos como la cafeína que representan un paso limitante en la extracción de su potencial energético confinado en sus lignocelulosas. La investigación biotecnológica es clave en acelerar el descubrimiento de genes nuevos con la “información” que le permite al microorganismo el producir enzimas catalíticamente activas aun desconocidas que puedan modificar estas materias primas cuando se las usa como fuentes de carbono y así poder transformarlas en productos de alto valor. Del total de la diversidad microbiana que existe en diferentes ecosistemas de nuestro planeta, solo el 1% de esos microorganismos son capaces de crecer bajo técnicas de laboratorio conocidas. Esto significa que el 99% de esos microorganismos con posiblemente nuevos genes que codifican para diferentes habilidades permanecen sin ser descubiertos. Es allí donde radica la importancia de la metagenómica. Después de investigar usando un grupo de Bibliotecas Metagenómicas construidas de muestras de suelo provenientes de bosques húmedos y secos de Puerto Rico y de tapetes microbiales híper-salinos de Puerto Rico buscando genes que codificaran para actividades que le permitieran a la célula huésped degradar lignina, celulosa, tween y cafeína como fuente de carbono; encontramos una lista de genes interesantes. Algunos de ellos están relacionados a biosynthesis y degradación de aminoácidos mientras otros tienen actividades más específicas. Estos genes pueden ser usados para la producción de biocombustibles o compuestos de valor

adquirido. Encontramos una posible glycosidase con posible actividad sobre saponinas y posible actividad de lipasa, genes relacionados a la ruta de diaminopimelato que posiblemente estén envueltos en metabolismo de aminoácidos, un gen conocido como LysR Hdfr que posiblemente se relacione a metabolismo de aminoácidos, desarrollo de flagelo y producción de pigmento y una Ketol acid reductoisomerasa con posible actividad sobre 2-acetolactato, relacionado a metabolismo de aminoácidos. Estos resultados prueban que el uso de metagenómica es útil para aislar genes de medios ambientes diferentes y de microorganismos cultivables y no cultivables. Finalmente, los datos apoyan el uso metagenómica para encontrar actividades que puedan ser aplicadas a procesos para la generación y producción de compuestos de valor adquirido y biocombustibles.

Palabras clave: Químicos de valor añadido, metagenómica, fuente de carbono

Dedication

I dedicate this work to God the Fahter, to Jesús the Son and to the Holy Spirit; the Biblical Diety:

1.) “Blessed be the name of God forever and ever: for wisdom and might are His: and He changeth the times and the seasons: He removeth kings, and setteth up kings: *He giveth wisdom unto the wise, and knowledge to them that know understanding”.* (Daniel 2:20-21)

2.) “He giveth *power to the faint*; and to them that *have no might* he increaseth strength”. (Isaiah 40:29)

3.) “If any of you *lack wisdom*, let him *ask of God*, that giveth to all men liberally, and upbraideth not; and it shall be given him. But let him *ask in faith*, nothing wavering. For he that wavereth is like a wave of the sea driven with the wind and tossed. For let not that man think that he shall receive anything of the Lord”. (James 1:5-7)

4.) “Whether therefore ye eat, or drink, or whatsoever ye do, *do all to the glory of God”*. (1 Corinthians 10:31)

Without the Strength and the Knowledge that comes from you, I couldn’t have done this my Lord; gratefully I thank you and I write these biblical texts to give you all the glory that you deserve; thank you.

Aknowledgements

I want to give tanks to my advisor Dra. Patricia Ortiz Bermúdez for her mentoring, guidance and supervision and for the opportunity of doing research with her; I want to give tanks Dr. Carlos Ríos Velázquez for his support, counseling and friendship; I want to give thanks to Magda Latorre for her kindness in accepting being part of the thesis comitee; and to the Department of Chemical Engineering for the opportunity to be part of the program.

Also I want to give tanks to some additional individuals:

I want to give tanks to my mother Elizabeth Matías, to my father Norberto Figueroa, to my sister Melany Figueroa and to my wife Anaida Ramos Lopez who supported me economically and emotionally and helped me “stay the course” when I was drained and tired and wanted to get out.

I want to give thanks to the member of the laboratory that helped me allot in preparing experiments and discussing and interpreting the results: María del Pilar Sierra Gómez, Camilo Mora, Irimar Torres and Mónica Medina.

I want to give special thanks to Keila Flores because even when she was not present at any time for this master degree, it was from her that I learned most of the techniques I used in the experiments that conducted me to achieve this goal, so I really wanted to give her thanks.

I wanted to share my success with all of you with this special acknowledgement. To all of you thanks from the bottom of my hearth. My best wishes for you all. I want to bless you all with the blessing that God commanded Moses to bless Israel with:

> “The Lord bless thee, and keep thee: The Lord make his face shine upon thee, and be gracious unto thee: the Lord lift up his countenance upon thee, and give thee peace. (Numbers 6:24-26)”

Glossary

1. Analysis In Silico: experimental results analyzed by computer databases.
2. Biofuel: fuel derived from organic matter by the action of chemical or biological processes.
3. Biomass: the total mass of a collection of organisms or microorganisms. For example the collection of all the lignocelluloses from crop waste or the collection of all the microorganisms in a cultive.
4. Biotechnology: application of biology, chemistry and engineering to produce industrial enzymes; to genetically improve microbes, plants and animals; to discover new enzymes, new bioactive agents like new medicines and antibiotics; from living organisms or their constitutive parts.
5. Biotransformation: chemical alterations of a compound caused by the action of enzymes produced by an organism.
6. Blast: Basic Local Alignment Search Tool. BLAST uses statistical methods to compare a DNA or protein input sequence to a database of sequences and return those sequences that have a significant level of similarity to the query sequence.
7. Blast n (nucleotide collection): compares a nucleotide query sequence against a nucleotide sequence database.
8. Blast x: compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
9. Blast x mategenome (env_nr database): database that compares the sequence submitted to blast with protein sequences isolated from environmental samples.
10. Blast x non redundant: database that compares the sequence submitted to blast with protein sequences found in GenBank CDS translations + PDB + SwissProt + PIR + PRF databases, excluding env_nr database.
11. Cellulose: a complex carbohydrate composed of glucose units found in plant cell walls. Is the most abundant organic compound in our planet.
12. Clones: a group of cells derived from a single ancestor. If a cell harbors a stable self replicable foreign DNA molecule, the derived cells will also contain the foreign DNA insertion.
13. Cosmid: a type of hybrid plasmid that contains a λ phage *cos sequence.*
14. Cultivation: to grow a microorganism or a single cell from other types of tissues (mamal, insect, plants) in vitro in a defined growing media under specific and controlled conditions.

15. e value: a parameter that describes the number of hits one can ‘expect’ to see by chance when searching a database of a particular size. The lower or the closer to zero the e-value is, the more ‘significant’ the match is.
16. Enzyme: proteins that work as biological catalyst to transform a substrate into a product.
17. Fosmid: an f-factor cosmid comonly used in metagenomics, which is capable to clone aproximatelly 45 kb of DNA.
18. Functional Based Metagenomics: branch of the field of Metagenomics that explores the products that microbes in a community can produce by randomly cloning fragments of DNA extracted from an environment under study in laboratory cultivable hosts and screen those clones to identify a function of interest by growing them under known laboratory techniques searching for a specific phenotype that indicate that a new genotype is present in the host cell containing the environmental insert.
19. Genotype: the genetic makeup of an organism or microorganism.
20. Hemicellulose: polysaccharides found in plants that are more complex than simple sugars but not as complex as celluloses.
21. Host cell: a cell that has been introduced with DNA (or RNA). Also can be described as a cell that harbors foreign molecules like foreign DNA or a cell that harbors viruses or microorganisms.
22. Lignin: a compound with recalcitrant nature, a major component of plant biomass, the second most abundant organic compound in the planet and the most abundant aromatic subastance in the world. Lignin gives plants its structural strength and rigidity by linking to both hemicelluloses and celluloses, creating a barrier and inhibiting the penetration of enzymes inside of the matrix where celluloses and hemicelluloses are protected.
23. Lignocellulose: substances that constitute the essential part of plant cell walls consisting of cellulose intimately associated with lignin.
24. Master pool: combination of all clones isolated from a metagenomic library.
25. Metagenomics: emerging discipline that combines disciplines such as engineering, chemistry and biology as tools leading to the extraction and cloning of total DNA from a community of microorganism collectively found in any environment that cannot be cultivated in a laboratory by known cultivating conditions (1). The objective of the field is to understand the characteristics of the microbial world present in the environment and to extract a benefit from that understanding. Kevin Chen et al.,

2005, defines metagenomics as the "application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species" (2).

26. Metagenomic Library: collection of clones each containing a random fragment of DNA isolated from all the microbes present in the sampled community under study.
27. Microbial Mats: laminar organosedimentary structures that are typically organized with colored layers dominated by different microbes, which are influenced by the oxygen, light and sulfur across the mat (3).
28. Microflora: bacterial population present in a specific environment.
29. Metagenome: collectively the genomes of the total microbiota found in nature (4).
30. NCBI Orf Finder (Open Reading Frame Finder): is a graphical analysis tool which finds all open reading frames in a user's sequence or in a sequence already present in the database.
31. Non-renewable source: a natural source that if is totally consumed, cannot be replenish.
32. Nucleotide: the basic building block of nucleic acids, such as DNA and RNA. It is an organic compound made of a nitrogenous base, a sugar, and a phosphate group.
33. Orf (Open reading frame): a section of a sequenced piece of DNA that begins with an initiation (methionine ATG) codon but doesn't have a termination or end codon. ORFs all have the potential to encode a protein or polypeptide, however many may not actually do so.
34. Phenotype: the observable characteristic of an organism or microorganism that indicates the presence of a genotype that enables the organism to have a specific ability; like the ability to degrade a specific substrate.
35. Plasmid: is a circular DNA molecule able of replicating in the cytoplasm of some microbes.
36. Primers: a segment of DNA or RNA that is complementary to a given DNA sequence and that is needed to initiate replication by DNA polymerase.
37. Primer Walking: "A method for sequencing long (>1 kbp) cloned pieces of DNA. The initial sequencing reaction reveals the sequence of the first few hundred nucleotides of the cloned DNA. Using this, a new primer of about 20 nucleotides is synthesized, which is complementary to a sequence near the end of sequenced DNA, and used to sequence the next few hundred nucleotides of the cloned DNA.

This procedure is repeated until the complete nucleotide sequence of the cloned DNA is determined" (5).

38. Query: each individual sequence submitted in Blast for analysis in specific databases or in a group of databases.
39. Query coverage: percent of the query sequence that overlaps the sequence of a gene or protein in the database.
40. Raw material: basic substance in its natural, modified, or processed state, that can be used as building block for subsequient modification to produce a value added compound.
41. Screening: the process of testing clones to find out if they carry inserted genes that confers them an ability that they originally were unable to perform.
42. Selection: a natural or artificial process that favors or induces survival and perpetuation of one kind of organism or microorganism over others that dies or fails to produce offspring.
43. Sequence: order of nucleotide bases in a DNA molecule.
44. Sequencing: laboratory technique used to find out the order of nucleotide bases in a DNA fragment.
45. Sequence Based Metagenomics: branch of the field of Metagenomics that focuses on identifying genes and metabolic pathways by analyzing and comparing the genomic information of an individual microbial species or the complete genome of a comunity obtained by sequencing the DNA present in an environmental sample.
46. Subpools: distribution of the clones presents in a master pool in different samples.
47. Substrate: compound on which an enzyme acts.
48. Transformation: The genetic alteration of a host cell by the inclusion of foreign DNA.
49. Transposon: discrete DNA segments that can move between different, non homologous, genomic loci.
50. Value Added Chemicals: compounds that can serve as raw materials or building blocks for the production of diverse products of economical importance like fabrics, vitamins, fuels, polymers, detergents, pharmaceuticals, disinfectants, antibiotics, cleaners, chelators, and other chemicals of interest.

Table of Contents

Table list:

List of Images and Figures:

1 Introducción

1.1 Motivation

We live in a time where our dependence on fossil fuels must be reduced to stop depleting these non-renewable energy sources. Biofuels are less contaminant, and are a renewable source of energy. Different types of alcohols, biodiesel, and methane are examples of the great variety of chemicals that can be produced by the intervention of microorganism; prokaryotes and eukaryotes alike (6). The bioproduction of fuels and chemicals presents a desirable alternative to the utilization of strictly chemical processes due to the socioeconomic relief associated to it. For instance, the reserves of fossil fuels that are being drained off by human consumption, have caused the increase in the cost of energy production and as a result an adversely effect on our economy is inminent. Because of environmental concerns and because of economical reasons we must increase our knowledge on the microbe's ability to produce chemicals that can help us cease our dependence to fossil fuels to be able to obtain cost effective energy sources that can help sustain our way of living. There are already known microbes that contain enzymes that can be used for the production of fuels and chemicals. However, we need to find efficient enzymes to be able to produce fuels and chemicals, in a cost-effective way to be able to compete with the prices of current technologies, and to avoid a negative economical impact. Puerto Rico has a great diversity of ecosystem with an immense potential for discovery. Here we can search for these enzymes in diverse environments that range from tropical rainy forest soils and tropical dry forest soils; to mangroover soils near beaches and hypersaline microbial mats. The variability in ecosystems increases the diversity in microorganism which augments the possibility of finding new enzymes and possible unknown alternative metabolic routes for the production of value added chemicals and fuels. With that thought in our mind our motivation for this research is to search for genes that codifies for enzymes that can be used to produce biofuels and add knowledge to our tool box for the future.

1.2 Literature Revision

1.2.1 Metagenomic Science

1.2.1.1 An unknown genetic treasure chest buried in the soil.

Terrestrial and aquatic ecosystems contains a vast quantity of microorganisms of which less than 1% can be cultivated using known standard laboratory techniques. Calcagno et al., 2005 estimates that only 0.1 - 1.0% of microorganisms can be cultivated using current techniques; leaving the vastness of microbial lifestyles remaining to be explored (7). Take soil for example; Torsvik et al., 2002 explains that microbial diversity in soil ecosystems exceeds, by far, that of eukaryotic organisms. He continues by citing that one gram of soil may harbor up to 10 billion microorganisms of possibly thousands of different species (8). He also states that soil diversity exceeds that of aquatic environments, and that is a great resource for biotechnological exploration of novel organisms, products and processes (8). Handelsman et al., 1998 explains that because microbes, generally, have great genetic diversity; soil carries the highest population of microbes of any habitat (9). With that exceptionally large amount of microorganisms in soil microflora, is evident that there is a remarkable amount of genetic material, unknown to us; that contains the “information” that enables the cell that enclose it, the ability to produce enzymes for the modification of a variety of substrates, for the production of a variety of compounds or for the development of mechanisms to resist and survive in the presence of a biocide agent. If we want to access that “treasure chest” of genetic material, we must overcome the cultivation problem to be able to work with the 99.9% - 99% uncultivated portion. Voget et al., 2003 explains that this problem can be solved by “direct isolation and cloning of metagenomic DNA, thereby circumventing cultivation” (10).

The metagenomic DNA or metagenome can be defined or described as collectively the genomes of the total microbiota found in nature (4); so using the same case as later, the soil; a soil metagenome sample can be defined or described as the collective genomes of the total microbiota or microflora found in the soil sample selected to be tested. The microbial ecogenomics, also called environmental genomics or more known as metagenomics is an approach that combines engineering, chemistry and biology as tools to unveil the genomes of soil microorganisms that as we said cannot be, or better said; have not been cultivated. So as Handelsman explains the discovery involved in metagenomic science goes by isolating the

environmental DNA from known and unknown microorganisms, the fragmentation of that DNA, the cloning of those DNA fragments into a culturable organism and the screening of the resultant clones for the production of new enzymes or chemicals (9). In other words the strategy is to isolate the metagenomic DNA directly from the soil sample, clone it in into a readily cultured organism such as *Escherichia coli*, and screen the clones for biological activity (9). Taking that sentence as the tip of the spear we have to explain how the environmental DNA was extracted from the soil samples selected and how that DNA was cloned in the metagenomic libraries that we are going to test, why was *Escherichia coli* the example cited and why it was the organism selected for expression of the environmental genes, which biological activities are we interested in and how to screen for those biological activities of our interest.

1.2.1.2 Constructing a Metagenomic Library

In our case, the Metagenomic Libraries that we are going to use were constructed from three types of soil samples: soil form the Dry Forest of Guánica, Puerto Rico; soil from the Rainy Forest from El Yunque, Puerto Rico and from the Hypersaline Microbial Mats of Cabo Rojo salterns, Puerto Rico. These metagenomic libraries were constructed in the Dr. Carlos Ríos Velazquez microbial biotechnology laboratory at the biology department of the University of Puerto Rico, Mayaguez Campus. These libraries were constructed using fosmids as the exogenous DNA carrying vectors. But how Dr. Ríos Velázquez Laboratory team constructed these libraries?

They had first to think on how big they wanted the environmental DNA inserts for the metagenomic library. Kakirde et al., 2010 explains that when extracting metagenomic DNA from a soil sample, the first consideration is DNA size (11). Kakirde keeps on explaining that if the goal of the study is high troughtput sequencing, PCR amplification, or small-insert clone libraries, then a harsh extraction method that results in substantial shear can be used if the extracted DNA is by the process highly purified (11). But in the case of designing a metagenomic library with big size inserts an alternate extraction protocol have to be selected to provide that big, pure and intact metagenomic DNA (11). Cruz et al., 2010 explains the use of a gentle process to obtain the approximate 40kb DNA fragments used to construct the metagenomic libraries that we have by using “freezing and thawing; gentle mix of the sample with SDS and pulse field gel electrophoresis in a low melting agarose gel” (12). That’s a process very similar of what Liles et al., 2008 cites as good for the recovery and purification of high

molecular weight DNA. Liles says: “High molecular weight genomic DNA was isolated using a combination of chemical and enzymatic lysis within an agarose plug” (13). Why this gentle treatment is important? Even when Rajendhran et al., 2008 explains that for DNA extraction from soil samples mechanical treatment is more effective and less selective than chemical lysis (14); they needed not only high quality and purity DNA but also high size with approximate 40kb in size, and harsh methods tend to produce DNA fragments of smaller size. Rajendhran presents examples of the mechanical methods used: thermal shocks, bead mill homogenization, bead beating, microwave heating and ultrasonication (14). However as we mentioned earlier it produces small fragments of DNA. The idea of having a metagenomic library composed of big DNA fragments is as Kakirde explains to have the potential of producing clones that contain “intact biosynthetic pathways involved in the synthesis of antimicrobial compounds, multiple enzymes with catabolic activity, or operons encoding other complex metabolic functions” (11). That’s why Rajendhran explains that no single method of cell lysis is suitable for all soils, different combinations and modifications of lysis protocols may be needed (14). So for the construction of the three metagenomic libraries that we have, three principal objectives had to be followed: 1.) isolate large fragments of environmental DNA from the soil samples not using the mechanical methods mentioned; 2.) shearing or cutting the DNA with methods that can produce large inserts that contain possibly intact biosynthetic pathways; and 3.) producing a DNA that contains the representation of Gram+ and Gram- communities. Gram+ bacteria wall is resistant to the chemical treatment and not always the DNA extraction can be obtained. Rajendhran says that “in a process using the combination of the gentle methods of grinding, freezing-thawing followed by SDS-based lysis; the DNA yield will contain Gram- bacteria DNA and two to six folds higher yield for most Gram + bacteria than using only a chemical approach” (14). This quotation is in accord with the method presented by Cruz et al., 2010 and explains that the process used for the extraction of environmental DNA of 40kb approximate size was a correct one. Now let’s answer a couple more of questions. Why the use of a fosmids as a carrying vector? How the vector was inserted into the host microorganism? Why to use *Escherichia coli* as the host microorganism?

1.2.1.3 Why is *Escherichia coli* a good election as host cell?

The system of expression for a metagenomic study is composed of a vector and a host organism that carries and expresses the genetic information encoded in the vector (15). Taupp et al., 2011 explains that currently, *Escherichia coli* is the dominant screening host for functional metagenomics. He also explains that most commercially available large insert library production systems utilize *E. coli* as a replication host (15). The genome of this microbial host is known, the growing conditions for this host are known. *E. coli* is able to grow under standard laboratory techniques and the scientist community has developed methods to manipulate it genetically. *E. coli* can grow in aerobic and anaerobic conditions using different electron acceptors like oxygen in aerobic conditions and nitrate in anaerobic conditions. This capability of being able to use a variety of electron acceptors is an advantage using *E. coli* as host organism. G. Unden et al., 1997 explains that because biochemical, physiological and molecular studies using *E. coli* had provided detailed knowledge of its respiratory chains and adaptability to the environment and to different enegetic demands; *E. coli* is one of the preffered bacteria to study energetics and regulation of respiration (16). And for the same reason it is a great host for metagenomic studies based on experiments focused on phenotype expression; because those types of experiments depends greatly on bioenergetic demands and electron acceptors to display the genes encoded in an insert.

Once we have selected a host cell we can start the process to construct the metagenomic library. To construct a metagenomic library to later test the clones in designed growth mediums; DNA fragments can be attached to self-replicating units, transferred into appropriate host cells, cloned and amplified (17). Chemical synthesis presents the possibility of constructing and cloning DNA sequences that may not occur in nature (18). There are a couple of options when selecting a vector for the construction of a metagenomic library. Some of them are λ bacteriophage (λ phage), *E. coli* F factor based plasmids, bacterial artificial chromosomes (BAC's), cosmids and fosmids.

1.2.1.4 Why was a fosmid elected as carrying vector and how it was incorporated in the host cell?

A λ phage virus (virion) has a head, which contains the viral DNA genome, and a tail, which functions in infecting *Eschericia coli* host cells like a syringe, by injecting the DNA contained in the virus head into the host cell. When λ DNA enters the host cell cytoplasm following infection, it undergoes either lytic or lysogenic growth. In lytic growth, the viral DNA is replicated and assembled into more than 100 new viruses in each infected cell, exploding the cell (killing it) in the process and releasing the replicated virions. In lysogenic growth, the viral DNA inserts into the bacterial chromosome, where it is replicated along with the host cell chromosome as the cell grows and divides but don't die because more virions are not produced (19). When a bacteriophage λ is used as a cloning vector, it must be capable of lytic growth, but other viral functions are not needed. Because of that, the genes involved in the lysogenic pathway and other viral genes not essential for the lytic pathway can be removed from the viral DNA and can be replaced with the DNA to be cloned (19). To prepare a set of infectious λ virions carrying recombinant DNA (like environmental DNA joined to the DNA of the virus); the λ DNA first is cutted with a restriction enzyme to produce fragments called λ vector arms, which have sticky ends without disrupting the genes necessary for lytic growth. This step takes out the dispensable non essential region in the middle of the λ genome; which is separated from the λ vector arms and thrown away (19). Genomic DNA then is extracted from a cell or a community of microorganism that contains all the genetic information that someone wants to study. The extracted DNA then is cleaved by a restriction enzyme to produce aproximatelly 20 kilobases (kb) fragments with sticky ends complementary to the sticky ends on the λ vector arms being used (19). The λ arms with sticky ends and the collection of genomic DNA fragments with sticky ends are mixed in about equal amounts. The complementary sticky ends on the fragments and λ arms hybridize and then are joined covalently by DNA ligase. Each of the resulting recombinant DNA molecules contains a DNA fragment different form the original viral DNA that was located between the two arms of the λ vector DNA. The ligated recombinant DNAs then are packaged into λ virions in vitro (19). One method that can be used to prepare these recombinant DNA molecules is to infect *E. coli* cells with a λ mutant virion defective in A protein, one of the two proteins required for packaging λ DNA into preassembled phage heads. These cells will accumulate preassembled "empty" heads with no tails attached to them; since tails attach only to heads "filled" with DNA. Preassembled tails also accumulate in these cells. Then these cells are lysed and an extract containing high concentrations of empty heads and tails is prepared.

When this extract is mixed with the missing protein A (obtained from lysing λ-infected cells) and recombinant λ DNA containing a cos site, the DNA is packaged into the empty heads. Once the heads are filled with the recombinant DNA, the tails in the extract combine with the filled heads; yielding complete virions carrying the recombinant λ DNA. This procedure yields fully infectious recombinant virions that can efficiently infect *E. coli* cells (19). “After the cells are infected the identification of the recombinant phages can be done by hybridization of the recombinant DNA or by means of its antigenicity” (17).

A plasmid is a circular DNA molecule able of replicating in the cytoplasm of some microbes. Some plasmids have not essential sites for their replication that can be recognized by certain restriction enzymes. A plasmid in linear form may be joined to another DNA fragment, recircularized, and re-inserted into a host with no loss of ability to replicate. Plasmids with this property can be used as cloning vectors (18). The features of the plasmids vectors to be considered include: the mode of plasmid replication, the presence of markers that help in the selection of the host microbe that contains the plasmid (such as antibiotic resistance), the presence of one or a number of sites at which DNA can be cloned without destroying the ability to replicate or the selectability of the plasmid, the disponibility of sites at which DNA may be cloned to produce an specific and identifiable phenotype (such as fluorescence or ability to degrade an specific compound), the presence of regulatory elements which may be used to cause expression of a gene or genes on a cloned DNA fragment, and the presence of mutations that boost the biological containability of the plasmid permiting to have a stable insert in the host cell (18). There are two classes of plasmids that have been used as cloning vectors in *E. coli*. One class is only present in a few copies per cell and undergoes stringent replications; the other present in many copies per cell undergoes relaxed replication (18). One of the best known plasmids is the F (fertility) factor plasmid. The F factor plasmid of *E. coli* K-12 of approximately 100 kb in size was the first plasmid to be fully described. Important deductions in describing the F factor plasmids were that it was able to replicate autonomously and it was stable (20). Some important functional regions on the F factor plasmid are: a transfer region, a leading region, an origin of conjugative transfer (*OriT*), replication regions (*Rep FIA*, *Rep FIB*, *Rep FIC*) and transposable elements (*Tn*1000, *IS*2, *IS*3). The replication regions act in concert to maintain the plasmid at one to two copies per cell helping to stabilize the DNA cloned in the vector (20). *E. coli* F factor based plasmids have some or all of these important characteristics that make this stable plasmid that follow stringent replication a good instrument in building DNA libraries.

Both λ phage vectors and the more commonly used *E. coli* plasmid vectors are useful for cloning DNA fragments up to aproximately 20 to 25 kb. However, cloning of much larger fragments is wanted for sequencing of long DNAs such as the DNA in a eukaryotic chromosome or such as sequences isolated from environmental samples enriched by a multitude of prokaryotic and eukaryotic microorganism (19). Remember that we want to have "clones that contain intact biosynthetic pathways involved in the synthesis of antimicrobial compounds, multiple enzymes with catabolic activity, or operons encoding other complex metabolic functions" (11). Consequently to construct a metagenomic library able to hold large insert we need a different vector other than λ phage vectors or *E. coli* F plasmid vectors. Large insert libraries can be constructed by the use of alternative vectors like bacterial artificial chromosomes (BAC's), cosmids and fosmids.

A bacterial artificial cromosome or BAC is a DNA construct, based on the *Escherichia coli* F factor (21). The bacterial artificial chromosome has an approximate insert size that goes between 150-350 kbp. Some inserts in high copy vectors tend to suffer from insert instability, having rearrangements or deletions in the cloned DNA causing difficulties in maintaining large intact DNA in bacteria; especially DNA from eukaryotic organisms. Conveniently the F factor codes for genes that are essential to regulate its replication controlling its copy number to one or two copies per cell. The regulatory genes in an F factor base plasmid include *oriS*, *repE*, *parA*, and *parB*. The *oriS* and *repE* genes mediate the replication in one direction of the F factor while *parA* and *parB* maintain the number of copies of the plasmid at a level of one or two copies per *E. coli* cell. The BAC vector contains all these essential genes as well as an antibiotic resistance selection marker and a cloning site (21). "The cloning segments include the bacteriophage λ cosN sites, cloning sites and rich C + G restriction enzyme sites for potential excision of the insert. The cloning site can be flanked by promoters for generating RNA probes for chromosome walking, and for DNA sequencing of the inserted segment" (21).

A cosmid vector is a type of hybrid plasmid that contains a λ phage c*os sequence*. "Is produced by inserting the C sequence from λ phage DNA into a small *E. coli* plasmid vector about 5 kb long" (19). Like the *E. coli* F factor based plasmids and the BACs; cosmid vectors contain a replication origin (*Ori*), an antibiotic resistance selection gene marker, and numerous restriction enzyme recognition sites for cloning of DNA (19). This vector is used in combination of a λ virus to infect the host cell. Cosmid plasmids cannot be packaged because they are circular. Monomeric circular DNA that contains only one cos site cannot be packaged in vivo or

in vitro because λ phage proteins required for packaging a DNA molecule don't recognize a packable element of DNA unless it contains a duplicated cos site (22). There are precise recognition regions on the DNA that are required for packaging a DNA molecule into an λ phage particle to be able to infect a cell: a site close to the left cos site or cohesive end, which might be recognized by the p(A-Nu 1) proteins that λ needs for packaging λ DNA into preassembled virion heads; and the cos site as substrate for the pA containing terminase. Only a phage DNA particle with cohesive ends is infectious thus the cos site is very important (22). For this reason the cosmid vector is cut with a restriction enzyme and then ligated to 35 to 45 kb restriction fragments of foreign DNA (like environmental DNA) with complementary sticky ends (19). Lambda in vitro packaging system apparently doesn't care of the type of DNA inserted adjacent to the small λ region required for packaging. This can be proven by the fact that most of the λ DNA, including all regions required for λ replication and lysogeny can be removed and replaced by foreign DNA and efficient packaging will still occur. The packaging then depends only on the joined λ DNA plus the exogenous DNA being of a certain minimum size and on the presence of concatemeric DNA formed by in vitro restriction endonucleases cleavage and ligation at high DNA concentrations (23). "In the packaging reaction, the λ Nu1 and A proteins bind to cos sites in the ligated DNA and direct the insertion of DNA between two adjacent cos sites into empty phage heads. Opened forms of cosmids, ligated to a fragment of DNA to be cloned, resemble concatemeric DNA and are thus packaged in a λ coat" (22). Packaging will occur always if the distance between contiguous cos sites does not exceed about 50 kb. One reason that can explain the 50 kb maximum limit is that 50kb is the approximate size of the λ genome. (19) Once the plasmid and the DNA desired for cloning is ligated and the heads of the phage are filled; phage tails recognizes filled heads and then are attached to those filled heads, producing viral particles that contain a recombinant cosmid DNA molecule instead of the λ genome. During the packaging of the cosmid hybrid DNA, the DNA of the bacteriophage used to produce the packaging mix is also packaged and some viral particles are formed. For this reason the injection of the recombinant DNA must be carried in a virus resistant *E. coli* host cell (23). Once these viruses are plated on a lawn of grown *E. coli* cells, they attach to phage receptors on the surface of the cells and injects their recombinant DNA to the cells (19). Once injected the recombinant DNA (a cosmid hybrid DNA in linear form), it circularizes when the separated cos sites join together (23). Since the injected cosmid hybrid DNA doesn't have any λ genes except the cos sites, no viral particles form in infected cells, no cells die by the infection of phages containing cosmid hybrid DNA and no plaques develop on the plate. Instead of that the injected DNA forms a large circular plasmid that is composed of the cosmid vector and the inserted DNA

fragment; a different circular plasmid in each host cell. This plasmid replicates to daughter cells like other *E. coli* plasmids, and the colonies that come from transformed cells can be selected on antibiotic plates because they have the resistance marker. The main advantages of the cosmid system over other plasmid cloning vectors are: the high efficiency of hybrid clone formation; the fact that essentially only hybrid plasmids are formed, and the fact that the system selects particularly for large fragments (23).

With all that backgrownd in mind we can now better understand what a fosmid is. A fosmids is an f-factor cosmid, which is capable to contain aproximatelly 45 kb of DNA. “Fosmids contain replicons derived from the *E. coli* F factor for DNA replication and segregation. Because of this they are more stable than cosmids and are suitable for the rapid generation of genomic or chromosome specific libraries” (24). There are reports of instability of complex mammalian eukaryotic DNA in cosmids whereas in fosmids those instabilities had a greater reduced frequency (25). Our metagenomic libraries can contain genes from prokaryotes and eukaryotes like bacteria and yeast respectively. The DNA of those microbes is not as complex as the DNA found in complex mammalian genomic DNA nonetheless fosmids clearly presents an advantage in sustaining a complex insert of metagenomic DNA undergoing changes at greatly reduced frequencies and holding its integrity from deletions or rearrangements even with multiple generations of the clone that contains the insert and most of the time multiple generations are needed to perform the experiments and to recover the insert information. Unlike plasmids, *E. coli* can only hold and replicate one fosmid. For that reason fosmids can hold larger pieces of DNA than plasmids, but fewer of them. Béjá explains that since 1992 genomic cloning efforts have dramatically improved because of the introduction of bacterial artificial chromosomes (BACs) and F1 origin based cosmid vectors (fosmids) as tools used to clone bigger fragments of DNA with more stable inserts (26). Fosmids libraries with insert sizes ranging between 32 and 45 kb are packed in phage lambda extracts and adsorbed directly to host cells in the same process as we described for the cosmids (15). All these arguments presented point to the conclusion that for the construction of large insert libraries, the vector of choice is among fosmids, cosmids or BAC’s. Is important to mention that “the low copy number status of fosmids or BAC’s within *E. coli* promotes stable replication and minimizes over expression of toxic genes” (15). But I think that the group of Dr. Carlos Ríos Velázquez chose fosmids as vectors for the reason explained by Béjá: “Environmental BAC libraries enable screening of larger inserts when compared to fosmids libraries (up to 200 kb vs. 40 kb), but demand a larger amount of starting material and can also contain rather small inserts (there is no selection

against small fragments as in the fosmids system). Using the fosmids system, one can extract enough DNA from less starting material and have enough material to construct a good library with uniform sized inserts of about 40kb" (26).

1.2.1.5 Designing a metagenomic library screening method

Once the metagenomic libraries are constructed; the next step is design the screening method to identify a clone or a set of clones that have inserts with the genes that interest us. How to screen the metagenomic libraries we have for the activities we are interested in? There are two different approaches to screening for target gene-containing clones: activity and sequence-based screening. An activity based screening will require analyzing in a single experiment several hundred thousand clones in order to detect the phenotype of a few functionally active clones by observing the growth of the clones in a specific media preparation (27). Sequence based screening uses known gene families and conserved DNA regions to design PCR primers or hybridization probes to target genes with similarities with the genes known on current databases (27). Labor intensive analyses and experimental preparation will still be required for both activity-based and sequence-based screening procedures. We selected to use an activity-based approach because is less expensive than the sequence based screening. Another reason is that gene-specific PCR has two major drawbacks. First the design of primers is dependent on existing sequence information and bias the search in favor of known sequence types and unknown sequence types will have a decrease probability of being detected, most of them not being detected. Another reason is that, only a fragment of the gene of interest will be amplified by this method, making more difficult and requiring more steps to get to full length genes (28). For the activity–based screening approach there are two mainly possibilities to isolate "one book from the entire library"; one clone from the entire community of clones to further study the genetic insert of the isolated clone: screening the libraries with minimal mediums designed with a specific and defined carbon source to select and isolate the clones that grows on the media or screening the libraries with colorimetric methods in rich mediums that enable the researcher to select and isolate by visual means the clones with a different colony color or halo around the colony that indicates the presence of a genotype that is giving the clone the ability to present a specific phenotype. We selected to use the minimal media approach. Even when is a stressful method for the host cell (*E. coli*); clones in minimal medium face the need to synthesize all of their building blocks from a single carbon and energy source, and that burden is reflected in the turning of the clone biosynthetic pathway and in the

expression of regulators of cell processes and regulons involved in stress tolerance and pushes the cell to use a carbon source not used by the cell unless an insert provide the clone the mechanism to process the carbon source being tested (29). Also for the same reason that Taupp says that “the low copy number status of fosmids minimizes over expression of toxic genes”; the minimization of the over expression of a gene present in an insert could affect a colorimetric assay that expects the over expression of an enzyme to be able to detect a genotype present via a visual reference. Instead of that the minimal media approach will tell us that there is a possibility of having the genetic information we are searching for by a simple method; if the clone grew or not.

1.2.1.6 Selecting biological activities to test for.

At last we have to identify which biological activities are we interested in and which carbon sources are we going to use for the preparation of the minimal media to screen the metagenomic libraries we have selected to use. Industrial or White Biotechnology is a term stamped in 2003 by the European Association for Bioindustries (EuropaBio), that divides all bio-based processes that are not included in the Red Biotechnology (medical) or Green Biotechnology (plant) labels and includes them in a new clasification (30). The new classification will be focused on the production of biobased fuels and on the production of new high value chemicals like nutraceuticals and bioactives (30). We are interested in finding genes that codifies for enzymes useful for replacing fossil fuels with bio fuels and in genes that can use waste materials and convert them to high value products. There is a high interest in the use of waste crop biomass as the renewable resource or “raw material” to be used to replace fossil fuels with bio fuels. This interest comes because there are countries like Brazil that uses sugar cane to produce ethanol via microbial fermentation and countries like the United States that instead of sugar cane uses corn to produce ethanol via microbial fermentation. But those (sugar and corn), are products that have value in our diet, which can be consumed by the society. The use of “food” for the production of a bio fuel has an impact on the availability of the resource for uses different to bio fuel production. Also the use of food crops with value in our diet to produce bio fuels impact the price of the product as itself, as food; raising it and making difficult for poor people to obtain the product and for companies that need the product at a reasonable price for animal feeding to maintain a cost effective process. Therefore there is a high interest in the use of “raw materials” without value in our diet, because we don’t want to use food crops to convert them to bio-fuels; instead of that, we want to use waste crop biomass, switch grass, wood chips

and other sources of lignin, cellulose and hemicelluloses to convert to ethanol and other biofuels.

Recently we are seen an increasing tendency toward the efficient utilization of agro industrial residues such as coffee pulp and husk, cassava bagasse, sugar cane bagasse, sugar beet pulp, apple pomace, etc (31). If we take coffee waste crop biomass for example to explain the potential of crop waste microbiological treatment for the production of biofuels and/or value added fine products the reason for this tendency will look clearler. Some waste crop biomass like coffee crop waste in addition to the content of lignin, cellulose and hemicelluloses contains caffeine and additional carbon sources that can be used by microorganisms as building blocks for the production of different products. Currently about one million tons of coffee is produced yearly in more than 50 countries (31). That amount is a large quantity of wasted carbon sources. There are two methods for processing coffee: dry and wet processing. Depending upon the method of coffee cherries processing, wet or dry; the solid residues obtained are termed as pulp or husk respectively. Coffee pulp and husk that are un-utilized waste; are becoming a pollution problem, and are composed of carbohydrates, proteins, fibers, fat, caffeine, tannins, polyphenols and pectins in addition to the content of lignin, cellulose and hemicelluloses (31). For example thinking on the lignin content of the coffee crop waste we found in literature that it contains allot of aromatic compounds. Martinez studied the composition of phenolic compounds in coffee pulp and revealed the main constituents of it: chlorogenic acid [5-caffeoylquinic acid] (42.2%) as the main constituent, epicatechin (21.6%, isochlorogenic acid I, II and III, 5.7%, 19.3%, 4.4%, respectively), catechin (2.2%), rutin (2.1%), protocatechuic acid (1.6%), and ferulic acid (1.0%) (31), (32). Other crops waste has different nutritional content but the important thing is that we are wasting allot of carbon sources that can be used to produce value added fine chemicals. Thinking in the composition of the crop wastes that are out there we decided to select as carbon sources for our study: soluble cellulose, lignin alkali, veratryl alcohol, caffeine and tween. If a clone from the library grows on a minimal media with one of these carbon sources it may mean that the clone has an insert that is enabling the host cell to use a carbon source that once was unable to use and to transform it in a product of interest for the society.

1.2.1.6.1 Cellulose

Cellulose is an interesting compound to select as carbon source to screen the metagenomic libraries searching for cellulases producing clones for multiple reasons. Cellulose is the primary product of photosynthesis in terrestrial environments, and the most abundant renewable bioresource produced in the earth (33). Cellulose is, next to chitin, in the list of the most abundant renewable energy sources; plants usually contain from 35% to 50% (dry weight) cellulose (34). Cellulose is the most abundant renewable source of fixed carbon produced by plants in tons each year (35). Cellulose, the most abundant biopolymer on earth; can be described as a linear polymer composed of β-glucose molecules joined together (36). To complete full degradation of cellulose an organism will need: endoglucanases (1,4-β-D-glucan-4-glucanohydrolases); exoglucanases including cellodextrinases (1,4-β-D-glucan glucanohydrolases) and cellobiohydrolases (1,4-β-D-glucan cellobiohydrolases); and β-glucosidases (β-glucoside glucohydrolases) (34). Cellulases have many industrial applications from the generation of bio-ethanol, to the finishing of textiles (37). Cellulases are used in the textile industry for cotton softening and denim finishing; in the detergent market for color care, cleaning and anti-deposition; in the fuel industry as biofuel or fuel additive; in the food industry for mashing; and in the pulp and paper industries for deinking, drainage improvement, and fiber modification (33). The potential cellulose market has been estimated to be as high as US $400 million per year if cellulases are used for efficiently hydrolyzing the available corn stover in the Midwestern United States (33).

The majority of the known investigated prokaryotic cellulases have been isolated from cultured microorganisms (34). Up to 2011 cellulases are classified into 14 glycoside hydrolase families (GHF) (GHF-5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 61 and 74) (38). At this moment, most commercial cellulases (including β-glucosidase) are produced by *Trichoderma* species and *Aspergillus* species (33). Even when remains much interest in the isolation of cellulases from fungal sources, there has been a recent increase in the isolation of diverse novel cellulases from prokaryotic organism. Some examples of bacterial isolates that have cellulases activities are: *Cellulomonas pachnodae*, *Clostridium cellulovorans*, *Bacillus subtilis* (39), (40), (41). Nevertheless even when most of the success of finding cellulases has been using culturable microbes, some are betting on finding the next generation of industrially applicable celluloses on metagenomic library constructed using samples from harsh environments like

hypersaline microbial mats, acidic and termal pools and contaminated sites. Finding enzymes in such environments can imply a higher possibility in finding enzymes tolerant to harsh conditions also found in industrial processes like high substrate concentrations, tolerance to organic solvents, pH variations, salinity variations and temperature variations. For example S. Voget et al., 2006 reports a metagenome derived halotolerant cellulose that is highly stable, salt and pH tolerant, an ideal candidate for industrial applications (37).

In bacteria there are two different structural types of cellulose degrading enzymatic systems. Those systems are termed noncomplexed and complexed systems. Diferent from aerobes some anaerobes produce an extracellular multienzyme complex called cellulosome. “A cellulososme consist of different hydrolases organized on a noncatalytic scaffolding protein that mediates the attachment to cellulose” (34). We selected Carboxymethyl Cellulose as carbon source for the minimal media for testing the clones in search for cellulose degrading candidates. Carboxymethyl cellulose (CMC) can be used for determining endoglucanase activity, because endoglucanases cleave intramolecular β-1,4-glucosidic bonds randomly, resulting in a reduction in the degrees of polymerization of CMC, resulting in a reduction of the viscosity of the solution and in the liberation of free sugars (33). Also by the use of CMC as carbon source we can detect enzymes with exoglucanases activities. The modes of action of endoglucanases and exoglucanases are different between them in that endoglucanases decrease the specific viscosity of CMC significantly with little hydrolysis (probably in less time) because of cleavages of the molecule of cellulose on the inside of the chains; whereas exoglucanases hydrolyze long chains from the ends in a continuous process (33). Also even when CMC is soluble cellulose, it can help in the discovery of a cellulase with activity over crystalline cellulose because some endoglucanases act on crystalline cellulose. The biodegradation of crystalline cellulose most of the time involves the action of both endo and exo acting cellulases. “Classical endoglucanases nick the cellulose internally, thus disrupting its crystalinity and generating new free ends in the polymer. Exoglucanases act processively from these free ends, remaining attached to the cellulose and releasing soluble cellulose molecules, which are subsequently hydrolyzed to assimilate glucose by β-glucosidases” (42). If a clone is found to be positive for the degradation of CMC, we can test later if the enzyme encoded in the insert can degrade substances corresponding to exoglucanases and β-glucosidases to be able to localize where the enzyme belongs; in a new category or in one of the 14 known categories. Considering all those aspects of the degradation of cellulose and the objectives for our research, we elected CMC as the carbon source to be tested in our minimal media to screen for cellulose degrading clones.

1.2.1.6.2 Lignin and veratryl alcohol

Lignin is an interesting compound to select as carbon source to screen the metagenomic libraries searching for lignin degrading enzymes also known as ligninases. Lignin is a major component of plant biomass and is the most abundant aromatic substance present in the globe (43). Lignin is a compound with a recalcitrant nature that makes it very difficult to be degraded by enzymes. Lignin creates a protection barrier by linking to both hemicelluloses and celluloses inhibiting the penetration of enzymes to the inside of the matrix where hemicelluloses and celluloses are guarded, drastically reducing the access to a more easy degrading energy source to microorganisms (44). “Chemically lignin is a heterogeneous, optically inactive polymer, consisting of phenylpropanoid interunits linked by different types of covalent bonds” (43). The predominant building blocks of lignin are β-O-4-linked ethers arranged in a complex non repeating three dimensional array (43). Lignins are generally classified into the three major groups: gymnosperm, angiosperm and grass lignins. The structural composition units of all of the major groups consist of different proportions of coniferyl, sinapyl, and p-coumaryl alcohol polymers. For this reason many researchers have used different phenols with structural simmilarity to the lignin building blocks for isolation, selection, and degradation studies of lignin decomposition by microbes (45). Because of the complexity and because of the recalcitrant nature of lignin polymers, most lignin degradation enzymes studies have been realized using lignin model compounds (46). After researching for lignin models we selected lignin alkali and veratryl alcohol as lignin models and as carbon sources for the minimal media experiments related to finding ligninases in the metagenomic libraries.

Lignocellulosic wastes are produced in large amounts by many industries including forestry, pulp and paper, agriculture, and food industries. Wastes with lignocellulosic material are also present in municipal solid waste, and animal wastes. These wastes were thrown away in many countries in the past, and still today are thrown away in some developing countries, which raise many environmental concerns (44). However those materials considered as waste can be used to produce fine valuable products, solving the environmental problem and creating bio-substitutes to some important chemicals. “Effective conversion of recalcitrant lignocelluloses to fermentable sugars requires three sequential steps: 1.) size reduction, 2.) pretreatment/fractionation, and 3.) enzymatic hydrolysis (33)”. The most difficult step is step 3 because lignin has to be chemically or enzymatically hydrolyzed to produce fermentable sugars

and aromatic compounds that can be transformed in value added products. Ligninases have an assortment of industrial applications such as: biofuel production, animal feed preparation, biological pulping of paper, fiber bleaching, remediation of organopolutants and aromatic pollutants like pesticides, halogenated aromatics and other chemicals that are detrimental for the environment like pentachlorophenol (46).

A diverse arrange of species are involved in lignin biodegradation, including fungi, plants, animals and also bacteria (43). “In nature, efficient lignin degradation during the process of wood decay became possible mainly by basidiomycetes white-rot fungi” (44). Brown rot fungi, can degrade wood carbohydrates, but cannot oxidized lignin. Most ascomycetes are able to degrade cellulose and hemicelluloses, yet their ability to convert lignin is very limited (44). Experiments that measure the rate of mineralization of ^{14}C labeled lignin have discovered that ligninolysis or lignin degradation is triggered by nutrient limitation (46). This nutrient limitation makes the microbes produce those enzymes required for the mineralization of lignin to make more accessible the celluloses and hemicelluloses contained inside which can be further transformed to simpler sugars that can be used by the microbes that are feeling that nutrient restraint, as preferred carbon sources. Fungi degrade lignin by secreting ligninases that can be classified as either phenol oxidases (laccases) or heme peroxidase (MnP) and versatile peroxidases (VP) (44). “Lacases or phenol oxidases are glycosylated blue multi-copper oxidoreductases that use molecular oxygen to oxidize various aromatic and non aromatic compounds through a radical catalyzed reaction mechanism” (44). Laccases have been found in white rot fungi like *Lentinus tigrinus, Pleurotus ostreatus, Cerrena unicolor, and T. versicolor*. Laccase has also been found in brown rot fungi like *Coniophora puteana*. There are not allot of ascomycetes with the ability to degrade lignin yet an examples is *Melanocarpus albomyces* a microbe able to produce laccases (44). Peroxidase contains a heme group and can be clasified in two: lignin peroxidases and manganese peroxidases. “Lignin peroxidases (LiP) are heme containing glycoproteins that catalyze the H_2O_2-dependent oxidative depolymerization of a variety of non-phenolic lignin compounds (diarylpropane), β-O-4 non-phenolic lignin model compounds and a wide range of phenolic compounds like guaicol, vanillyl alcohol, catechol, syringic acid, acteosyringone” (44). “Manganese peroxidases (MnPs) are extracellular glycoproteins secreted in multiple isoforms which contain one molecule of heme as iron protoporphyrin IX. MnP catalyzes the peroxide dependent oxidation of Mn(II) (as the reducing substrate) to Mn(III), which is then released from the enzyme surface in complex with oxalate or with other chelators” (44). Another type of lignin degrading enzymes called versatile

peroxidases (VP) are glycoproteins that have dual oxidative ability and are able to oxidize Mn(II) and also phenolic and non phenolic compounds (44). “Other fungal lignin degrading accessory enzymes include oxidases generating H_2O_2, which provide the hydrogen peroxide required by peroxidases, and mycelium-associated dehydrogenases, which reduce lignin-derived compounds. Oxidases generating H_2O_2 include aryl-alcohol oxidase (AAO) and glyoxal oxidase (GLOX). In addition aryl-alcohol dehydrogenase (AAD) and quinine reductase (QR) are also involved in lignin degradation by fungi. Moreover, it has been shown that cellobiose dehydrogenase (CDH), which is produced by many different fungi under cellulolytic conditions, is also involved in lignin degradation in the presence of H_2O_2 and chelated Fe ions” (44).

Fungi; specially white-rot as demonstrated by the papers and reviews quoted have a specialized “tool box” full of enzymes that enables them to degrade lignin more efficiently than any other organism. Nonetheless there are examples where actinomycetes and other bacteria have been identified as lignocelluloses degrading microorganism. These strains have been isolated from a great variety of aerobic and anaerobic sources, including compost soil, terrestrial environments, and aquatic ecosystems (43). “Of those strains, three main morphological forms of plant cell wall degradation by bacteria have been discovered: tunneling, erosion, and cavitation” (43). Tunneling bacteria are rare in waterlogged wood because they appear to require a good supply of oxygen and waterlogged wood is located mainly in an anoxic or anaerobic environment. Tunneling bacteria produce minute tunnels that they use as way to migrate through the cell wall. Oposite to tunneling bacteria, erosion bacteria that attack the wall from the lumen are responsible for the predominant form of degradation in waterlogged archaeological wood, because they seem to tolerate near anaerobic or fully anoxic environments. Cavitation bacteria have been found in the wood cell lumen, and are believed to utilize products derived from the degrading activities of lignin degraders (43).

Most products produced by wood degraders biodegradation activities are aromatics compounds. Pathways for the catabolism of aromatic compounds play a crucial role in the mineralization of organic matter, in the carbon cycle of plant biomass because a big part of plant biomass is composed of lignin subunits. For that reason is not surprising to find enzymatic routes in different microbes, in prokariotes and eucariotes capable of using as carbon sources the diverse aromatic compounds resulting from the initial depolymerization accomplished mostly by fungi (47). Bacteria of several genera including *Alcaligenes*, *Arthrobacter*, *Nocardia*, *Pseudomonas*, and *Streptomyces*, have been found to degrade single ring aromatic substrates

(43). Abd-Elsalam et al., 2009 quotes that Perestelo et al., 1996 and Morii et al., 1995 have studied lignin degradation with unicellular bacteria and that bacteria isolated from compost soil (*Azotobacter*, *Bacillus megatarium* and *Serratia marcescens*), were capable of decolourizing, mineralizing or solubilizing lignin (48). Known bacteria that are related to the degradation of aromatic compounds product from lignin degradation are aerobic but there are reports of degradation of ^{14}C-labeled lignin in anaerobic conditions. Even when the microorganisms that are causing anaerobic decay of lignin subunits have not been characterized; both $^{14}CO_2$ and $^{14}CH_4$ (a bio fuel) have been detected. This information tells us that lignin degradation occurs even in natural anoxic conditions, adding to the recycling cycle of carbon and adding to the diversity in enzymes involved in the degradation of lignin or the monomeric compounds produced after its first depolimerization (43).

There are two major groups of intracellular enzymes of aerobic bacteria that are involved in lignin degradation: peroxidases and phenol oxidases. Peroxidases have been found in bacteria, fungi, plants, and animals. “In consideration of the sequence similarity and structural divergence, they are viewed as belonging to a super family consisting of three major classes: mitochondrial yeast cytochrome c peroxidase, chloroplast and cytosol ascorbate peroxidases, and gene duplicated bacterial peroxidase (class I); secretory fungal peroxidases (class II); classical, secretory plant peroxidases (class III)” (43). Phenol oxidases can be divided into two subgroups: laccases and polyphenol oxidases. Lacases are oxidoreductases able to catalyze the oxidation of various aromatic compounds with the reduction of oxygen to water (like phenols). Because of its big size, lacases alone can only oxidize phenolic lignin units at the substrate surface because they can’t penetrate the recalcitrant interliked units of lignin. For that reason laccases are often used with an oxidation mediator, a small molecule that is able to extend the effect of laccase to non-phenolic lignin units and to overcome the accessibility problem penetrating into the recalcitrant matrix. For example the mediator, called LMS, is first oxidized by laccase and then diffuses into the cell wall to oxidize the inaccessible parts of lignin where laccases alone cannot get (43). “Polyphenol oxidases or tyrosinases (PPO), containing a dinuclear center, are able to insert oxygen in a position *ortho* to an existing hydroxyl group in an aromatic ring with the concomitant oxidation of the diphenol to the corresponding quinine. For lignin degrading anaerobic bacteria all oxygen dependent reactions are replaced by a set of alternative enzymatic processes that uses nitrate, sulfate or Fe(III) as terminal electron acceptors for its electron respiratory chains. It is suggested that some of the enzymatic processes involved in the degradation of phenol by anaerobes uniquely exist in the aromatic

metabolism of anaerobic bacteria" (43). "The anaerobic phenol metabolism route is composed of the following reactions: Two different enzymes are involved in the carboxylation of phenol: a phenylphosphate synthase that transforms phenol with ATP into phenylphosphate, forming AMP and phosphate and a phenylphosphate carboxylase that transforms phenylphosphate with CO_2 into 4-hydroxybenzoate and phosphate, using Mn as the metal cofactor. By a specific AMP-forming carboxylic acid coenzyme A ligase, 4-hydroxybenzoate is first activated to 4-hydroxybenzoyl-CoA. Then the phenolic coenzyme A ester is dehydroxylated to benzoyl-CoA by 4-hydroxybenzoyl-CoA reductase. Benzoyl-CoA is reduced to a non-aromatic, cyclic dienoyl-CoA compound by benzoyl-CoA reductase. According to the hydrolyzation of two molecules of ATP, benzoyl-CoA reductase couples aromatic ring dearomatization to a stoichiometric hydrolysis. Then the product is further converted to three molecules of acetyl-CoA and one molecule of CO_2". (43)

1.2.1.6.3 Caffeine

Caffeine is an interesting compound to select as carbon source to screen the metagenomic libraries searching for caffeine degrading enzymes. Caffeine ($C_8H_{10}N_4O_2$) is an alkaloid naturally occurring in coffee beans, cola nuts and tea leaves, and is also called 1, 3, 7-trimethyl xantine (49). "Currently about one million tons of coffee is produced yearly in more than 50 countries" (31). Brazil is the largest producer of coffee in the world, contributing approximately 25% of the world's production. During 1999, its production reached two million tons. Only 6% of the coffee cherries (fresh weight basis) were the portion produced as coffee powder; the remaining 94% was by-product of the process or waste (50). "Caffeine is one of the major toxic compounds generated by solid waste in the coffee and tea industries. In spite of the fact that these wastes are enriched with carbohydrates and proteins, they cannot be used as animal feed because of the presence of caffeine and other toxic compounds. The caffeine in liquid effluents of coffee and tea industries cannot be allowed to be fed into lakes and rivers as it would affect the aquatic ecosystems. Hence, caffeine degradation is of importance in view of health as well as general environmental concerns" (51). A variety of processes have been developed to use these agro industrial residues as raw materials for the production of bulk chemicals and value-added fine products such as: ethanol, single cell protein (SCP), mushrooms, enzymes, organic acids, amino acids, biologically active secondary metabolites, methane, etc. (50). Even when there are some caffeine catabolising fungal strains isolated, caffeine catabolic pathway in fungi is not elucidated in detail because the intermediates of

caffeine degradation are rapidly metabolized to simpler diffusible elements and have been not detected by the instrumentation used in research. Researchers have assumed that the degradation of caffeine in fungi takes place via the demethylation route, with theophyline as the first formed intermediate. If that's true, the caffeine degradation pathway in filamentous fungi closely resembles the caffeine degradation pathway of plants (51). "In plants, degradation of caffeine occurs through sequential demethylation that finally results in the formation of xantine. The demethylation reactions have been found to be catalyzed by demethylase enzymes: N-1 demethylase, N-7 demethylase and N-3 demethylase. Xantine is then converted to CO_2 and ammonia by purine catabolism" (52). Even when fungi resembles the mechanism of plants, in yeast, caffeine degradation is brought about by cytochrome P450 enzymes and therefore the degradation pathway would be more similar to the mechanism in animals. (53) "The degradation pathway in animals; like in mammals result in the formation of methyl xanthenes and methyl uric acids by cytochrome P450 enzymes CYP1A2, CYP3A4, CYP2E1 zanthine oxidase and N-acetyl transferase. Methylated xanthenes and the respective uric acid formed due to degradation are excreted from the body through urine" (52). Even when caffeine catabolic pathway in fungi is not elucidated in detail, bacterial caffeine catabolism is well elucidated. (51) In aerobic bacteria, two major classes of enzymes are responsible for the degradation of caffeine: N-demethylases and Oxidases (which include caffeine oxidases and xantine oxidases) (53). For example in *Pseudomonas*, caffeine is demethylated to dimethylxanthines which undergo further enzymatic modifications to produce at the end glyoxalic acid, urea, formaldehyde and methanol. In bacteria like *Serratia marcescens, Klebsiella and Rhodococcus,* in addition to demethylation caffeine can also be converted to other metabolites by an oxidative route where caffeine is oxidized at the C-8 position resulting in the formation of 1,3,7-trimethyluric acid which can be further degraded to 3,6,8-trimethylallantoin which, on further degradation, forms dimethylurea (53). In certain species of bacteria like *Rhodococcus*, caffeine can also be directly oxidized to methyluric acid (51).

Finding efficient caffeine degrading enzymes can be a very valuable tool because they can be used to detoxify the caffeine rich substrates to then use another organism or set of organisms to degrade the rest of the compounds present in the waste of the coffee and tea industries processes or feed animals with the detoxified waste. A good idea can be treating the coffee pulp or husk waste with caffeine degrading enzymes and then add an organism that can use the rest of the compounds present in the waste as raw material for the production of value added product. A good example of this proposal can be to make an adaptation of the process

cited by Pandey et al., 2000. They tells us about a novel approach on value addition to coffee husk, using it as substrate for the production of aroma compounds for food industry application with yeast and fungi in solid state fermentation. They tell us that Soares et al., 1999 used the yeast *Pachysolen tannophilus* in solid state fermentation for synthesizing aroma compounds. In his experiments his results showed that using coffee husk and coffee husk extract was better than using steam treated coffee husk because he obtained a superior yield of aromatic compounds. The yeast culture produced a strong alcoholic aroma with fruity flavor useful in the food industry. Also the culture produced ethanol, as the major compound produced, acetaldehyde, ethylacetate, isobutanol, isobutyl acetate and ethyl-3-hexanoate and isoamyl acetate, giving a strong pineapple aroma (isobutanol and ethanol are examples of alcohols that can be used as biofuels if a different set of applications for the process was in mind). When leucine was added to the medium, a strong banana odour was found with increased amounts of isoamyl alcohol and isoamyl acetate (31). For example we can adapt that idea by first treating the waste with enzymes that degrade caffeine and then use a know fungi that has efficiency in producing alcohols and aromas and we can produce a biofuel or a food additive component. In case we want to treat directly the waste without decaffing it; we will have to select a host microbe that produces the value added products we are interested in, like ethanol and aromatic compounds; and by genetic engineering add the genetic information that enables the host cell to grow in the presence of caffeine dealing with its toxicity or using the caffeine as carbon source or nitrogen source as well as the other compounds present in the waste for the production of the components we are interested in producing. Another good reason to search for caffeine degrading enzymes can be to prepare a caffeine free hydrolysis of the coffee husk. Once the waste is free of caffeine or its concentration is reduced to a tolerable concentration to industrially useful known microbes; we can employ a method used by Urbaneja et al., 1996, where he used diluted sulphuric acid for hydrolyzing the coffee pulp and have a raw material useful for the production of a variety of products; for example ethanol. Urbaneja using diluted sulphuric acid for hydrolyzing the coffee pulp obtained: xylose, arabinose, fructose, glucose, sucrose and maltose. Arabinose was produced in highest concentration followed by glucose. The overall efficiency of the hydrolysis was 64 and 67% for total and reducing sugars respectively (31). And those carbon sources can be further processed into biofuels or added value chemicals or products depending on the product we want to obtain.

1.2.1.6.4 Tween 80 and Tween 20

The last carbon source we selected to test was tween. We are interested in tween degrading enzymes because a clone that grows on tween as carbon source points us toward the possibility of finding lipases or esterases. Tween 80 (polyoxyethylene sorbitan monooleate), mostly considered as a surfactant, can be used as a carbon source substrate for assay of lipases and esterases (54). I also elected to use Tween 20 for the preparation of the minimal media because of the similar structure it has with tween 80 to give the clones a broader selection for lipases/esterases activities if they were to be present. Esterases and lipases are enzymes with activity over lipids that are able to hydrolyze long chain and short chain acylglycerols (55). Lipases and esterases belong to the α/β hydrolase superfamily group. They share characteristic sequence motif GXSXG pentapeptides and a catalytic triad consisting of serine-histidine-aspartic or glutamic acid. Lipases are being used as catalyst to enhance chemical reactions related with the production of dairy products, pharmaceutical compounds, detergents, textiles, biodiesel, cosmetics, fine chemicals, agrochemicals, and polymeric materials (56). Most lipases and esterases are relatively small and most of them have activity over different substrates whithout the need of cofactors which increases the range of industrial applications that lipases have accounting for billions of dollars in revenue (57). There even exists a report that tells about lipases that can degrade caffeine and related methylxanthines demonstrating that these enzymes are promiscuous in that they have activity on wide-range of substrates (58). Another application for lipases that have received recent attention is the enzymatic transesterification of lipids for biodiesel production, because it produces a highly pure biodiesel without the presence of harsh contaminants and with the ease of separation of the byproduct of the enzymatic transesterification, glycerol; that being free of toxic contaminants can be also used for further processing to develop added value chemicals and products (59).

“Biodiesel is defined as the non-petroleum-based diesel fuel consisting of short chain alkyl (methyl or ethyl) esters, typically made by transesterification of vegetable oils or animal fats, which can be used (alone, or blended with conventional petroleum diesel) in unmodified diesel-engine vehicles” (59). Biodiesel have some major advantages. It is biodegradable and non toxic and it has a low emission profile which is environmentally beneficial. Nevertheless biodiesel have a major disanvantage. The cost for the production of biodiesel is high in comparison to the production of diesel which makes it hard for biodiesel to compete comertially

with diesel. There are four main ways to make biodiesel: direct use and blending, microemulsions, thermal cracking (pyrolysis) and transesterification (alcoholysis). The method mostly used by biodiesel production companies is the transesterification method in which oils or fat reacts with monohydric alcohol (an organic solvent) in the presence of a catalyst such as acid, base or lipases (59). The industrial process of biodiesel production is usually made by heating an excess of alcohol (usually methanol or ethanol) with vegetable oils under different conditions in the presence of an inorganic catalyst. “The most commonly used catalysts are alkali hydroxides (NaOH, KOH), carbonates and corresponding sodium and potassium alkoxides” (60). This industrial process have major disadvantages: 1.) the catalysts used for the reaction are homogenous and are dissolved in the reaction solution and for that reason are removed with the glycerol and cannot be recovered and reused, 2.) toxic waste are present in the production of the biodiesel, 3.) difficulties in the remove and purification of glycerol, 4.) partial saponification reaction which produces soap which lowers the yield of esters and makes the separation of esters and glycerol difficult, 5.) expensive cost of the required oil for the process; an oil with low free fatty acid content (inferior to 1%) (60). Enzymatic reactions involving lipases as catalyst in transesterification can be an excellent alternative to produce biodiesel as it produces high purity product and enables easy separation from the byproduct, glycerol; overcoming some of the above mentioned problems (59). And the product glycerol a carbon source in itself being uncontaminated with toxic substances can be used for further production of fine value chemicals by the activity of microbes on the substrate. But to overcome the cost problems of producing the lipases that will be used in the process we must find new highly active lipases that can perform in diverse conditions using fewer enzymes getting higher yield of production than the ones now known. For that main reason we are also interested in searching in the metagenomic libraries for lipases because they are part of the set of enzymes useful for the production of biofuels or value added chemicals.

1.3 Objectives

The possibility of finding new enzymes with better adaptability to established processes for the production of biofuels and value added chemicals increases greatly with the new access that metagenomic research gives us to the vast uncultured microbes population found in soil. There are reports of genes for the degradation of caffeine being cloned and expressed in *Escherichia coli* (61) demonstrating that *E. coli* as host cells can express this type of genes. There have been reports of genes that contain the information to produce enzymes related to the degradation of cellulose (38), (37), (36); lignin models (62), and lipids (30), (63), (64); being found by metagenomic studies demonstrating the potential of this technique. With the intention to capitalize on the latter, our main objectives for this research are posed as follows:

a. To develop screening assays to test a group of metagenomic libraries in search for clones that are able to grow using a test compound as carbon source in a minimal media.

b. To identify and isolate candidate clones, and assess the genomic information contained in their inserts.

c. To perform analyses *in silico* of the genomic information in the cloned fragment involved in the activity in order to propose genes that could be responsible for conferring the clone the ability to grow on the corresponding test compound.

2 Materials and Methods

2.1 Selection of the metagenomic libraries used.

Park et al., 2008 explains that there are two different approaches to screen for target gene containing clones: activity and sequence based screenings. We selected to use an activity based screening approach. Park continues by telling that in the case of activity based screening, several hundred thousand clones need to be analyzed in a single screen in order to detect a few functional active clones (27). For that reason we selected to test three metagenomic libraries developed from three different types of environments that contains a few hundred thousand clones to increase the chance of finding a positive hit for the degradation of one or all of the carbon sources we selected to test. We chose to use some of the metagenomic libraries that were developed by Dr. Carlos Ríos Velázquez Microbial Biotechnology Labroatory Team using samples from different environments of Puerto Rico: from El Yunque Rainy Forest; from Guanica Dry Forest and from Cabo Rojo hypersaline salterns Microbial Mats. From the Rainy Forest Metagenomic library's collection we selected 4 libraries pools: MM, PI, PII, PIII; adding up to 14,571 clones tested. From the Dry Forest Metagenomic library's collection we selected 6 libraries pools: $D1_3$, $D1_6$, $D1_{10}$, $D2_1$, DM; adding up to 87,617 clones tested. From the Microbial Mats Metagenomic library's collection we selected 24 libraries pools: CH1, CH2, CH3, CS1, CS2, CS3, CS4, CS5, CS6, CS7, CS8, FH1, FH2, FH3, FH4, FH5, FH6, FS1, FS2, FS3, FS4, FS5, FS6, FS7; adding up to 30,000 clones tested. At the end the final number of clones tested equals 132,188 clones.

2.2 Selection of the carbon sources

For the experiments associated to finding enzymes related to the degradation of lignin we selected two lignin degradation models to be tested, individually, as carbon sources in a minimal medium formulation: lignin alkali (Sigma Aldrich, Saint Louis MO) and veratryl alcohol (Sigma Aldrich, Saint Louis MO). For the experiments associated to finding enzymes related to the degradation of cellulose we used the soluble cellulose carboxymethyl cellulose (Sigma Aldrich, Saint Louis MO). For the experiments associated to finding enzymes related to the degradation of caffeine we used caffeine (Sigma Aldrich, Saint Louis MO). Finally, for the experiments associated to finding enzymes related to the degradation of lipids we decided to use a mixture

of Tween 80 (Acros Organics, New Jersey) and Tween 20 (USB Corporation, Cleveland OH) as carbon sources for the preparation of the minimal medium.

2.3 Culture media and host cells used.

We used two main growth media: Luria-Bertani (LB), a rich media and the M9 Salts Minimal Medium (M9). LB is a well-known standard laboratory media. The M9 medium is a rather specialized chemically defined medium, which we supplemented according to our culture conditions. For the preparation of M9, we used the following ingredients for 1L: 1X M9 salts (added as a 5X stock solution), 0.002 M $MgSO_4$, 0.0001 M $CaCl_2$, supplements, antibiotic, and a carbon source. The supplements are: 100 µg/ml l-leucine (added as a 10,000 µg/ml stock), and 10 µg/ml thiamine (added as a 500 µg/ml stock). We used 15 µg/ml chloramphenicol (added as a 1,500 µg/ml stock in absolute ethanol) as the selective antibiotic (Molecular Cloning, A Laboratory Manual 1st ed. Maniatis. 1982. P.68). The supplements and the antibiotic are needed specifically to cultivate the host selected for our study, which will be described below. The carbon sources were added as follows: soluble cellulose, lignin alkali and veratryl alcohol were added to a final concentration of 0.4% (w/v) from a 2% stock solution; caffeine was added to a final concentration of 0.1% to avoid toxicity levels; and Tween 20 and Tween 80 were added as a 1:1 mixture to a final concentration of 0.4% each. For the preparation of solid media agar was added to final concentration of 1.5% (w/v). In the preparation of the M9 medium all components were sterilized individually, using the autoclave or filtration as appropriate prior to be combined into the final preparation. The water (and the agar if a solid medium is desired) was adjusted to pH 7.0 using KOH and HCL depending on variations in pH levels observed in the solutions. The cells used as hosts for the production of the metagenomic libraries were the genetically engineered cell line EPI 300 (Epicentre Biotechnologies, Madison WI). This cell line has a genotype (*F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ- rpsL nupG trfA tonA dhfr*) incapable of growing in the absence of leucine and thiamine, for which we compensated by adding leucine to a final concentration of 100 µg/ml and thiamine to a final concentration of 10 µg/ml as we mentioned above. It is important to note that this cell line was also engineered to be used with copy control vectors. This required inserting the *trf A* gene into the *E. coli* genome. The company did this insertion using an EZ-Tn5 Transposome which carried both the *trf A* gene and the Dihydrofolate reductase (DHFR) gene. Thus, these cells will grow in the presence of Trimethoprim. As a

result, trimethoprim cannot be used as carbon source and cannot be present in the media for any screening assay.

For transposon mutagenesis (which will be described in detail below) we prepared the LB and M9 media as decribed above but added Kanamycin B sulfate salt (Sigma Aldrich, Saint Louis MO) to a final concentration of 50 µg/ml instead of chloramphenicol as the antibiotic for the media. The kanamycin stock solution was made dissolving the compound in the powder form in water. A stock of 500 µg/ml kanamycin was used for M9 medium.

2.4 Growth of libraries, isolation and preservation of selected candidates

We grew our collection of libraries on the corresponding media at a temperature of 37 ^{0}C from 1 to 2 weeks for each of the carbon sources we tested. We used the following type of media preparations: LB + Chloramphenicol (liquid media), LB + Chloramphenicol + agar (solid media), LB + Kanamycin (liquid media), LB + Kanamycin + agar (solid media), M9 + Leucine/Thiamine + selected carbon source (solid media), M9 + Leucine/Thiamine/Chloramphenicol + selected carbon source (solid media), M9 + Dextrose + Leucine/Thiamine/Kanamycin (solid media), and M9 + Leucine/Thiamine/Kanamycin + selected carbon source (solid media). All experiments were made in triplicates and with one metagenomic library collection at a time.

To identify our candidate clones the metagenomic libraries were grown on the corresponding carbon source. To do this first, we used LB + chloramphenicol liquid media to grow the metagenomic library collection we have selected to test. Then, in sterile 1.5 ml-eppendorf tubes a volume of this culture was centrifuged. After centrifugation the supernatant was removed by decantation and the pellet of cells was washed by adding 1 ml of an isotonic solution (0.85% sterile saline) to remove traces of carbon sources from the rich media. We resuspended the cells gently, followed by centrifugation to recover the cells. This washing procedure was repeated twice for a total of three washes; and it was done for every library tested. After washing the cells, they were serially diluted (1/10; 1/100; 1/1000; 1/10,000) in the saline solution and 50 µl of each dilution was plated in triplicate on a solid media plate composed of M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + the carbon source. Then we incubated the plates at 37^{0}C for 1 week. At the end of the first week the plates were evaluated for growth. If no evidence of growth was observed, the plates were left in

the incubator for another week. At the end of this period the best growing colonies were isolated from the plates and grown on LB + 15 µg/ml chloramphenicol to be able to identify if there was any contaminant that grew on the minimal media, and to isolate and preserve the possible candidate having an insert that conferred the host cell line EPI 300 the ability to grow on a minimal media. After the isolated colony grew on a solid media of LB + 15 µg/ml chloramphenicol (approximately in 2 days at 37^0C), a pure colony was selected to be grown in liquid media of LB + 15 µg/ml chloramphenicol and 600 µl of the fresh culture was then preserved in a sterile eppendorf tube using 400 µl of a 50% sterile glycerol solution and stored in a -80^0C freezer. All the preserved clones were identified and curated in a collection to be able to use them in subsequent steps in this research and in future experiments. This procedure was followed for all the carbon sources tested, except for the minimal media used: i) for lignin alkali-degrading experiments we used M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.4% lignin alkali; ii) for veratryl alcohol-degrading experiments we used M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.4% veratryl alcohol; iii) for caffeine-degrading experiments we used of M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.1% caffeine; iv) for cellulose-degrading experiments we used M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.4% carboxymethyl cellulose; and v) for tween 20/80-degrading experiments we used M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.4% tween 20 + 0.4% tween 80.

2.5 Fosmid DNA Extraction

After having a collection of candidates preserved; the next step was to extract the fosmid DNA from each candidate to check if the isolated clones had an insert and to use the fosmid DNA from the ones with a proven insert for further processing. The first step we did for the extraction of Fosmid DNA was to grow the isolated clones in liquid rich media with the appropriate antibiotic (LB + 15 µg/ml chloramphenicol). Then, after the cells grew (between 12 to 24 hours), each individual clone culture was induced for the overproduction of fosmids using Epicentre 1000X Copy Control Induction Solution. After the cells were induced, another period of growth between 12 to 24 hours was needed for the amplification of copies of the fosmid inside each clone. At this point the cells were ready for the extraction of fosmid DNA. The protocol we used for extracting the fosmid DNA from the clones is an adaptation from Qiagen miniprep protocol and standard DNA extraction/precipitation procedures by Lynn Williamson & Heather Allen called Best Miniprep Protocol Ever (Handelsman Lab Manual, 2008).

2.6 Restriction Enzymes Digest

Once we extracted the DNA from the preserved clones and identified that the fosmid DNA from those clones had an insert by comparing it with the fosmid DNA from EPI 300 cell line with no environmental insert; we proceeded to perform a restriction enzyme digest with the restriction enzyme *Not* I. The vector used for the development of the metagenomic libraries was pCC1FOS Copy Control Vector from Epicentre. This vector has multiple sites which can be recognize by different restriction enzymes. *Not* I recognizes two sites in the "empty" pCC1FOS fosmid which correspond to the following palindromes: 5' …GC$^{\Diamond}$GGCC GC…3'; 3'…CG CCGG$^{\Diamond}$CG…5'. By analyzing the restriction pattern generated by the action of *Not* I we can assess if the fosmid from a particular clone is "empty" (i.e. contains only two restriction sites) or if the fosmid has an environmental insert (i.e. multiple restriction sites are observed). Moreover, the restriction analysis helps us determine whether an isolated clone from a library is the same clone as another one isolated from the same library or if they are different. For the restriction digest reaction we followed the manufacturer's instructions for *Not* I (Promega, Madison, WI). We heat inactivated the enzyme at 65 °C for 15 min, centrifuged each sample and proceeded to conduct agarose gel electrophoresis to separate and visualize the digestion products.

2.7 Retransformation

The purpose of retransformation was to demonstrate if a fosmid found in a clone isolated from one of the metagenomic library collections can be extracted and transferred to a new host cell and if so, to see if the new host cell can express the phenotype behavior seen on the isolate. The observed phenotype behavior or growth (in this case) in the presence of the corresponding carbon source not used by the host cell, is due to the genotype information contained in the environmental insert enclosed in the fosmid extracted from the isolated clone from the library. The process we selected to do this was to extract the fosmid DNA and transfer it into a new host cell by electroporation. The host cell we selected to do the transfer is the same host cell used for the development of the metagenomic libraries, EPI 300. But instead of using the EPI 300 TI resistant line (or the line resistant to the infection of the virus T1) we selected to use the cell line EPI 300 Electrocompetent *E. coli.* The electroporation procedure followed was according to the manufacturer's recommendations (TransforMax™ EPI300 ™ Electrocompetent E. coli; Epicentre Biotechnologies). After the retransformation procedure was finished, the retransformants made were preserved in the freezer at -80^{0}C as was done for the isolated

clones from the libraries and were grown in the corresponding minimal media to verify the expected phenotype corresponding to an environmental insert present in the fosmid transferred to the host cell.

2.8 Transposon Mutagenesis

Transposable elements or transposons can be described as discrete DNA segments that are able to move between different, non homologous, genomic loci. Transposition of an element is thus a recombination reaction involving three separate sites: the two transposon ends and the new target locus (65). One important characteristic of transposons is its ability to disrupt gene function by inserting in or near genes (66). Bacterial transposons utilize two major modes of transposition: non-replicative or ‘cut-and-paste' transposition and replicative transposition. In a non-replicative or ‘cut-and-paste' mechanism, the element is excised from its original location and inserted into the new target locus. In replicative transposition, the element is copied such that insertion of the element into the same DNA molecule leads to deletion and/or replicative inversion, while transposition from one circular molecule to another generates a cointegrate structure in which the donor and target backbones are joined by directly repeated copies of the transposon at each junction. This cointegrate may be subsequently resolved by recombination between the two copies of the element, giving rise to a restored donor molecule and a target molecule in which one copy of the transposon has inserted (65).

We chose to use the EZ-Tn5 <T7/KAN-2> Promoter Insertion Kit (Epicentre, Madison WI) to insert transposable elements in the isolated fosmids of our interest. This was done to be able to sequence the possible gene or genes responsible for the phenotype behavior of the clone in a more efficient manner. We selected this promoter insertion kit because: i) the kit contains a Tn5 transposon that works in a non-replicative or ‘cut-and-paste' mechanism in a random way and that allowed us to randomly insert a transposable element in a fosmid; ii) the kit contains a Tn5 transposase that carries out transposition without the need for host cell factors. This allowed us to carry out the reaction *in vitro*, and enabled us to insert the transposable element directly into the fosmid and not in the genome of our *Escherichia. coli* host cell line; iii) the kit contains an antibiotic resistant gene for kanamycin which enabled us to prepare selective growth media to differentiate a clone with a transposable element from any other clone; and iv) the transposable element from the kit contains mosaic ends with recognition

sequences useful for the development of forward and reverse primers that facilitated sequencing of the interrupted (mutagenized) gene.

The main objective of having a transposable element that disrupts the gene or genes in the fosmid involved in the degradation of the carbon source, was to take that fosmid with the inserted transposon via an *in vitro* reaction and electroporate it into a new *recA*$^{-}$ *E. coli* host cell, grow it again in M9 minimal media with the respective carbon source to verify if we have eliminated the initial phenotype observed in our original screen. We finally selected those that had a possible gene "knockout" to sequence. In theory if a clone with a transposon inserted in the fosmid was unable to grow was because the transposable element disrupted the gene itself or close genes needed for the phenotype behavior of the clone, preventing the clone to use the carbon source that initially was able to use. Once the cadidates were selected we used the primers recommended in the EZ-Tn5 <T7/KAN-2> kit to sequence the gene of interest evading in that way the need to sequence all the 40kb environmental insert present in each isolated fosmid. The protocol followed for the development of transposons was the one published by the manufacturer (*EZ-Tn5™ <KAN-2>Tnp Transposome™ Kit; Cat. No. TSM99K2*).

2.9 Sequencing

2.9.1 Using primers from the flanking sites

The developers of the metagenomic libraries we used; made the library using the fosmid pCC1FOS Copy Control Vector from Epicentre. This fosmid contains sites recognized by a specific set of forward and reverse primers at the corners of the site for the insertion of the DNA under study to ease the sequencing of the inserted segment of DNA. We sequenced the fosmids present in our collection of isolates from the metagenomic libraries growing experiments to see which genes were present at the flanking sites of the fosmids and if those genes were related to the degradation of the carbon source in which the clone from the metagenomic library was grown. The primers used are the following (CopyControl™ Fosmid Library Production Kit with pCC1FOS™ Vector; Epicentre):

a. pCC1/pEpiFOS Forward Sequencing Primer:
b. 5′ - GGATGTGCTGCAAGGCGATTAAGTTGG - 3′
c. pCC1/pEpiFOS Reverse Sequencing Primer:
d. 5′ - CTCGTATGTTGTGTGGAATTGTGAGC - 3′

2.9.2 Using primers from the transposon

We selected two clones from the collection of isolated from the metagenomic library to extract the fosmids and introduce in each individual fosmid transposable elements to make a collection of clones with transposable elements to try to hit the gene responsible for the phenotype behavior of the clone in the minimal media in the presence of the respective carbon source trying to create a gene "knockout" to be able to later sequence the gene the transposon was "disrupting". The primers used to sequence the genes that were being disrupted by the transposable element inserted were present in the kit for the transposon development and were the following:

2. Kan-2 FP-1 Forward primer:
    a. 5' - ACCTACAACAAAGCTCTCATCAACC - 3'
3. Kan-2 RP-1 Reverse primer:
    a. 5' - GCAATGTAACATCAGAGATTTTGAG - 3'

2.9.3 Using developed primers for Primer Walking

For the fosmid isolates in which we used transposable elements, we performed also primer walking to increase the number of bases sequenced to have a broader look of the genes that were disrupted by the transposons. For the generation of the new primers needed for the primer walking sequencing we used a free online program called Primer 3 (http://frodo.wi.mit.edu/).

2.10 Analysis in silico

2.10.1 Sequence analyses

DNA samples from the clones isolated and from the transposons selected were sent to sequence at the Wisconsin Biotechnology Center (http://www.biotech.wisc.edu/). We used a free download program called Chromas Lite 2.01 (http://www.softpedia.com/get/Science-CAD/Chromas-Lite.shtml) to visualize and read the sequences and the sequence chromatogram obtained from the sequencing facility.

2.10.2 Blast-Basic Local Alignment Search Tool

2.10.2.1 Nucleotide Blast

We used the online program Blast (http://blast.ncbi.nlm.nih.gov/) and applied for the sequences we obtained Nucleotide blast to see if the sequence has similarities with known sequences of known microorganisms or if the sequence looks to be from an uncultivable unknown microorganism.

2.10.2.2 Blast X

2.10.2.2.1 Non-redundant protein sequences (nr)

We used the online program Blast and applied for the sequences we obtained Blast x, using the option for Non redundant protein sequences (nr) to see if the sequence has similarities with known genes to compare the results with the enzymatic degradation behavior of the clone in the presence of the carbon source under study.

2.10.2.2.2 Metagenomic proteins (env_nr)

We used the online program Blast and applied for the sequences we obtained Blast x, using the option for Metagenomic proteins (env_nr) to see if the sequence has similarities with known metagenomic genes or sequences to compare the results with the enzymatic degradation behavior of the clone in the presence of the carbon source under study.

2.10.2.3 NCBI Open Reading Frame Finder

To make predictions of the direction for the gene transcription of the sequenced genes that called our attention; we used the free online program NCBI Orf finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). This was important to understand if a gene or group of genes were involved in the behavior or phenotype expression of the clone that contained a fosmid with a particular environmental insert.

3. Results:

3.1 Clones Isolated

The M9 medium with the addition of various carbon sources of interest constituted our main clone isolation instrument. We were able to isolate and preserve clones that grew in all respective carbon sources tested except for veratryl alcohol. Even when the isolated clones grew weakly in the minimal media, in comparison with their growth in LB (figure 1), we were successful in isolating candidate clones with sequencing data that indicated possible genes that could be correlated to the ability to use as carbon source the chemicals we tested: Carboxymethyl Cellulose, Lignin Alkali, Tween 20 and Tween 80, and Caffeine. As an example we showed in figure 2 the growth of clones in M9 + 100 µg/ml leucine + 10 µg/ml thiamine + 15 µg/ml chloramphenicol + 0.4% Carboxymethyl Cellulose as carbon source; from the Rainy Forest library subpool PIII.

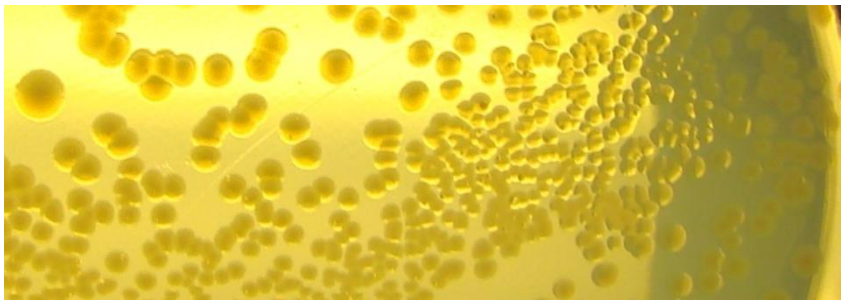


Figure 1: Rainy Forest library subpool PIII in rich media.
Clones of the metagenomic library PIII growing in a rich media (Lb + Chloramphenicol) portrays a healthy growth.

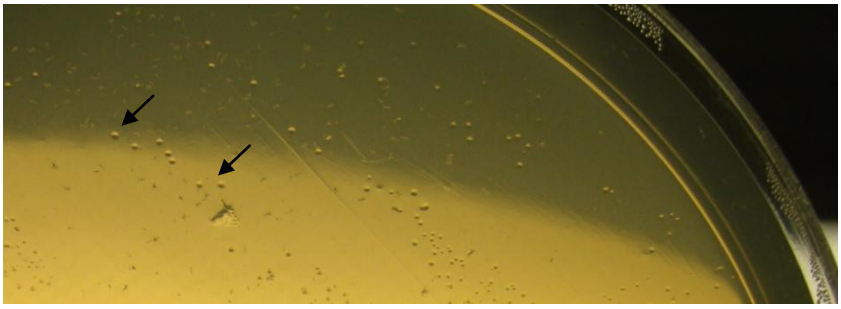


Figure 2: Rainy Forest library subpool PIII in M9 + cellulose minimal media

Clones of the metagenomic library PIII growing in chemically defined minimum media (M9) + carboxymethyl cellulose as sole carbon source. The blak arrows indicate growing colonies.

As can be observed in the image (figure 2), the growth on the minimal medium + carbon source is weak. Translucent colonies can be observed indicated by the black arrows in the image. All the colonies that grew in the minimal media for cellulose and lignin alkali as carbon sources had this similar appearance; translucent colonies presenting a weak growth. In contrast, for the case of caffeine as carbon source and Tween 80 and Tween 20 as carbon source a slightly different result was observed. If we consider the growth of the clones that grew on the minimal media with cellulose or lignin alkali as the default growth for these type of experiments; for caffeine the cellular growth was the weakest one and for Tween the cellular growth was the strongest one.

We isolated the largest colonies we observed in each plate and proceeded to grow them in LB rich media to preserve each single clone at -80^{0}C. After we did this for all the libraries and for all the chemicals we were testing; we ended with a list of possible degrading candidates. The list of isolated clones is as follows:

For the cellulose degradation experiments we isolated and preserved the following clones:

Table 1: Cellulose degrading candidates preserved

Clone ID	Library from where it was isolated
9	D2-1 (Dry Forest)
10	PI (Rainy Forest)
11	FH5 (Microbial Mats)
12	D1-12 (Dry Forest)
13	MM (Rainy Forest)
14	PIII (Rainy Forest)
15	FS3 (Microbial Mats)
16	CS2 (Microbial Mats)

Clones isolated from different metagenomic libraries that grew on M9 minimal media + Carboxymethyl cellulose as carbon source. These clones were preserved at -80^{0}C.

For the lignin degradation experiments we isolated and preserved the following clones:

Table 2: Lignin alkali degrading candidates preserved

Clone ID	Library from where it was isolated
1	Master pool (Microbial Mats)*
2	Master pool (Microbial Mats)*
3	Master pool (Microbial Mats)*
4	Master pool (Microbial Mats)*
5	Master pool (Microbial Mats)*
6	Master pool (Microbial Mats)*
7	Master pool (Microbial Mats)*
17	Master pool (Microbial Mats)*
18	CS5 (Microbial Mats)
19	FH2 (Microbial Mats)
20	FH1 (Microbial Mats)
21	CS6 (Microbial Mats)
22	CS3 (Microbial Mats)
23	CS4 (Microbial Mats)
31	Master pool (Microbial Mats)*
32	Master pool (Microbial Mats)*
33	Master pool (Microbial Mats)*
34	Master pool (Microbial Mats)*
35	Master pool (Microbial Mats)*
36	Master pool (Microbial Mats)*

Clones isolated from different metagenomic libraries that grew on M9 minimal media + Lignin alkali as carbon source. These clones were preserved at -80^{0}C.
*The Master pool is a combination of the 24 subpools of the Microbial Mats metagenomic libraries joined in one single tube for the experiment.

For the Tween 20 & Tween 80 degradation experiments we isolated and preserved the following clones:

Table 3: Tween degrading candidates preserved

Clone ID	Library from where it was isolated
A	FS6 (Microbial Mats)
B	CH3 (Microbial Mats)
C	CH2 (Microbial Mats)
D	FS2 (Microbial Mats)
E	CH1 (Microbial Mats)
F	CS4 (Microbial Mats)
G	FS6 (Microbial Mats)
H	CH1 (Microbial Mats)
I	CS5 (Microbial Mats)

Clones isolated from different metagenomic libraries that grew on M9 minimal media + Tween 20/80. These clones were preserved at -80^{0}C.

For the Caffeine degradation experiments we isolated and preserved the following clones:

Table 4: Caffeine degrading candidates preserved

Clone ID	Library from where it was isolated
8	PIII (Rainy Forest)
25	D1-6 (Dry Forest)
26	D1-12 (Dry Forest)
27	D1-3 (Dry Forest)
28	D2-1 (Dry Forest)
29	D1-10 (Dry Forest)
30	DM (Dry Forest)

Clones isolated from different metagenomic libraries that grew on M9 minimal media + Caffeine. These clones were preserved at -80^{0}C.

After all these clones were isolated and preserved, DNA extraction was performed to identify that an insert was present comparing the fosmid DNA weight from the clone, with the λ ladder and with the fosmid without insert from the control cell line EPI 300 T1 resistant. The fosmid DNA from a clone should be over the largest λ ladder size band of 23,130 bp because it's approximate size is around 40,000 bp with the environmental insert. The fosmid DNA without the insert should be below the 23,130 bp band because it's approximate size is around 8,000 bp, thus it should be seen between the λ ladder size band of 9,416 bp and the size band of

6,557 bp. We found that all isolated listed had a fosmid insert. An example of a DNA extraction and gel electrophoresis done for this research project are shown below to illustrate what we have explained:

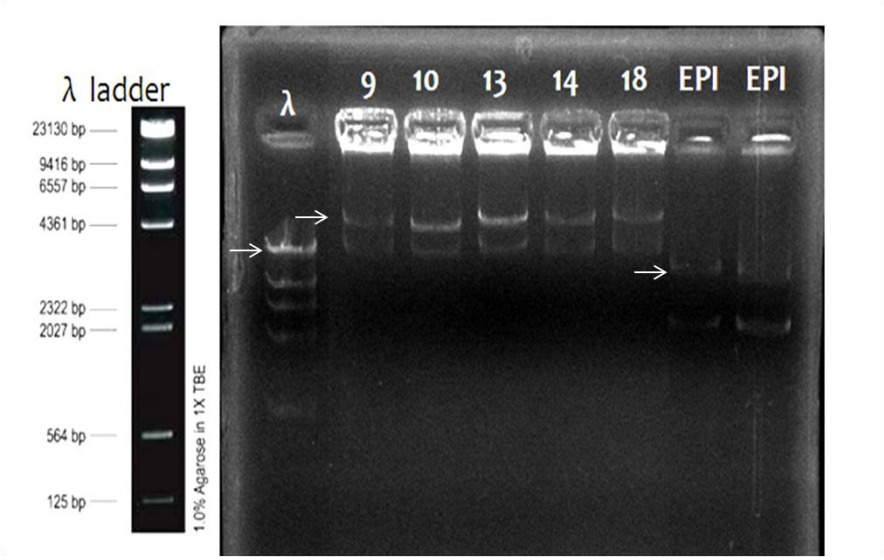


Figure 3: Example of an electrophoresis of DNA in a 1% gel of agarose

Electrophoresis of a DNA sample from some of the clones isolated from the different metagenomic libraries in a 1% agarose gel. The sample named EPI is the control of this experiment and describes the DNA of a fosmid without an insert. The samples 9, 10, 13, 14 and 18 are clones isolated from different metagenomic libraries that contain an environmental insert and as the white arrows indicate for them, the weight is over the taller band of the lambda ladder confirming the presence of an insert. The white arrow below the control DNA (EPI) indicates that the DNA without insert is below the taller band of the lambda ladder confirming the absence of an insert.

3.2 Retransformation

After we knew that all the clones we have isolated had an insert, the next step was to assess the insert stability by performing a process of re-transformation. We selected various clones from the list shown before to do retransformation by extracting the fosmid of the clone and by electroporation transforming an electrocompetent EPI 300. The clones selected and transformed are listed below:

Table 5: Retransformation of isolated clones were developed and preserved

Clone ID	Retransformant ID	Library from where it was isolated	Carbon source tested in M9 minimal media
A	AS	FS6 (Microbial Mats)	Tween 20 & Tween 80
C	CS	CH2 (Microbial Mats)	Tween 20 & Tween 80
D	DS	FS2 (Microbial Mats)	Tween 20 & Tween 80
H	HS	CH1 (Microbial Mats)	Tween 20 & Tween 80

13	13S	MM (Rainy Forest)	Carboxymethyl Cellulose
14	14S	PIII (Rainy Forest)	Carboxymethyl Cellulose
25	25S	D1-6 (Dry Forest)	Caffeine
28	28S	D2-1 (Dry Forest)	Caffeine
18	18S	CS5 (Microbial Mats)	Lignin Alkali
19	19S	FH2 (Microbial Mats)	Lignin Alkali
20	20S	FH1 (Microbial Mats)	Lignin Alkali
21	21S	CS6 (Microbial Mats)	Lignin Alkali
22	22S	CS3 (Microbial Mats)	Lignin Alkali
23	23S	CS4 (Microbial Mats)	Lignin Alkali

This is a list of retransformants developed by inserting the fosmid extracted from a clone isolated from a metagenomic library into an electrocompetent cell. For example we extracted the fosmid contained in clone 23 and inserted it in a new electrocompetent EPI 300 *Escherichia coli* obtaining the retransformant 23S. Once each of the retransformants were made, they were tested growing them in the same minimal media formulation used to isolate the original clone. The objective was to identify if the phenotype observed in a clone was passed to the electrocompetent host cell by retransformation which implies that the fosmid is giving the host cell the necessary "enzymatic machinery" to grow in that chemically defined media.

All the clones listed in table 5 were successfully retransformed. After the retransformation process was done, we again grew the retransformants in the minimal media to confirm if the phenotype we observed in the clones was replicated in the retransformants by transferring the fosmid or genotype to a new host cell via electroporation. As expected the phenotype was replicated (data not shown) and we proceed to preserve the transformants prepared.

3.3 Restriction Enzyme Digest

We regrew the retransformantss preserved in rich media and proceeded to perform a DNA extraction to have DNA material for a restriction enzyme digestion reaction with *Not* I. As we explained in the literature review the enzymatic reaction with *Not* I helped us identify differences between clones that came from the same library and that grew on the same substrate as carbon source by looking at the electrophoresis gel band pattern. This fingerprinting process was done to avoid sequencing the same clone more that once, which help us save materials and resources to further process different samples that presented different band patterns. Having different band patterns means that with a high probability the differences presented by the cut fosmid DNA of the isolated clones in a library are because the clones have different environmental inserts. An example of a restriction enzyme reaction electrophoresis gel follows:

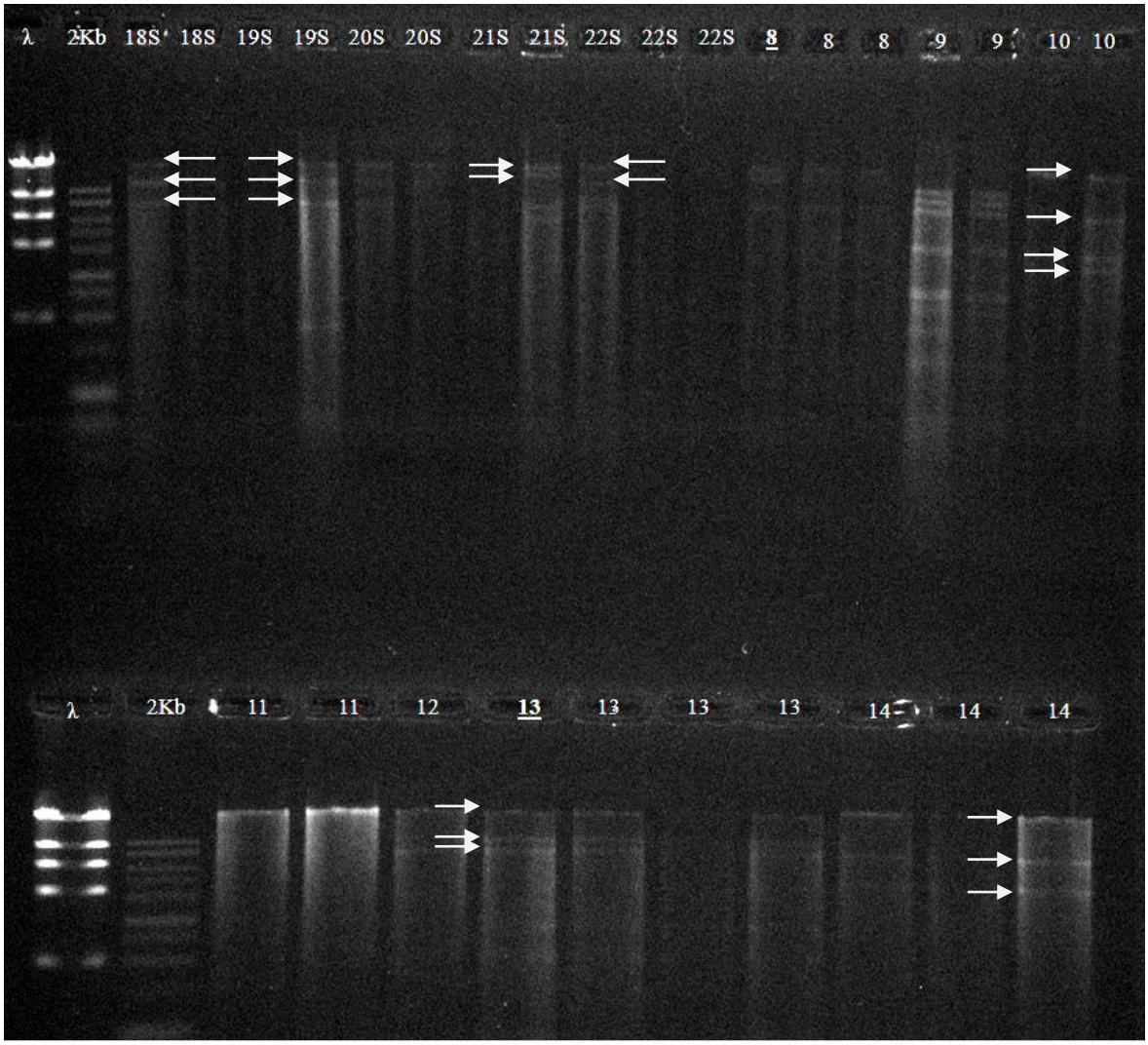


Figure 4: Example of a restriction enzyme reaction band pattern for the isolated clones in a 1% gel of agarose

Electrophoresis in a 1% agarose gel of a restriction enzyme reaction using *Not* I to react with fosmid DNA samples extracted from clones isolated from the metagenomic libraries. The white arrows indicate band patterns. We can observe to the superior part of the figure, to the left; that the arrows point toward similar band patterns from the 18S and 19S clones DNA samples. In the superior part of the figure to the center, we can observe that the arrows point toward different band patterns from the 21S and 22S DNA clone samples. Clones 18S, 19S, 21S and 22S are clones isolated from M9 minimal media + lignin alkali. By comparing their insert between each other we can assume that clone 18S and 19S have a high probability of having the same insert while clone 19S, 21S and 22S have different inserts. To the superior right of the figure and to the bottom of the figure we find clones 10, 13 and 14. These three clones were isolated from M9 minimal media + carboxymethyl cellulose. By comparing the band patterns between them we see that these clones probably contain a different insert that differentiates between them.

We can observe in the figure above the "fingerprinting" identification using *Not* I restriction enzyme digest reaction. For example, as the lignin degradation isolates table shows (Table 2); the clones 18, 19 and 21 in this image were isolated from the lignin alkali degradation experiments. If we compare the band pattern of clones 18 and 19 we see that the pattern is identical, possibly meaning that we isolated the same clone twice. However, if we compare the band patterns between clones 19 and 21, we see that the patterns between them are different,

discriminating in that way between two clones that have different environmental inserts. Another example that can be noted in figure 4 is for the candidates for cellulose degradation. As the cellulose degradation isolates table tells; the clones 10, 13 and 14 in this image were isolated from the cellulose degradation experiments. All the band patterns between the three of them are different, signaling us that we have three different environmental genetic insert present in three different clones.

After the restriction enzyme reaction experiments using *Not* I were finished and clones with interesting band patterns were identified, we selected the following clones for sequencing, using the primers developed to recognize the corners of the fosmid:

Table 6: Clones selected to be sequenced

Clone ID	Retransformant ID	Library from where it was isolated	Carbon source tested in M9 minimal media
A	AS	FS6 (Microbial Mats)	Tween 20 & Tween 80
C	CS	CH2 (Microbial Mats)	Tween 20 & Tween 80
D	DS	FS2 (Microbial Mats)	Tween 20 & Tween 80
H	HS	CH1 (Microbial Mats)	Tween 20 & Tween 80
13	13S	MM (Rainy Forest)	Carboxymethyl Cellulose
14	14S	PIII (Rainy Forest)	Carboxymethyl Cellulose
28	28S	D2-1 (Dry Forest)	Caffeine
18	18S	CS5 (Microbial Mats)	Lignin Alkali
19	19S	FH2 (Microbial Mats)	Lignin Alkali

List of clones selected to be sequenced. Here we have a list of representative clones that include examples from all the carbon sources investigated in the screening of the metagenomic libraries.

3.4 Transposon Mutagenesis

From this list of clones we selected two clones to do transposon mutagenesis; in our attempt to identify the genes responsible for the growth of the clones in the minimal media. We selected the clone 19 from the list of possible lignin alkali degraders and the clone 28 from the list of possible caffeine degraders. We developed transposon mutagenesis for each of the two selected clones (19 and 28) to be able to have a collection of transposons with a possible gene knockout contained in the collection. From that collection we send to sequencing the clones with the inserted transposon on the fosmid that did not grow on the minimal media expecting the lack of growth to be caused by a gene knockout in the gene related to the initial phenotype of the clone. The process for the isolation of those clones with a transposable element interfering a gene related to the growth of the clone in the minimal media can be better explained using a image. For the case of the clone 28 the explanatory image is as follows:

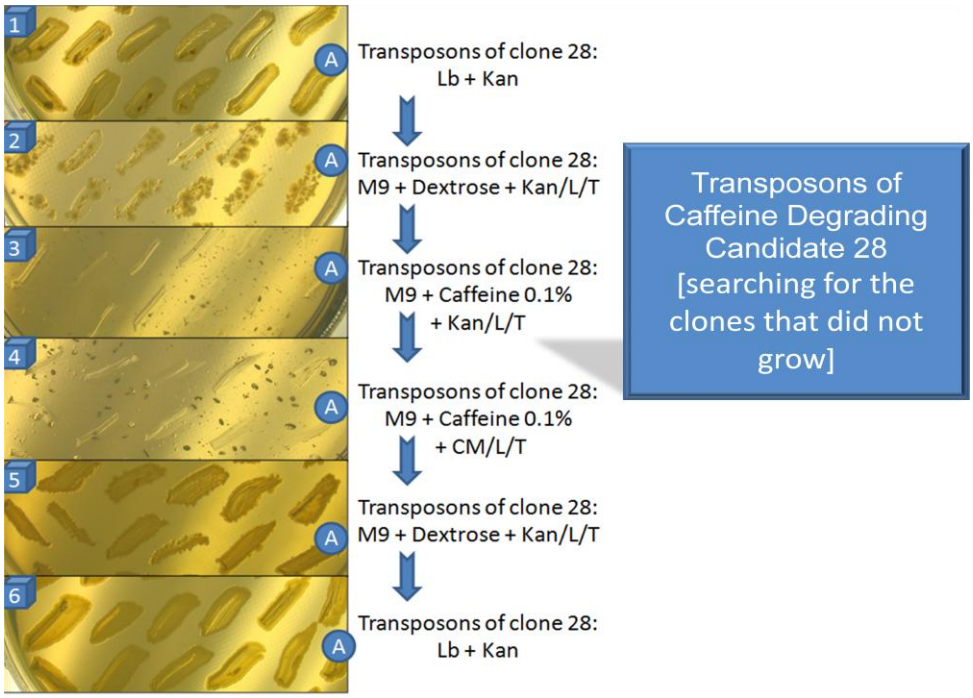


Figure 5: Example of a set of plates for transposons growing test for caffeine degrading candidate 28

These images give details about the screening procedure for caffeine degrading transposons with a possible gene knockout. The objective is to identify a clone with a transposable element interfering with the genes present in the fosmid environmental insert causing the clone to not grow under the same conditions that the original clone grew. In plate 1 and 6 we grew a list of clones with transposable elements inserted in the fosmid in presence of a rich media (Lb). Plate 2 and 5 were M9 minimal media but with a simple carbon source (dextrose). Plate 3 and 4 were M9 minimal media with caffeine as carbon source. Plates 3 and 4 were the ones we were looking to see absence of growth. The clones who didn't grew in the minimal media with the carbon source, in this case caffeine; were isolated from the rich media plates and selected to sequence because of the possibility that the absence of a phenotype replication was due to a gene knockout from the transposon that was inserted invitro.

Is important to mention that we used the same sterile stick to pick up healthy colonies from a LB rich media with kanamycin and that same stick was used for a single line in all of the plates in the depicted order to assure that if a clone did not grow it was not due to lack of inocula. We show in figure 5, 12 patches or lines in each plate. Each patch corresponds to a different clone that contains a different transposon insertion in the fosmid. To do this experiment in this way we used a sterile stick with fresh inocula, a fresh colony of the clone with the transposon insertion and we took that fresh colony with the sterile stick and passed it through the plate 1 to 6 only in the line or patch A. After this, the stick was discarded and a new sterile stick was used for a second patch or line. First we have to grow the clones with transposable elements inserted in a rich media like in the panel 1. All clones with transposable elements inserted grew on the rich media. Then, we grew the same clones with transposable elements inserted in a minimal media with dextrose, expecting them to grow as well. All the clones grew but the difficulty of the clone to grow in a minimal media in comparison to the growth in a rich media was evident. Then, the next step was to grow those clones in the minimal media with the carbon source under study; in this case with caffeine to a concentration of 0.1%. As the image demonstrates the clones did not grow, which is telling us that the transposon insertion is affecting the fosmid in a way that is disabling the cell to grow in the presence of this carbon source. We re-grew the clones in the minimal media with dextrose and back to a rich media in this case LB to make sure that the lack of growth we saw was caused by a reason different to lack of inocula. Because all the clones had the same position in all the plates, the clones that did not grow on the minimal media were isolated from the rich media and set apart for DNA extraction to sequence those clones because they had a better possibility to have a transposable elements inserted in the fosmid blocking the gene that was conferring the clone the phenotype expression we saw initially, namely, the ability of growing in caffeine as carbon source.

For in the case of the clone 19, one of our putative lignin alkali degrading candidates the explanatory image is as follows:

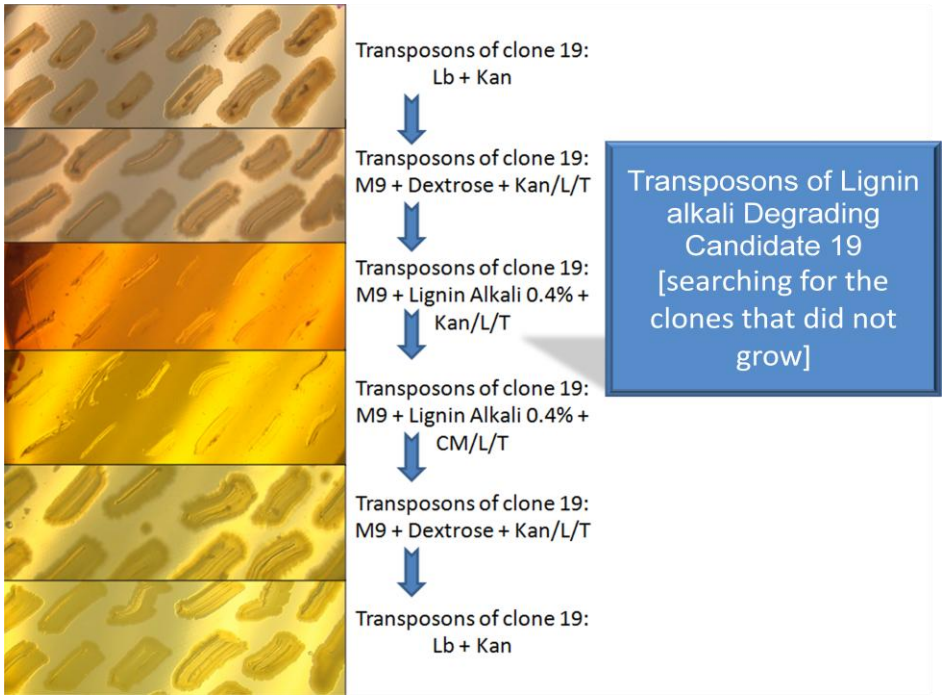


Figure 6: Example of a set of plates for transposons growing test for lignin alkali degrading candidate 19

These images give details about the screening procedure for lignin alkali degrading transposons with a possible gene knockout. The objective is to identify a clone with a transposable element interfering with the genes present in the fosmid environmental insert causing the clone to not grow under the same conditions that the original clone grew. In plate 1 and 6 we grew a list of clones with transposable elements inserted in the fosmid in presence of a rich media (Lb). Plate 2 and 5 were M9 minimal media but with a simple carbon source (dextrose). Plate 3 and 4 were M9 minimal media with lignin alkali as carbon source. Plates 3 and 4 were the ones we were looking to see absence of growth. The clones who didn't grew in the minimal media with the carbon source, in this case lignin alkali; were isolated from the rich media plates and selected to sequence because of the possibility that the absence of a phenotype replication was due to a gene knockout from the transposon that was inserted invitro.

The process was the same already described and for this case; we also separated the clones that did not grow in the minimal media for sequencing, expecting that the lack of growth was due to having a transposable elements inserted in the fosmid that blocks the gene that was conferring the clone the phenotype expression we observed in our initial screen.

We were able to produce a number of clones with different transposable elements inserted in different places in the fosmid DNA by following Epicentre protocol for the production of clones with transposons. We grew those clones in the format depicted in the image process we described above using a grid with 32 divisions per plate. From that number of clones produced we preserved the clones that did not grow on the minimal media. The following table contains the names of the transposons developed from the clone 28 and from the clone 19 that were preserved:

Table 7: Transposons for the clone 19 and for the clone 28

Clone Id	Transposon Id	Transposon Id	Clone Id	Transposon Id	Transposon Id
19	19 (1-3)	19 (2-1)	28	28 (1-3)	28 (5-16)
	19 (1-5)	19 (2-2)		28 (1-4)	28 (5-22)
	19 (1-8)*	19 (2-4)		28 (1-6)	28 (5-27)
	19 (1-9)	19 (2-5)		28 (1-16)	28 (6-1)
	19 (1-11)	19 (2-6)		28 (1-29)	28 (6-2)
	19 (1-12)	19 (2-7)		28 (2-2)	
	19 (1-13)	19 (2-8)		28 (2-10)	
	19 (1-14)	19 (2-9)		28 (2-11)	
	19 (1-15)	19 (2-10)		28 (2-16)	
	19 (1-16)	19 (2-11)		28 (2-19)	
	19 (1-19)	19 (2-12)		28 (2-26)	
	19 (1-20)	19 (2-13)		28 (3-7)	
	19 (1-23)	19 (2-14)		28 (3-15)	
	19 (1-24)	19 (2-15)		28 (3-22)	
	19 (1-25)	19 (2-16)		28 (4-1)	
		19 (2-17)		28 (4-7)	
		19 (2-20)		28 (4-9)	
		19 (2-23)		28 (4-18)	
		19 (2-24)		28 (5-11)	

List of transposons with a possible gene knockout that were preserved for the growing experiment presented in figure 5 and 6. The objective of sequencing a fosmid insert with a possible gene knockout was to obtain the sequence of the gene responsible for the phenotype presented by the clone without the gene knockout.
Leyend*: Clone Id (number of plate from which it was isolated-number of patch or line) Example 19(1-8).

After we finished the transposon mutagenesis experiments for the two clones, the clone 19 and the clone 28; we extracted DNA of all the clones with transposon that were isolated to prepare the DNA for sequencing and to perform a restriction enzyme digest to confirm the presence of an insert and to see the band pattern we already know and can identify as fingerprint for the clone. An example image of this is below.

DNA Extraction:

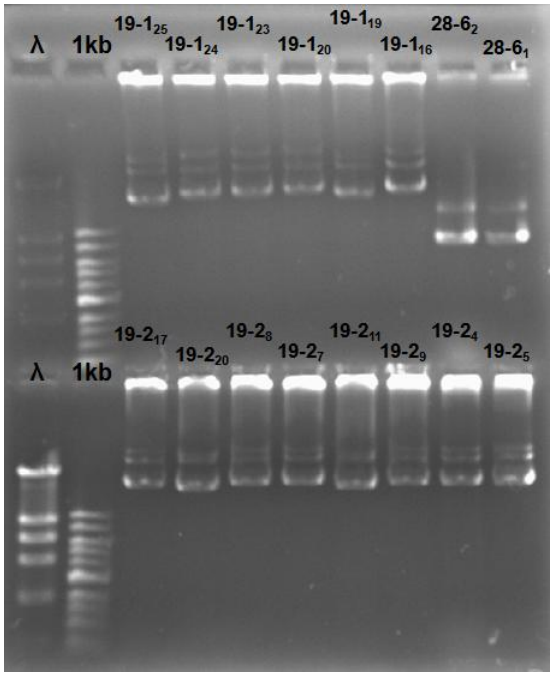


Figure 7: DNA extraction of the transposons isolated

Electrophoresis of Fosmid DNA with inserted transposable elements extracted from the different isolated and preserved transposons in a 1% agarose gel. The transformed cells were isolated and preserved for sequencing searching for the gene that was possibly interfered by the insertion of the transposable element that possibly caused the clone to loose its potential to grow in the minimal media with lignin alkali as the test compound.

Restriction Enzyme Reaction with *Not* I:

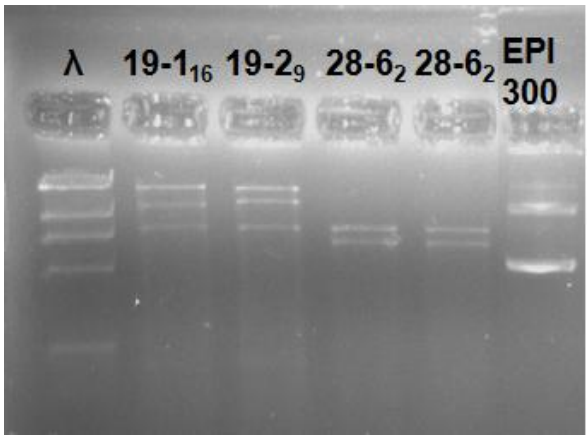


Figure 8: Restriction Enzyme Reaction with *Not* I of the DNA extracted from the isolated transposons

Electrophoresis of a DNA sample from transposons from clone 19 and clone 28 in a 1% agarose gel. This electrophoresis gel demonstrates the difference in the insert between these two clones, evidencing two different inserts.

After all this was done we ended with 2 final collections of clones to be sequenced: the list of retransformants and the list of clones with transposons. The list of retransformants were partially sequenced using the corners of the fosmid with the primers forward and reverse developed by Epicentre. The list of transposons were sequenced using the corners of the transposon with the primers forward and reverse developed by Epicentre and in addition using the corners of the fosmid as in the case of the retransformants. The next table has the list of transposons that were sequenced:

Table 8: Transposons for the clone 19 and for the clone 28 chosen to be sequenced

Clone Id	Transposon Id	Transposon Id	Clone Id	Transposon Id
19	19 (1-3)	19 (2-7)	28	28 (1-1)
	19 (1-5)	19 (2-8)		28 (1-2)
	19 (1-12)	19 (2-9)		28 (1-4)
		19 (2-10)		28 (1-6)
		19 (2-12)		28 (2-10)
		19 (2-16)		28 (2-11)
				28 (2-16)
				28 (2-19)

List of transposons with a possible gene knockout that were selected from the transposons preserved (Table 7) to send to sequence to identify the possible gene or genes responsible for the phenotype observed in the clone. Transposons 19(1-3), 19 (1-12); 19(2-9) had an almost unperceptive growth in the smear leaved by the stick used for the inoculation in the minimal media with lignin alkali as test compound, much less than the control growth observed for the isolated clone (Figure 14); while transposons 19(2-8), 19(2-12), 19(2-16) did not presented growth. Transposons 28 (1-6), 28(2-16) presented a an almost unperceptive growth in the smear leaved by the stick used for the inoculation in the minimal media with caffeine as test compound, much less than the control growth observed for the isolated clone (Figure 13); while transposons 28(1-1), 28(1-2), 28(1-4), 28(2-10), 28(2-11), 28(2-19) did not presented growth.

3.5 Sequencing

The sequencing results for our candidates with and without transposons inserted are summarized in the tables below. Some of the fosmid DNA samples with transposnons of the candidate 19 and 28 sent for sequencing resulted in the insertion of the transposon in the fosmid DNA and not in the environmental DNA. These results were not included in table 9 or 10.

Sequencing results for the transposons of the candidate for lignin alkali degradation; clone 19:

Table 9: Sequencing results for the clone 19

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (1-3)	Forward and Reverse sequences Joined	FH2	2,386 bp	Transcriptional Regulator Hdfr; Lys R
					Blast x non redundant results
					transcriptional regulatory protein [*Serratia odorifera* 4Rx13]
					Blast x metagenome results
					hypothetical protein GOS_4173434 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					FAD dependent oxidoreductase (-)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (1-12)	Forward and Reverse sequences Joined	FH2	2,389 bp	CHL I & Mg Chelatase CHL & Mg Chelatase 2
					Blast x non redundant results
					Mg chelatase subunit ChlI [*Serratia proteamaculans* 568]
					Blast x metagenome results
					hypothetical protein GOS_2981880 [marine metagenome]
					magnesium chelatase subunit ChlI [gut metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					Mg chelatase subunit ChlI [*Serratia proteamaculans* 568](Total score 863) (+)

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (2-8)	Forward and Reverse sequences Joined	FH2	2,381 bp	AsmA & EAL superfamily putative diguanylate phosphodiesterase
					Blast x non redundant results
					AsmA family protein [*Serratia sp*. AS12]
					Blast x metagenome results
					Hypothetical protein GOS_223740 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					AsmA family protein (-)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (2-9)	Forward and Reverse sequences Joined	FH2	2,339 bp	AdoMet-Mtases superfamily
					Blast x non redundant results
					tRNA (uracil-5-)-methyltransferase [*Serratia odorifera* DSM 4582]
					Blast x metagenome results
					hypothetical protein GOS_7232520 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					tRNA (uracil-5-)-methyltransferase [*Serratia odorifera* 4Rx13] (-)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (2-12)	Forward and Reverse sequences Joined	FH2	2,394 bp	DUF 4324 (DUF = Domain of unknown function)
					Blast x non redundant results
					Hyp. Prot. PROSTU_00109 [*Providencia stuartii* ATCC 25827]
					Blast x metagenome results
					hypothetical protein GOS_243598 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					hyp. Prot. PROSTU_00109 [*Providencia stuartii* ATCC 25827]; note = (DUF4324)

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (2-16)	Forward and Reverse sequences Joined	FH2	2,387 bp	Putative ATP dependent protease & Mg Chelatase superfamily 2 & DUF 413 superfamily & HTH superfamily & PBP2_LTTR_substrate superfamily; LysR
					Blast x non redundant results
					LysR family transcriptional regulator HdfR [*Serratia proteamaculans* 568]
					Blast x metagenome results
					hypothetical protein GOS_1913202 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					transcriptional regulator HdfR [*Serratia proteamaculans* 568] (+)
					uncharacterized DUF413 family protein [*Pantoea stewartii subsp stewartii* DC283] (-)
					Mg chelatase subunit ChlI [*Serratia proteamaculans* 568] (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	19 (2-16) and 19 (1-12)	Forward and Reverse sequences joined of the primer walking for the 19 (2-16) and 19 (1-12)	FH2	4,974 bp	CHL I & Mg Chelatase CHL & Mg Chelatase 2 & COG3085, Uncharacterized protein conserved in bacteria [Function unknown] & Bacterial regulatory helix-turn-helix protein, LysR family & PRK03601, transcriptional regulator HdfR; Provisional
					Blast x non redundant results
					Mg chelatase subunit ChlI [*Serratia proteamaculans* 568]
					Blast x metagenome results
					hypothetical protein GOS_2981880 [marine metagenome]
					Cont…

					Gene direction (+ or -) NCBI Orf Finder
					Mg chelatase subunit ChlI [*Serratia proteamaculans* 568] (+)
					transcriptional regulator HdfR [*Serratia proteamaculans* 568] (+)
					YifE like protein [*Serratia odorifera* DSM 4582] (-)
					protein of unknown function DUF413 [*Pantoea sp. aB*] (-)
					Blast n nucleotide collection results
					Serratia sp. AS13, complete genome
					Serratia sp. AS12, complete genome
					Serratia plymuthica AS9, complete genome
					Serratia proteamaculans 568, complete genome
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	No transposon	Forward	FH2	940 bp	Virulence factor BrKb (*Bordetella pertusis*)
					Blast x non redundant results
					Ribonuclease [*Serratia sp. AS12*]
					Blast x metagenome results
					hypothetical protein GOS_8480964 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					ribonuclease [*Serratia sp. AS12*] (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
19	No transposon	Reverse	FH2	382 bp	AdoMet_MTase Superfamily
					Blast x non redundant results
					tRNA (uracil-5-)-methyltransferase [*Serratia plymuthica* PRI-2C]
					Blast x metagenome results
					hypothetical protein GOS_1980219 [marine metagenome]; Cont…

					Gene direction (+ or -) NCBI Orf Finder
					tRNA (uracil-5-)-methyltransferase [*Serratia odorifera* 4Rx13] (+)

Sequencing results for the clone 19. We used different databases from Blast (Basic Local Alignment Search Tool) to identify and analyze the genes codifying for enzymes possibly responsible for the phenotype observed. "Putative conserved domains" are domains found in the sequence submitted to Blast; is the first result that appears. Blast x non redundant database (nr) compares the sequence submitted to blast with protein sequences found in GenBank CDS translations + PDB + SwissProt + PIR + PRF databases, excluding env_nr database. Blast x metagenome database (env_nr) is a database that compares the sequence submitted to blast with protein sequences isolated from environmental samples. The NCBI Orf Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames in a user's sequence or in a sequence already present in the database.

Sequencing results for the transposons of the candidate for caffeine degradation; clone 28:

Table 10: Sequencing results for the clone 28

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (4)	Forward and Reverse sequences Joined	D1-12	2,357 bp	Major Facilitator Superfamily (MFS); glycerol-3-phosphate transporter putative substrate translocation pore
					Blast x non redundant results
					glycerol-3-phosphate-transporter [*Desulfovibrio magneticus* RS-1]
					Blast x metagenome results
					hypothetical protein GOS_1092582 [marine metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					glycerol-3-phosphate-transporter [*Desulfovibrio magneticus* RS-1] (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (6)	Forward and Reverse sequences Joined	D1-12	2,257 bp	LbH_THP_succinylT active site; CoA binding site; substrate binding site; trimer interface.
					Zinc_peptidase_like superfamily; peptidase dimerisation Metal Binding sites
					Cont…

					Blast x non redundant results
					dapD gene product [*Candidatus Solibacter usitatus* Ellin6076]
					acetylornithine deacetylase/succinyldiaminopimelate desuccinylase-like deacylase [*Terriglobus roseus* DSM 18391]
					Blast x metagenome results
					2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase (dapD) [mine drainage metagenome]
					Peptidase dimerisation [mine drainage metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					acetylornithine deacetylase/succinyldiaminopimelate desuccinylase-like deacylase [*Terriglobus roseus* DSM 18391] (+)
					putative 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase [*Gemmatimonas aurantiaca* T-27] (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (10)	Forward and Reverse sequences Joined	D1-12	2,314 bp	Sigma-54 interaction domain HTH_8; Bacterial regulatory protein, Fis family AtoC; Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains UhpC: Sugar phosphate permease; glpT: glycerol-3-phosphate transporter
					Blast x non redundant results
					Fis family transcriptional regulator [*Desulfovibrio vulgaris* str. 'Miyazaki F'] Cont…

					Blast x metagenome results
					two component, sigma54 specific, transcriptional regulator, Fis family [gut metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					two component, sigma54 specific, transcriptional regulator, Fisfamily [*bacterium Ellin514*] (+) cont… glycerol-3-phosphate-transporter [*Desulfotomaculum reducens* MI-1] (-)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (11)	Forward and Reverse sequences Joined	D1-12	2, 348 bp	LacZ & aspartate-semialdehyde dehydrogenase & Amino Acid Kinases (AAK) superfamily
					Blast x non redundant results
					aspartokinase [*Gemmatimonas aurantiaca* T-27]
					hypothetical protein S18_873_0036 [uncultured Flavobacteriia bacterium]
					Blast x metagenome results
					Hyp. Prot. GOS_3366660 [marine metagenome]
					or bifunctional protein: aspartokinase I homoserine dehydrogenase [gut metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					Aspartokinase (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (16)	Forward and Reverse sequences Joined	D1-12	2, 320 bp	REC: Active site; dimerization interfase & REC: Signal receiver domain & Pseudouridine synthase
					Blast x non redundant results
					outermembrane protein [*Candidatus Methylomirabilis oxyfera*] Cont…

					truA gene product; tRNA pseudouridine synthase A [*Ignavibacterium album* JCM 16511]
					Blast x metagenome results
					hypothetical protein GOS_9564869 [marine metagenome]
					hypothetical protein GOS_8398108 [marine metagenome]
					or pseudouridylate synthase I [mine drainage metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					truA gene product; tRNA pseudouridine synthase A [*Ignavibacterium album* JCM 16511] (-)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (19)	Forward and Reverse sequences Joined	D1-12	2,324 bp	glycerol-3-phosphate dehydrogenase
					glpK: glycerol kinase
					DapB_C: Dihydrodipicolinate reductase, C-terminus
					DapB_N: Dihydrodipicolinate reductase, N-terminus
					Blast x non redundant results
					FAD dependent oxidoreductase [*Chthoniobacter flavus Ellin428*]
					or glycerol-3-phosphate dehydrogenase [*Terriglobus saanensis SP1PR4*]
					Blast x metagenome results
					glycerol-3-phosphate oxidase [mine drainage metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					glycerol-3-phosphate dehydrogenase [*Terriglobus saanensis* SP1PR4] (-)
					dihydrodipicolinate reductase [*Candidatus Solibacter usitatus* Ellin6076] (-)

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	28 (19)	Forward and Reverse sequences joined of the primer walking for the candidate 28 (19)	D1-12	3,178 bp	glycerol-3-phosphate dehydrogenase
					glpK: glycerol kinase DapB_C: Dihydrodipicolinate reductase, C-terminus DapB_N: Dihydrodipicolinate reductase, N-terminus
					ACT domain
					Amino Acid Kinase Superfamily (AAK)
					Blast x non redundant results
					glycerol-3-phosphate dehydrogenase [*Terriglobus roseus* DSM 18391]
					Blast x metagenome results
					glycerol-3-phosphate oxidase [mine drainage metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					FAD dependent oxidoreductase [*Verrucomicrobium spinosum* DSM 4136] (-) or glycerol-3-phosphate dehydrogenase [*Terriglobus roseus* DSM 18391] (-)
					aspartate kinase [*Candidatus Solibacter usitatus Ellin6076*] (-)
					dihydrodipicolinate reductase [*Candidatus Solibacter usitatus* Ellin6076] (-)
					Blast n nucleotide collection
					No significant similarity found

Cont…

Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	No transposon	Forward	D1-12	480 bp	aspartate-semialdehyde dehydrogenase
					Blast x non redundant results
					phage terminase large subunit [*Environmental Halophage* eHP-12]
					or dihydrodipicolinate synthetase [uncultured Flavobacteriia bacterium]
					aspartate-semialdehyde dehydrogenase [*Herpetosiphon aurantiacus* DSM 785]
					or aspartate semialdehyde dehydrogenase [*Candidatus Chloracidobacterium thermophilum* B]
					Blast x metagenome results
					hypothetical protein GOS_3928583 [marine metagenome]
					or Aspartate-semialdehyde dehydrogenase (ASA dehydrogenase) (ASADH) [mine drainage metagenome]
					Gene direction (+ or -) NCBI Orf Finder
					aspartate-semialdehyde dehydrogenase [*Herpetosiphon aurantiacus* DSM 785] (+)
Clone Id	Transposon Id	Primer	Library	Sequence analyzed	Putative Domains results
28	No transposon	Reverse	D1-12	528 bp	aspartate-semialdehyde dehydrogenase
					Blast x non redundant results
					phage terminase large subunit [*Environmental Halophage* eHP-12]
					or dihydrodipicolinate synthetase [uncultured Flavobacteriia bacterium]
					hypothetical protein CPAR2_210860 [*Candida parapsilosis*]
					Aspartate-semialdehyde dehydrogenase [*Synergistetes bacterium* SGP1]; Cont…

				Blast x metagenome results
				hypothetical protein GOS_3928583 [marine metagenome]
				or Aspartate-semialdehyde dehydrogenase (ASA dehydrogenase) (ASADH) [mine drainage metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				hypothetical protein S18_858_0001 [uncultured Sphingobacteria bacterium] (+)
				aspartate-semialdehyde dehydrogenase [Herpetosiphon aurantiacus DSM 785] (-)

Sequencing results for the clone 28. We used different databases from Blast (Basic Local Alignment Search Tool) to identify and analyze the genes codifying for enzymes possibly responsible for the phenotype observed. "Putative conserved domains" are domains found in the sequence submitted to Blast; is the first result that appears. Blast x non redundant database (nr) compares the sequence submitted to blast with protein sequences found in GenBank CDS translations + PDB + SwissProt + PIR + PRF databases, excluding env_nr database. Blast x metagenome database (env_nr) is a database that compares the sequence submitted to blast with protein sequences isolated from environmental samples. The NCBI Orf Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames in a user's sequence or in a sequence already present in the database.

Sequencing results for the candidate clones that were not submitted to transposon mutagenesis (i.e. sequenced from the corners of the fosmid).

Table 11: Sequencing results from the corners of the fosmid

Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
A	Forward	FS6	678 bp	Glutamine synthetase, catalytic domain
				Blast x non redundant results
				glutamine synthetase, type I [*Ammonifex degensii* KC4]
				Blast x metagenome results
				hypothetical protein GOS_3956351 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				glutamine synthetase, type I [*Ammonifex degensii* KC4] (-)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
A	Reverse	FS6	505 bp	Acetohydroxy acid isomeroreductase
				Cont…

				Blast x non redundant results
				ketol-acid reductoisomerase [*Pyrococcus sp. NA2*] cont…
				Blast x metagenome results
				hypothetical protein GOS_3632436 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				ketol-acid reductoisomerase [*Pyrococcus sp. NA2*] (-)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
C	Forward	CH2	682 bp	Glycosyltransferase_GTB_Type superfamily; CoA-disulfide reductase
				Blast x non redundant results
				CoA-disulfide reductase [*Thiocapsa marina* 5811] (note = Pyridine nucleotide-disulphide oxidoreductase)
				or FAD-dependent pyridine nucleotide-disulfide oxidoreductase [*Isosphaera pallida* ATCC 43644]
				Blast x metagenome results
				hypothetical protein GOS_1694740 [marine metagenome]
				or FAD-dependent pyridine nucleotide-disulfide oxidoreductase, partial [gut metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				FAD-dependent pyridine nucleotide-disulfide oxidoreductase [*Isosphaera pallida* ATCC 43644] (+)
				or CoA-disulfide reductase [Thiocapsa marina 5811] (+)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
C	Reverse	CH2	604 bp	No putative conserved domains have been detected
				Blast x non redundant results
				beta-glycosidase [*Terrabacter ginsenosidimutans*]
				Blast x metagenome results
				no significant similarities found
				Gene direction (+ or -) NCBI Orf Finder
				beta-glycosidase [*Terrabacter ginsenosidimutans*] (-)

Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
D	Forward	FS2	1,126	No putative conserved domains have been detected
				Blast x non redundant results
				hypothetical protein [*Sphaerochaeta pleomorpha str. Grapes*]
				Blast x metagenome results
				hypothetical protein GOS_9737587 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				No gene recognized
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
D	Reverse	FS2	466	No putative conserved domains have been detected
				Blast x non redundant results
				transglutaminase-like domain protein [uncultured *Leeuwenhoekiella sp*.]
				Blast x metagenome results
				no significant similarities found
				Gene direction (+ or -) NCBI Orf Finder
				No gene recognized
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
13	Forward	MM	666	Lyase class I_like superfamily
				aspartase (L-aspartate ammonia-lyase) and fumarase class II enzymes
				Blast x non redundant results
				fumarate hydratase, class II [*Candidatus Nitrospira defluvii*]
				or fumarate lyase [*Thermobaculum terrenum* ATCC BAA-798]
				Blast x metagenome results
				hypothetical protein GOS_3606144 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				fumarate hydratase, class II [*Candidatus Nitrospira defluvii*] (-)

Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
13	Reverse	MM	512	ACT_AHAS: N-terminal ACT domain of the Escherichia coli IlvH-like regulatory subunit of acetohydroxyacid synthase (AHAS).
				ALS_ss_C is the C-terminal half of a family of proteins which are the small subunits of acetolactate synthase.
				Blast x non redundant results
				acetolactate synthase small subunit [*planctomycete KSU-1*]
				Blast x metagenome results
				hypothetical protein GOS_5989491 [marine metagenome]
				or Acetolactate synthase small subunit [sediment metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				strongly similar to acetolactate synthase regulatory subunit [*Candidatus Kuenenia stuttgartiensis*] (+)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
14	Forward	PIII	751	No putative conserved domains have been detected
				Blast x non redundant results
				N-acetyltransferase GCN5 [*Arcobacter nitrofigilis* DSM 7299]
				Blast x metagenome results
				hypothetical protein GOS_9614338 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				N-acetyltransferase GCN5 [*Arcobacter nitrofigilis* DSM 7299] (-) N-Acyltransferase superfamily domain found
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
14	Reverse	PIII	1,173	No putative conserved domains have been detected
				Blast x non redundant results
				conserved hypothetical protein [*Streptomyces sviceus* ATCC 29083]
				Blast x metagenome results
				hypothetical protein GOS_9454055 [marine metagenome]
				hypothetical protein GOS_7661837 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				conserved hypothetical protein [Streptomyces sviceus ATCC 29083] (-)

Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
H	Forward	CH1	747	ALDH_F18-19_ProA-GPR; The aldehyde dehydrogenase superfamily (ALDH-SF) of NAD(P)+-dependent enzymes & ATP-dependent RNA helicase RhlE; Provisional
				Blast x non redundant results
				gamma-glutamyl phosphate reductase [*Melioribacter roseus* P3M]; isolation_source="microbial mat, developing on the wooden surface of a chute, under the flow of hot water coming from an oil exploring well"
				Blast x metagenome results
				hypothetical protein GOS_6863771, partial [marine metagenome]
				or Gamma-glutamyl phosphate reductase (GPR) (Glutamate-5-semialdehyde dehydrogenase) [mine drainage metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				gamma-glutamyl phosphate reductase [*Melioribacter roseus* P3M] (-)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
H	Reverse	CH1	917	P-loop containing Nucleoside Triphosphate Hydrolases; Ferrous iron transport protein B (FeoB) family
				Blast x non redundant results
				ferrous iron transport protein B [*Thermodesulfatator indicus* DSM 15286]
				Blast x metagenome results
				hypothetical protein GOS_9655278, partial [marine metagenome]
				or ferrous iron transport protein B [gut metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				ferrous iron transport protein B [*Desulfarculus baarsii* DSM 2075] (+)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
18	Forward	CS5	1,180	No putative conserved domains have been detected
				Blast x non redundant results
				hypothetical protein HAL1_08410 [*Halomonas sp.* HAL1; isolation_source="soil of a gold mine"; note="high tolerance to arsenite"

cont….

				Blast x metagenome results
				hypothetical protein GOS_7096911 [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				hypothetical protein HAL1_08410 [*Halomonas sp. HAL1*] (-)
Clone Id	Primer	Library	Sequence analyzed	Putative Domains results
18	Reverse	CS5	1,105	Aconitase Superfamily
				Blast x non redundant results
				aconitate hydratase 1 [*Halomonas sp. HAL1*]
				Blast x metagenome results
				hypothetical protein GOS_3854902, partial [marine metagenome]
				Gene direction (+ or -) NCBI Orf Finder
				aconitate hydratase 1 [*Halomonas sp. HAL1*] (+)

Sequencing results for the clones A, C, D, H, 13, 14 and 18. We used different databases from Blast (Basic Local Alignment Search Tool) to identify and analyze the genes codifying for enzymes possibly responsible for the phenotype observed. "Putative conserved domains" are domains found in the sequence submitted to Blast; is the first result that appears. Blast x non redundant database (nr) compares the sequence submitted to blast with protein sequences found in GenBank CDS translations + PDB + SwissProt + PIR + PRF databases, excluding env_nr database. Blast x metagenome database (env_nr) is a database that compares the sequence submitted to blast with protein sequences isolated from environmental samples. The NCBI Orf Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames in a user's sequence or in a sequence already present in the database.

3.6 Genes of interest

The genes that encoded for activities of interest to the present study that were obtained from the sequencing results above are listed in table 12:

Table 12: Genes that called our attention

Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
19 (1-12)	Mg chelatase subunit ChlI [*Serratia proteamaculans 568*]	2,389 bp	64% (1528 bp)	0	Catalyses the insertion of Mg into protoporphyrin IX (Proto) being this simple reaction one of the most interesting and crucial steps in the (bacterio)-chlorophyll biosynthesys . (67)
19 (2-8)	AsmA family protein [*Serratia sp. AS12*]	2,381 bp	53%	0	Blast x note: "Uncharacterized protein involved in outer membrane biogenesis [Cell envelope biogenesis, outer membrane]

Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
19 (2-9)	tRNA (uracil-5-)- methyltransferase [*Serratia odorifera* DSM 4582]	2,339 bp	26%	1e-67	Blast x note: "S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase), class I; AdoMet-MTases are enzymes that use S-adenosyl-L-methionine (SAM or AdoMet) as a substrate for methyltransfer, creating the product S-adenosyl-L-homocysteine (AdoHcy). The enzyme tRNA (uracil(54)-C(5))-methyltransferase catalyses the reaction: S-adenosyl-L-methionine + uridine(54) in tRNA <=> S-adenosyl-L-homocysteine + 5-methyluridine(54) in tRNA
19 (2-12)	hyp. Prot. PROSTU_00109 [*Providencia stuartii* ATCC 25827];note = Domain of unknown function (DUF4324)	2,394 bp	22%	1e-82	Unknown Function
19 (2-16)	LysR family transcriptional regulator HdfR [*Serratia proteamaculans* 568]	2,387 bp	35%	3e-175	[Superfamily] cl11398: LysR-type transcriptional regulators (LTTRs) have diverse functional roles including amino acid biosynthesis, CO_2 fixation, antibiotic resistance, and degradation of aromatic compounds, oxidative stress responses, and nodule formation of nitrogen-fixing bacteria, synthesis of virulence factors, toxin production, attachment and secretion. (Cited from blast conserved domains on [lcl|18655]).
19 (2-16) & 19 (1-12)	Mg chelatase subunit ChlI [*Serratia proteamaculans 568*]	4,974 bp	30%	0	Mg-chelatase catalyses the insertion of Mg into protoporphyrin IX (Proto).
19 (forward)	Ribonuclease [*Serratia sp. AS12*]	940 bp	74%	8e-143	Blast x note: Source- host="rapeseed plant"; region- "inner membrane protein YhjD"; CDS- PFAM: Ribonuclease BN-related" The biosynthesis of a mature, functional tRNA requires a series of processing steps

					in which both 5'-leader and 3'-trailer sequences are removed. The 5'-leader sequence is removed by the universal endoribonuclease, RNase P (68). A single pathway for 3'-maturation does not exist. Rather, it appears to proceed in a stochastic manner such that any one of five exoribonucleases (RNase II, D, BN, T, or PH) may act to complete the 3'-maturation process (69) .
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
19 (reverse)	tRNA (uracil-5-)-methyltransferase [*Serratia plymuthica* PRI-2C]	382 bp	52%	2e-35	*Escherichia coli* tRNA (uracil-5-)-methyltransferase (RUMT) catalyzes the S-adenosylmethionine (AdoMet)-dependent methylation of a specific Urd residue to form the m^5U residue found in the T-loop of most prokaryotic and eukaryotic tRNA. (70)
28 (4)	glycerol-3-phosphate transporter [*Desulfovibrio magneticus* RS-1]	2,357 bp	38%	3e-100	Catalysis of the transfer of glycerol-3-phosphate from one side of the membrane to the other. Glycerol-3-phosphate is a phosphoric monoester of glycerol. (http://www.uniprot.org/uniprot/C4XNQ8). In *E. coli*, G3P serves both as a carbon and energy source and as a precursor for phospholipid biosynthesis. GlpT is an organic phosphate/ inorganic phosphate (Pi) antiporter that functions for G3P uptake and is driven by a Pi gradient. In reconstituted systems, this transporter can also mediate Pi/Pi exchange. (71)
28 (6)	dapD gene product [*Candidatus Solibacter usitatus* Ellin6076] 2,3,4,5-tetrahydropyridine-2,6-carboxylate N-succinyltransferase	2,257 bp	32%	7e-96	Blast x note: region_name="LbH_THP_succinylT"; note ="2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase (also called THP succinyltransferase) catalyzes the conversion of tetrahydrodipicolinate and succinyl-CoA to N-

					succinyltetrahydrodipicolinate and CoA. Also in the blast x note: CDS: note="catalyzes the formation of N-succinyl-2-amino-6-ketopimelate from succinyl-CoA and tetrahydrodipicolinate in the lysine biosynthetic pathway.
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
28 (6)	acetylornithine deacetylase/succinyldiaminopimelate desuccinylase-like deacylase [*Terriglobus roseus* DSM 18391]	2,257 bp	41%	1e-63	This family corresponds to several clans in the MEROPS database, including the MH clan, which contains 4 families (M18, M20, M28, M42). The peptidase M20 family includes carboxypeptidases such as the glutamate carboxypeptidase from Pseudomonas. Peptidase family M28 contains aminopeptidases and carboxypeptidases, and has co-catalytic zinc ions. However, several enzymes in this family utilize other first row transition metal ions such as cobalt and manganese. Peptidase families M18 and M42 contain metalloaminopeptidases. M18 is widely distributed in bacteria and eukaryotes. However, only yeast aminopeptidase I and mammalian aspartyl aminopeptidase have been characterized in detail. Some of M42 (also known as glutamyl aminopeptidase) enzymes exhibit aminopeptidase specificity while others also have acylaminoacylpeptidase activity (i.e. hydrolysis of acylated N-terminal residues).
28 (10)	Fis family transcriptional regulator [*Desulfovibrio vulgaris* str. 'Miyazaki F']	2,314 bp	25%	1e-39	Molecular function: ATP binding, DNA binding, nucleoside triphosphate activity, sequence specific DNA binding transcription factor activity, two component response regulator activity. http://www.uniprot.org/uniprot/E3IP06

Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
28 (11)	aspartokinase [*Gemmatimonas aurantiaca T-27*]	2,348 bp	17%	1e-15	The aspartate pathway uses L-aspartic acid as the precursor for the biosynthesis of the amino acids lysine, methionine, isoleucine, and threonine. This is an essential pathway in plants and microorganisms involving, as it does, one-fourth of the building block amino acids that are required for protein synthesis. In addition, there are several important metabolic intermediates and products from this pathway including diaminopimelic acid, a key component required for cross- linking in bacterial cell wall biosynthesis, and dipicolinic acid, important for sporulation in Gram-positive bacteria. (72) Taking *E. coli* as example the bifunctional enzymes aspartokinasehomoserine dehydrogenase I and II (AK-HDH I and AK-HDH II) catalyze a phosphorylation and then, after an intervening reduction catalyzed by a separate enzyme, a second reduction to produce the intermediate homoserine. A third monofunctional enzyme, aspartokinase III, also catalyzes the phosphorylation of aspartic acid that is the commitment step in this biosynthetic pathway. (72)
28 (16)	truA gene product [*Ignavibacterium album* JCM 16511]; tRNA pseudouridine synthase A	2,320 bp	25%	8e-50	Kyung-Seop Ahn and collegues found that in *Pseudomonas aeruginosa* the truA gene is required for the type III secretory gene expression. The type III secretion system of *P. aeruginosa* encodes about 20 proteins, including components of a secretory apparatus which is devoted to the direct translocation of effectors into the host cell cytoplasm, and four effector molecules, ExoS, -T, -U and -Y, which alter normal host cell processes. The same author also cites

					that since pseudouridination of tRNAs is known to be required for maturation of tRNAs from their precursors, aminoacylation and stabilization of the stem–loop structure through improved intramolecular base-pairing, their observations imply that truA controls tRNAs that are critical for the expression of type III genes or their regulators. (73)
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
28 (19)	FAD dependent oxidoreductase [*Chthoniobacter flavus* Ellin428]	2,324 bp	24%	4e-85	Uniprot KB general annotation: Belongs to the FAD-dependent glycerol-3-phosphate dehydrogenase family. Glycerol-3-phosphate dehydrogenase catalyses the reaction: sn-glycerol 3-phosphate + a quinone = glycerone phosphate + a quinol. (http://www.uniprot.org/uniprot/B4CXH1)
28 (19) Primer Walking II	glycerol-3-phosphate dehydrogenase [*Terriglobus roseus* DSM 18391]	3,178 bp	41%	4e-154	In biological systems, glycerol is metabolized to dihydroxyacetone phosphate (DHAP) by one of two routes: (i) phosphorylation by glycerol kinase and subsequent conversion of sn-glycerol-3-phosphate (G3P) into DHAP through G3P dehydrogenase (G3PDH) or (ii) oxidation by glycerol dehydrogenase to form dihydroxyacetone (DHA) (74).
28 (19) putative domain	Aspartate kinase [*Candidatus Solibacter usitatus* Ellin6076]	1,599 bp	42%	2e-56	Appears only in putative conserved domains or by cutting the sequence 3,178 bp sequence by the aprox 1500 bps and blast the second half of it again. This result was obtained and described for this clone's fosmid in the transposon $28_{(11)}$
28 (19) putative domain	Dihydrodipicolinate reductase [*Candidatus Solibacter usitatus* Ellin6076]	1,599 bp	40%	4e-54	Appears only in putative conserved domains or by cutting the sequence 3,178 bp sequence by the aprox 1500 bps and blast the second half of it again. Related to the aspartate pathway for the production of

					the aminoacid lysine. Dihydrodipicolinate synthase and dihydrodipicolinate reductase are two enzymes central to the diaminopimelate pathway for lysine biosynthesis (75). Dihydrodipicolinate reductase (DHPR), encoded by the dapB gene, catalyzes the pyridine nucleotide-dependent reduction of dihydrodipicolinic acid to form tetrahydrodipicolinic acid. The enzyme is a component of the biosynthetic pathway leading to mesodiaminopimelic acid (DAP) and L-lysine in bacteria. DAP is a component of the peptidoglycan layer of Gram-negative bacterial cell walls, and inhibition of its biosynthesis results in cell death, probably due to the instability of the peptidoglycan which suggest that DHPR is a target for the design of inhibitors which may exhibit antibiotic activity. (76) Sean R. A. Devenish and colleagues experimentations suggest that dihydrodipicolinate reductase can have also a previously enrecognized dehydratase activity. (75)
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
28 (19) putative domain	Glycerol Kinase	----------	----------	----------	Appears only in putative conserved domains. Glycerol Kinase is the rate-limiting enzyme in glycerol metabolism. Glycerol kinase catalyzes the Mg^{2+} -ATP- dependent phosphorilation of glycerol to glycerol -3-phosphate. (77)
28 (Forward)	aspartate-semialdehyde dehydrogenase [Herpetosiphon aurantiacus DSM 785]	480 bp	24%	4e-08	The aspartate biosynthetic pathway is responsible for the production of four essential amino acids (threonine, isoleucine, methionine and lysine) and important primary metabolites such as

					diaminopimelate (DAP) 4 in bacteria. The pathway is not found in mammals but is critical to the survival of bacteria and thus presents a potential target for novel antimicrobial agents. (78)
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
A (Forward)	glutamine synthetase, type I [*Ammonifex degensii* KC4]	678 bp	89%	4e-57	Also known as glutamate-ammonia ligase. Glutamine synthetase (GS) is a primary biological catalyst in the sense that it catalyzes the first step at which nitrogen is brought into cellular metabolism: glutamate + $NH4^+$ + ATP => glutamine + ADP + Pi. The product glutaminie is a source of nitrogen in the biosynthesis of many other metabolites. (79)
A (Reverse)	ketol-acid reductoisomerase [*Pyrococcus sp. NA2*]	505 bp	74%	4e-50	ketol-acid reductoisomerase (KARI, IlvCfrom E. coli; EC1.1.1.86) catalyzes the two step reaction from S-2-acetolactate (S-2-AL) to 2,3-dihydroxyisovalerate (DHIV), involving a $Mg2^+$ dependent alkyl migration followed by ketone reduction. (80)
C (Forward)	CoA-disulfide reductase [*Thiocapsa marina* 5811] (note = Pyridine nucleotide-disulphide oxidoreductase)	682 bp	58%	9e-46	CoA-disulfide reductase is an enzyme that catalyzes the chemical reaction 2 CoA + NAD(P)+ <=> CoA-disulfide + NAD(P)H + H^+ http:/ / enzyme. expasy. org/ EC/ 1. 8. 1. 14
C (Reverse)	beta-glycosidase [*Terrabacter ginsenosidimutans*]	604 bp	44%	4e-09	Hydrolases which attack glycosidic bonds in carbohydrates, glycoproteins and glycolipids. The glycosidases are not highly specific. Usually they distinguish only the type of bond, e.g. O- or N-glycosidic, and its configuration (alpha or beta). http://www.uniprot.org/keywords/KW-0326

Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
13S (Forward)	fumarate hydratase, class II [*Candidatus Nitrospira defluvii*]	666 bp	72%	4e-65	Fumarase catalyzes the reversible hydration of fumarate to L-malate and is a key enzyme in the tricarboxylic acid (TCA) cycle and in amino acid metabolism. Fumarase is also used for the industrial production of L-malate from the substrate fumarate. (81) Two classes of fumarases are known. Class I fumarases are composed of heat-labile, iron–sulfur (4Fe–4S) homodimeric enzymes that are only found in prokaryotes. Class II fumarases are made of thermostable homotetrameric enzymes found both in prokaryotes and in eukaryotic mitochondria, and belong to a superfamily that also includes aspartate-ammonia lyases, argininosuccinatases, d-crystallins and 3-carboxy-cis,cis-muconate lactonizing enzymes. All these enzymes participate in reactions in which fumarate is released from different substrates, ranging from adenylosuccinate to malate. (82)
13S (Reverse)	acetolactate synthase small subunit [*planctomycete KSU-1*]	512 bp	69%	6e-39	Acetolactate synthase (ALS) is an essential enzyme in plants and many microorganisms because it catalyses the first step in the biosynthesis of branched-chain amino acids. In some bacteria it also plays a catabolic role, supplying acetolactate for the butanediol fermentation. Many of the bacterial ALSs have been shown to be heterotetramers composed of two types of subunit, large and small. The role of the small subunit is not entirely clear and it may be that it is involved in more than one way. For the various *E. coli* isoforms it has been shown that this subunit affects sensitivity to branched-chain amino acids and specific

					activity, stability and the kinetic properties. (83) In the absence of the small subunits, the large catalytic subunits of acetolactate synthase enzymes exhibit only partial activity *in vitro* and are insensitive to product feedback inhibition. In the presence of their cognate small subunits, however, the holoenzymes gain full catalytic activity that is negatively regulated by the product valine. (84)
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
14S (Forward)	N-acetyltransferase GCN5 [*Arcobacter nitrofigilis* DSM 7299]	751 bp	53%	2e-34	GCN5-related N-acetyltransferases (GNAT), catalyze the transfer of the acetyl group from acetyl coenzyme A (AcCoA, the "donor") to a primary amine (the "acceptor"). (85)
H (Forward)	gamma-glutamyl phosphate reductase [*Melioribacter roseus* P3M] (isolation_source of hit = "microbial mat, developing on the wooden surface of a chute, under the flow of hot water coming from an oil exploring well")	747 bp	86%	3e-65	Gamma-glutamyl phosphate reductase is an enzyme that catalyzes the second step of proline biosynthesis, the reversible nicotinamide adenine dinucleotide phosphate (NADPH)-dependent reduction of L-gamma-glutamyl phosphate to phosphate and L-glutamate 5-semialdehyde (86).
H (Reverse)	ferrous iron transport protein B [*Thermodesulfatator indicus* DSM 15286]	917 bp	57%	3e-51	Many anaerobic or microaerophilic bacteria transport Fe2+ via Feo (''ferrous iron transport'') systems, sometimes accompanied by reduction of Fe^{3+} to Fe^{2+} through ferric reductase. Enterobacterial Feo systems are composed of three proteins: FeoA, a small, soluble SH3-domain protein probably located in the cytosol; FeoB, a large protein with a cytosolic N-terminal G-protein domain and a C-terminal integral inner-membrane domain containing two 'Gate' motifs, which likely functions as the Fe^{2+} permease; and FeoC,

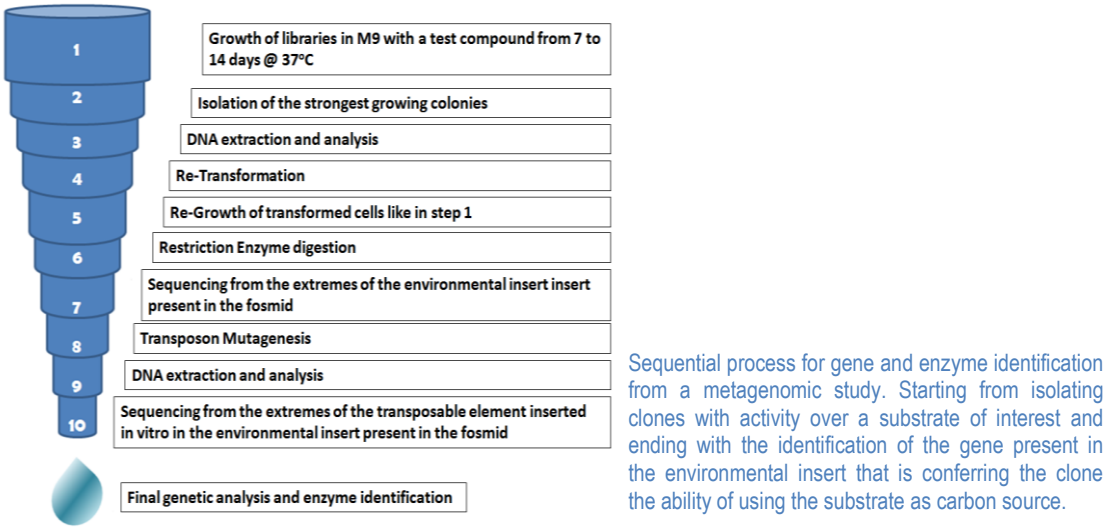
					a small protein apparently functioning as an [FeeS]-dependent transcriptional repressor. The feoABC genes constitute an operon. FeoB is responsible for ferrous iron transport, but the functions of FeoA and FeoC remain unclear. (87)
Clone Id	Gene Result obtained from blast x	Query	Query coverage	e value	Function
18 (Reverse)	aconitate hydratase 1 [*Halomonas sp. HAL1*]	1,105 bp	77%	1e-173	Aconitase is the second enzyme of the TCA cycle and catalyses the stereospecific and reversible isomerisation of citrate to isocitrate with cis-aconiate as intermediate. Aconitase is not only part of the TCA cycle, but also of the glyoxylate cycle, which serves as anaplerotic reaction during growth on acetate, ethanol or fatty acids, and of the methylcitrate cycle, which is responsible for the catabolism of propionate and odd-numbered fatty acids. In eukaryotes and some prokaryotes, aconitase also has a regulatory function by binding to certain mRNAs and inhibiting or increasing their translation. (88)

List of genes that called our attention. These genes were obtained by analyzing the sequencing data we have in the different databases contained in Blast (Basic Local Alignment Search Tool) and in the NCBI Orf Finder (Open Reading Frame Finder) tools. Each individual sequence submitted for analysis is termed query. The query coverage is the percent of the query sequence that overlaps the sequence of a gene or protein in the database. The e-value is a parameter that describes the number of hits one can 'expect' to see by chance when searching a database of a particular size. The lower or the closer to zero the e-value is, the more 'significant' the match is.

4. Discussion:

4.1 Funnel type process

We conducted our experimental design in a funnel-type design. Our goal was to obtain a large number of candidate clones to finally end up with a few interesting clones to focus on. It is important to mention that our intention was not to completely sequence all our candidate genes. The process we followed can be summarized in this way: 1.) Growth of clones in M9 medium *(132,188 clones tested)* plus the carbon source under study (lignin alkali or veratryl alcohol or caffeine or tween 80 and tween20); 2.) Isolation and preservation of the strongest growing clones present in the minimal media plate *(44 clones isolated)*; 3.) Extraction of the fosmid DNA of representative clones to check for the presence of a fosmid with an insert *(14 clones selected)*; 4.) Perform re-transformation from selected clones from each carbon source under study *(14 retransformants)*; 5.) Re-growth of the retransformants in their respective minimal media to confirm that the genotype that was giving the clone the observed phenotype could be replicated by transferring the fosmid to a new EPI 300 host cell; 6.) Perform a restriction enzyme digest to discriminate between different clones; 7.) Select from the set of representative clones to perform sequencing *(9 selected retransformants)*; 8.) Select from the final reduced group for sequencing from the corners of the fosmid, and a couple of clones for transposon mutagenesis to sequence them from the corners of the transposable elements inserted in the fosmid in vitro *(2 clones)*; 9.) Perform analisis “in silico” of the sequences obtained from the transposons and from the fosmids sequenced from the corners.



Sequential process for gene and enzyme identification from a metagenomic study. Starting from isolating clones with activity over a substrate of interest and ending with the identification of the gene present in the environmental insert that is conferring the clone the ability of using the substrate as carbon source.

Figure 9: Funnel type approach

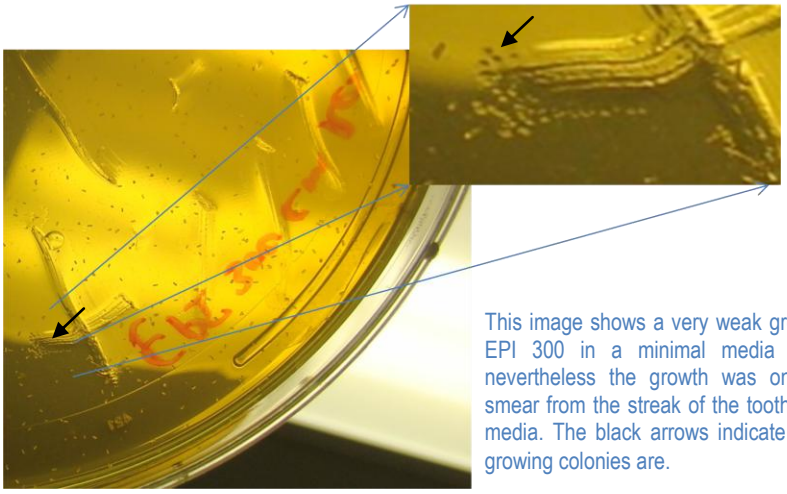
4.2 Isolation of clones discussion

At first the isolation of the list of clones we have was a challenge. We didn't know that the EPI 300 T1 resistant cell line we were using had a mutation that made it a cellular line auxotroph to Leucine and Thiamine. For that reason in the first experiments we didn't had growth because the clones could not grew in the absence of at least 100 ug/ml Leucine and 10 ug/ml Thiamine present in the minimal media. In addition to learn that the host cell used for the development of the metagenomic libraries we were testing needed Leucine and Thiamine for its growth, we also learnded that the EPI 300 T1 resistant cell line has a transposome mutation that carried into the cell the DHFR dene (Dihydrofolate reductase) which enables the cell to grow in the presence of Trimethoprim a compound that cannot be present in the minimal media of any degradation experiment that we wanted to run. We modify the minimal media adding the required concentration of Leucine and Thiamine and we were able to obtain growth. The growth was very weak but we were able to isolate 44 clones from the three types of metagenomic libraries we screened (8 possible cellulose degrading candidates, 20 possible lignin degrading candidates, 9 possible tween degrading candidates and 7 possible caffeine degrading candidates).

Having an additional carbon source in the minimal media (Leucine and Thiamine) even in small concentrations added a panorama that we didn't expected, it added to the equation the possibility that the growth of a clone was because of an effective use of the Leucine and/or Thiamine present in the media as carbon and nitrogen sources instead of the sole carbon source we were studying in the minimal media. On most of experiments the control EPI 300 without the insert didn't grew, however in some experiments it grew extremely weak and those who grew extremely weak only did it in the smear made by the tooth pick not in the rest of the plate; and for that reason we used the control weakest growth as example of a growth we didn't expect to see in a degrading clone.

In this image we cannot see any type of growth of the control EPI 300 in a minimal media with lignin alkali, not even in the smear from the streak of the toothpick used for the inoculatiuon process.

Figure 10: no growth of EPI 300 control growing on M9 media + Lignin Alkali



This image shows a very weak growth of the control EPI 300 in a minimal media with lignin alkali; nevertheless the growth was only present in the smear from the streak of the toothpick over the solid media. The black arrows indicate where the weakly growing colonies are.

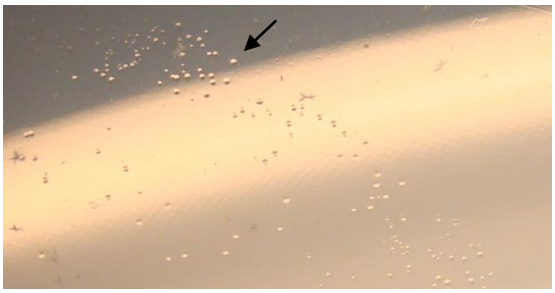
Figure 11: extremely week growth of EPI 300 control growing on M9 media + Lignin Alkali

Because of the change of the original minimal media recipe method when we discovered that we need to add Leucine and thiamine to the medium we decided to isolate from a plate only the strongest growing colonies present from the number of colonies that grew per plate. If a clone presented an extremely week growth like in figure 9 the clone was not selected. We wanted to isolate only degraders of the carbon source to be tested not degraders of the aminoacid on the media. Yet if the reason of the growth of the clone was the presence of a gene that enables the clone to use Thiamine and or Leucine effectively for its metabolic processes, it can be known once the insert of the clone is sequenced completely, or if in an assay with an isolated enzyme produced by the clone we observe activity over Leucine or Thiamine and not

over the carbon source we tested. Even degradation of Leucine or thiamine can be an interest finding because the host cell is auxotroph to those compounds and a strong growth will imply the presence of an insert with genes helping the microbe to grow under such harsh and difficult conditions.

We didn't obtained growth in the minimal media with the lignin degradation model veratryl alcohol as carbon source. We obtained growth for the rest of the chemicals tested; for carboxymethyl cellulose, for the lignin degradation model (lignin alkali) and for the mix of carbon sources tween 20 & tween 80. Below we present an example of an isolate from each chemical tested, growing in its particular minimal media:

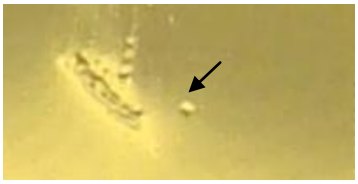
Possible Cellulose Degrading candidate # 10



In this image we can see a clone growing in a minimal media with cellulose. The black arrow indicates a clone that has the size to be isolated from the plate for further study.

Figure 12: Clone 10 in M9 minimal media + carboxymethyl cellulose + Chloramphenicol

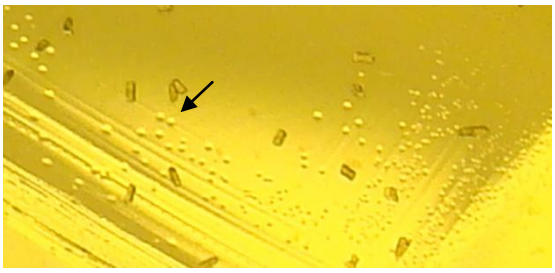
Possible Caffeine Degrading candidate # 28



In this image we can see a clone growing in a minimal media with caffeine. The black arrow indicates a single clone that grew in the media. Even when it is a weak growth, caffeine is known to be toxic to bacteria therefore we elected clone 28 for further study.

Figure 13: Clone 28 in M9 minimal media + caffeine + Chloramphenicol

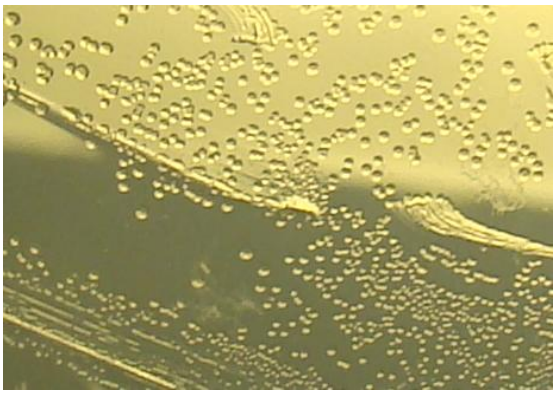
Possible Lignin Alkali Degrading candidate # 19



In this image we can see a clone growing in a minimal media with lignin alkali. If we compare the growth of this clone with the growth of EPI 300 in the same media (figure 10 and 11), we will see that clone 19 have a superior growth in presence of lignin alkali. The black arrow indicates the presence of colonies.

Figure 14: Clone 19 in M9 minimal media + lignin alkali + Chloramphenicol

Possible Tween 80 &20 Degrading candidate # C



In this image we can see a clone growing in a minimal media with tween 20/80. This was the strongest grow observed in the clones isolated. This image presents a very interesting possibility for this clone C to contain an insert that is conferring it the ability of using tween as carbon source.

Figure 15:Clone C in M9 minimal media + Tween 80 & Tween 20 + Chloramphenicol

All the colonies of the isolates had a weak growth and a translucid apareance. When they were transferred to a rich media the growth was strong but in the minimal media the growth was always weak, telling us that all the isolated clones had difficulties growing under these conditions. For the tween degrading candidates we had a special result. We had growth of the control EPI 300 T1 resistant cell line betther that the growth presented of figure 9. Yet EPI 300 T1 resistant cell line had a weak growth in comparison with the growth of the clones isolated from tween minimal media. For that reason even when the control cell line grew we again decided to use the EPI 300 weakest growth as example of the growth we will discard in the clone selection and isolate only the strongest growing colonies and preserved them for further analyses (taking the biggest colonies in the plate as the strongest). Below we present the difference of growth between the control EPI 300 T1 resistant cell line without an insert and a clone isolated from the metagenomic libraries under study.

Control: EPI 300 T1 Resistant cell line

Figure 16: EPI 300 T1 Res without the insert in M9 minimal media + Tween 80 + Tween 20 - Chloramphenicol

In this image we can see a week growth by EPI 300 in M9 + tween 20/80. The black arrow indicates the growing colonies.

Tween degrading candidate clone A

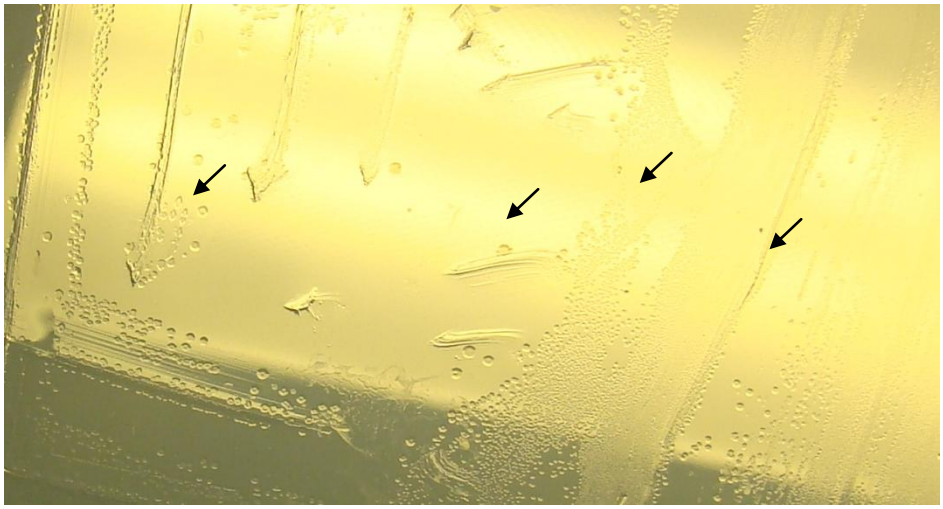


Figure 17: Clone A in M9 minimal media + Tween 80 + Tween 20 + Chloramphenicol

In this image we can see a strong growth by clone A in comparison with the growth of EPI 300 in the same media (figure 16). The black arrows indicate the growing colonies.

As can be seen the isolated clone A in figure 12 has an insert very interesting that is conferring the host *E. coli* EPI 300 T1 Resistant the ability of growing in the presence of Tween to a grade superior to the same cell line without an insert (figure 11).

After we isolated the clones that are tabulated in tables 1 to 4, we performed restriction enzyme digestion and retransformation. For resource disponibility we selected respresentative clones from each carbon source using the following criteria: 1.) a stronger growth than the other isolates; 2.) a different restriction enzyme pattern suggesting a different insert; 3.) representation from each substrate tested as carbon source. Once we had a list of clones we selected to sequence from the corners of the fosmid (table 6), we chose from that final list (also for resource disponibility) two clones to perform an additional step; transposon mutagenesis.

4.3 Transposon Mutagenesis discussion

We selected clone 19 that came from the Hypersaline Microbial Mat library; our lignin alkali degradation candidate, and clone 28 that came from the Dry Forest library; our caffeine degradation candidate, to do transposon mutagenesis. Our objective was to try to get by probability a transposon over or close to the gene responsible present in the environmental insert that is conferring the clone the ability of growing in the presence of caffeine (for the candidate 28) or lignin alkali (for the candidate 19) as sole carbon source. We used EPICENTRE Kit for in vitro transposon insertion and inserted the transposons on the fosmids contained by the clones we selected to use and then electroporate again the fosmids in a new EPI 300 host cell. We obtained a good number of transposons. The best approach to sequencing an unknown genetic insert of a big size of aproximete 40 kb like ours will be to sequence all the transposons developed and try to construct a map of the insert, localizing all the genes present in the insert, filling the gaps between genes by primer walking and then trying to adjudicate responsibility to one or a combination of genes by cloning only the segment with the gene or genes we think responsible for the phenotype in a new host cell and repeate the minimal media growing experiments searching for the phenotype behavior we saw in the clone witht the entire fosmid insert. Also another way of adjudicating the reason of the growth of the clone is to isolate the enzyme produced by the gene or group of genes we think responsible for the degradation of the carbon source and test the activity in vitro of the enzyme in the presence of the carbon source. Yet because of resource disponibility we decided to perform an additional growing experiment to send to sequence only those transposons that didn't grew in the minimal media, a phenotype search opposite to the phenotype expressed by the clone without the transposable element present on the insert in the fosmid. The rationale of this approach was that by choosing only the clones with inserted transposons that didn't grew, we will get a list of possibilities of having the transposon inserted and interfering or interrupting the genetic

transcription of the gene or group of genes responsible for the growth of the host cell in the minimal media with caffeine or lignin alkali as sole carbon source. Two main reasons for this rationale were: 1.) to increase the chance of finding the gene with a small number of sequencing steps (less primer walking) and 2.) to save resources by minimizing the number of fosmid DNA sent to sequence. So after doing this experiment we ended with the list presented in table 8. Therefore to summarize, at this moment we have two main collections that we selected to be sequenced, the retransformants of table 6 and the clones with transposons in their fosmid of table 8.

4.4 Discussing genes that called our attention

4.4.1 Lignin Alkali degrading candidate sequenced genes; clone 19.

For the lignin alkali degrading candidate we sequenced with primers that recognized the corners of the transposons and with primers that recognized the corners for the fosmid itself. We obtained the following hits:

Genes from the sequencing of the transposons after primer walking that appear to be close to one another in an unknown place in the approximately 40kb insert were:

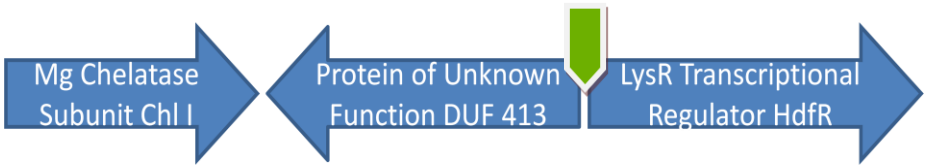


Figure 18: Genes direction in clone 19 insert.

In this image we can see the putative direction of the genes following the ORF analysis. The green insert is the localization where the transposon was inserted.

Genes from the sequencing of the transposons that were in an unknown place in the aproximalely 40kb insert:

Figure 19: Gene direction in clone 19 insert.

In this image we can see the putative direction of this gene following the ORF analysis. The green insert is the localization where the transposon was inserted.

Genes from the sequencing of the fosmid from the corners using the forward primer:

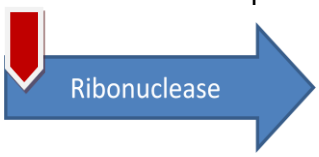


Figure 20: Gene direction in clone 19 inser forward primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the forward primer.

Genes from the sequencing of the fosmid from the corners using the reverse primer:

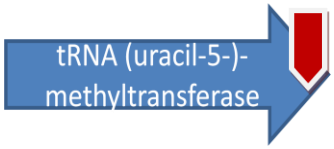


Figure 21: Gene direction in clone 19 inser reverse primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the reverse primer.

From these 6 genes 2 genes called our attention; the Magnesium Chelatase subunit I and the LysR Transcriptional regulator HdfR. The first hit analysed in blast x told us that the segment of a Magnesium Chelatase we sequenced resembles the Magnesium Chelatase subunit I (ChlI) of the microbe *Serratia proteamaculans.*

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
YP_001480983.1	Mg chelatase subunit ChlI [Serratia proteamaculans 568] >gb|ABV	787	787	30%	0.0	85%	G
YP_004503258.1	Mg chelatase subunit ChlI [Serratia sp. AS12] >ref|YP_004508210	781	781	30%	0.0	85%	G
EKF61881.1	competence protein ComM [Serratia plymuthica A30]	778	778	30%	0.0	85%	
ZP_06192842.1	Mg chelatase, subunit ChlI [Serratia odorifera 4Rx13] >gb|EFA148	777	777	30%	0.0	85%	

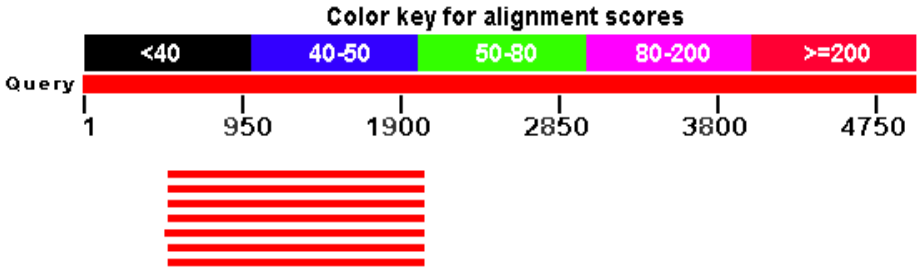


Figure 22: Blast x of the 4,974 bp sequence of clone 19.

Blast x hit that corresponds to a Magnesium Chelatase subunit I from *Serratia.*

The blast x hit covered 30% of the 4,974 bp we analyzed (query) with an e-value of 0. When we performed a blast n of the same sequence segment we obtained a hit that resembles the *Serratia AS13* as first hit, *Serratia AS12* as second, *Serratia plymuthica AS9* as third and

Serratia proteamaculans 568 as fourth hit. This is an interesting result because the metagenomic library from which the clone 19 was isolated was developed from the Microbial Mats of the hypersaline salterns of Cabo Rojo Puerto Rico using the anoxygenic phototropic (purple and green) bacteria layer (pink layer) and literature tells that Magnesium Chelatases had been found in purple non sulphur photosintetic bacteria like in *Rhodobacter capsulatus, Rhodobacter sphaeroides and Synechocystis* (89). This suggest that the insert that we have came from an anoxygenic phototropic bacteria that has the mechanism to produce its energy from a light harvesting complex, in other words from the light of the sun and that corresponds directly with the place where the DNA sample for the preparation of the library was isolated. Three main clases of chelatases are recognized: 1.) ATP dependent heterotrimeric chelatases like protoporphyrin IX magnesium chelatase; 2.) ATP independent chelatases like protoporphyrin IX ferrochelatase and 3.) multifunctional homodimeric chelatases like siroheme synthase (90). Our magnesium chelatase subunit falls in the category 1; an ATP dependent chelatase. Magnesium chelation requires 3 proteins, the subunits I, D and H found in plants and bacterio-chlorophyll producing prokaryotes (91). Our hit resembles the I subunit of bacteriochelatase. Researchers find Mg chelatase enzyme interesting because it lies at a branch point (see figure 21), for the biosynthesis of (bacterio)chlorophyll and haem. Magnesium chelatase is responsible for directing tetrapyrroles into the Mg branch of the pathway. (92) “Metal ion insertion steps are critical in the biosynthesis of the tetrapyrrole-derived pigments, which are involved in electron transport, light harvesting, oxygen transport and the assimilation of organic nitrogen and sulphur” (91).

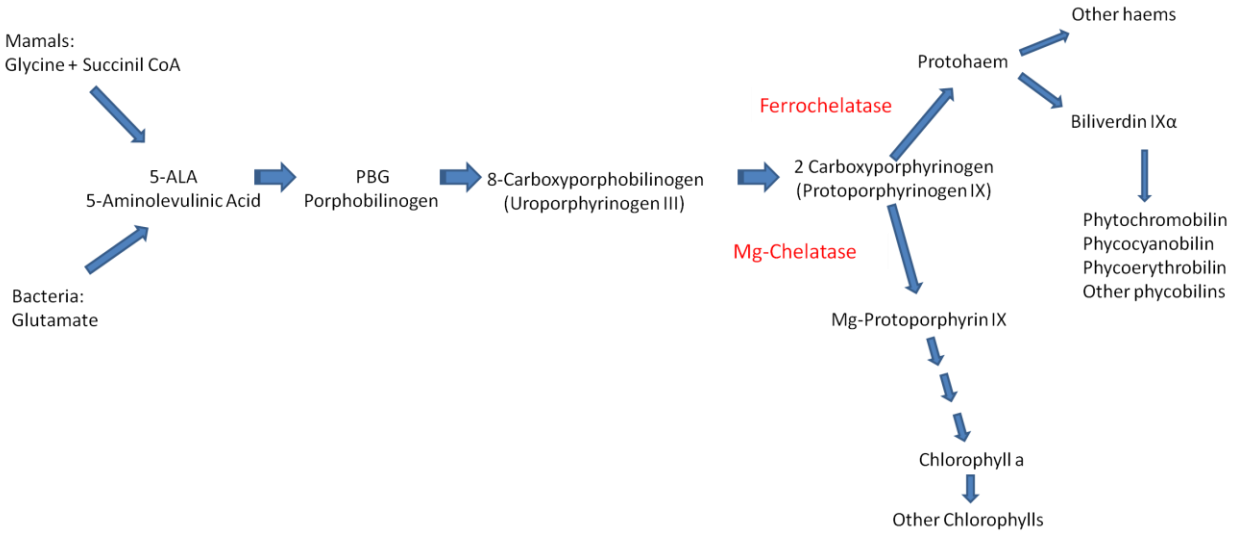


Figure 23: Position of Mg–chelatase at the branchpoint of haem and Bchl/Chl biosynthesis. {Ref (80); (81)}

Pathway for the biosynthesis of chlorophyll by the intervention of magnesium chelatase of haems by the intervention of ferrochelatase.

Magnesium chelatase is involved in light harvesting for energy production. Researchers have found that “under the light, the photosynthetic electron transport chains in thylakoid membranes employ diverse redox cofactors such as iron-sulfur clusters, quinines, and excitable systems in photosynthesis that can generate reactive oxygen species” (93). This is interesting because reactive oxygen species have been found to be part of lignin degradation by activating the production of lignin peroxidases. Is also interesting that our hit is closely resembles to *Serratia* genome, a microorganism that have been found to degrade lignocellulosic compounds like kraft pine lignin (94). Yet our result don’t appear to sustain that the growth of the clone 19 in the presence of lignin alkali was because Mg-chelatase subunit I acted in someway harvesting light and producing reactive oxygen species that acted upon the model for lignin degradation we were using (lignin alkali), because we only sequenced the Mg-chelatase subunit I and to have a full active magnesium chelatase we will need the other subunits being produced to act together over the substrate. In addition to that we found nothing in literature that directly relates magnesium chelatase or any of its subunits with the degradation of an aromatic compound or any lignin model.

The I and D subunits of magnesium chelatase have been shown to be members of the AAA^+ family of ATPases, and ATPases has a diverse set of cellular roles, including DNA replication, membrane fusion, microtubule processing and proteolisys (91). Therefore even when is not related to the degradation of lignin alkali, magnesium chelatase subunit I may have an important role in DNA replication of the genes close to it as part of the AAA^+ family of ATPases and the transposon that blocked it affected the DNA replication of the genes close to it explaining by that hypothesis why the clone with the inserted transposon didn’t grew.

The second gene that called our attention from these set of genes was the LysR transcriptional regulator HdfR. We obtained a hit for the LysR transcriptional regulator HdfR for the transposon 19 (2-16). The LysR transcriptional regulator HdfR we sequenced resembles the gene of an HdfR regulator in *Serratia proteamaculans.* The hit covered 35% of a 2,387 bp analyzed (querry) with an e-value of 3e-175. “The LysR-type transcriptional regulators (LTTRs) were first described by Heinkoff in 1988 and are present in diverse bacterial genera, archaea and algal chloroplast (95).” Blast conserved domains description for LTTRs transcriptional regulators tells us that they have a common helix-turn-helix (HTH) binding motif and that they perform diverse functional roles including amino acid biosynthesis, CO_2 fixation, antibiotic resistance, degradation of aromatic compounds, oxidative stress responses, nodule formation of

nitrogen-fixing bacteria, synthesis of virulence factors and toxin production. Individual family members of the LTTRs have specific names. Mark A. Schell talking about some LTTRs mentions a couple of examples like *IlvY, CatM, OxyR, NodDs* (96). *IlvY* can be found in *E. coli* and regulates the biosynthesis of leucine, Isoleucine and valine. *CatM* can be found in *Acinetobacter calcolaceticus* and regulates the catabolism of catechol, an aromatic compound. *OxyR* can be found in *E. coli* and regulates oxidative stress response. *NosDs* found for example in *Rhizobium* regulates the development of nitrogen fixation symbiosis. Literature confirms the diverse functions in which LTTRs are involved. When we obtained our LysR HdfR hit, we were expecting to cross link it to an LTTR related to the degradation of an aromatic compound. Searching trought the literature we found that LysR transcriptional regulator HdfR have different functions. Marcello Jakomin et al., 2008 suggests that HdfR directly or indirectly regulates the std fimbrial transcription operon in *Salmonella enterica* (97). Minsu Ko et al., 2000 presents that HdfR is a transcriptional regulator involved in the transcription of the flagelar master operons in *E. coli* (98). And the blast x information for the transcriptional regulator HdfR for Serratia proteamaculans 568 hit tells that it negatively regulates the transcription of the flagellar master operon *flhDC* by binding to the upstream region of the operons. In addition to that; Catherine A. Easom et al., 2012 found that HdfR can regulate up to 124 genes, including genes involved in aminoacid metabolism (arginine metabolism), hydroxyphenylacetate catabolism (compound with an aromatic ring) and pigment production in *Photorhabdus luminescens* (99). All this information is interesting because the LysR transcriptional regulator HdfR was close to a Mg chelatase who is related to pigment production and blast x told us that in *Serratia proteamaculans* it is related to a flagellum development. However the information presented by Catherine A. Easom et al., 2012 shows that it can be involved in much more. But to sustain that the HdfR gene is involved in the degradation of a compound with an aromatic ring like lignin alkali we will have to obtain the sequence of the rest of the genes present in the insert and see if there is one gene or a group of genes that produces an enzyme regulated by HdfR that can act over the lignin alkali in a catabolic way.

This LysR HdfR gene was close to the Mg-chelatase gene divided only by a protein of unknown function that also appeared in the magnesium chelatase transposon 19 (1-12), reason why we joined the sequences to finish with a sequence of 4,974 bp. By the Orf analysis we performed, we get to the conclusion that the Mg-chelatase and the LysR HdfR gene go in the same direction in the fosmid; toward the + direction and that the DUF gene go in the opposite direction.The LysR HdfR hit didn't apeared in the blast x results of the 4,974 bp we joined and

submitted to blast x (figure 20) where only the magnesium chelatase appears; but the LysR gene appeared in the putative conserved domains present in the sequence of clone 19 (2-16).

Figure 24: Putative conserved domains in 2,387 bp sequence from clone 19

Putative conserved domains from NCBI blast results. Here the green hit is a protein of unknown function, the blue hit is a HTH conserved motif and the red hit is a LysR transcriptional regulator. These results are from the clone 19 which was isolated from the lignin alkali degradation experiments.

Therefore when we took out the querry of 4,974 bp a segment of the left of the sequence where the magnesium chelatase subunit I gene hit stopped; ending with a sequence of 4,139 bp's long and resubmitted it to blast and the Mg chelatase, the LysR and the DUF genes appeared in the putative conserve domain but in the blast x hits only appeared the Mg chelatase and the Lys R genes:

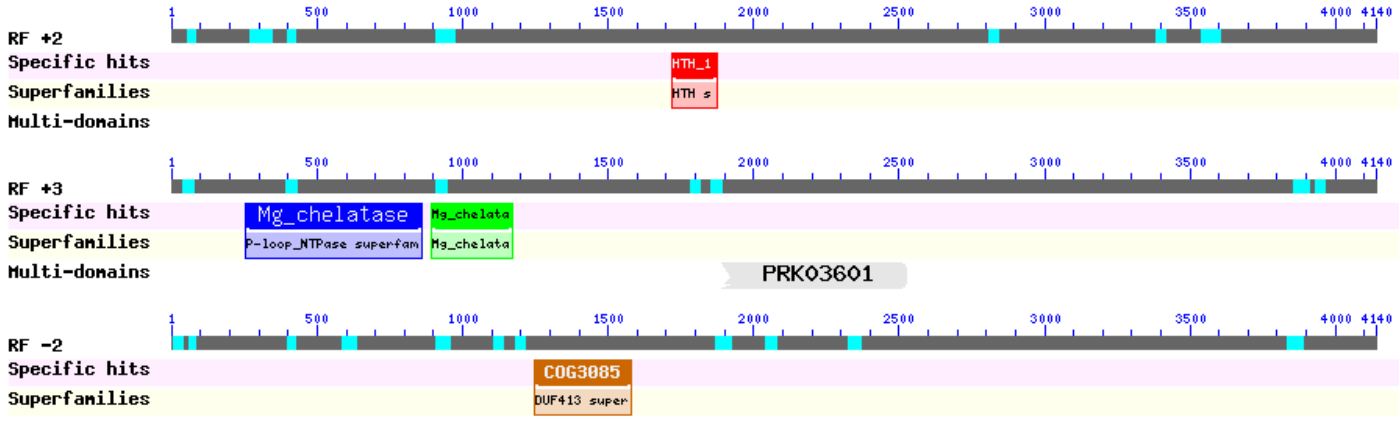


Figure 25: Putative conserved domains in 4,139 bp sequence from clone 19.

Putative conserved domains from NCBI blast results. Here the brown hit is a protein of unknown function DUF, the blue and green hits are related to a magnesium chelatase; the red hit is a HTH conserved motif and the greyhit is a LysR transcriptional regulator HdfR. These results are from the clone 19 which was isolated from the lignin alkali degradation experiments.

YP_001480983.1	Mg chelatase subunit ChlI [Serratia proteamaculans 568] >gb|ABV	590	590	29%	0.0	83%	
YP_004503258.1	Mg chelatase subunit ChlI [Serratia sp. AS12] >ref|YP_00450821C	584	584	29%	0.0	83%	
EKF61881.1	competence protein ComM [Serratia plymuthica A30]	581	581	29%	0.0	83%	
ZP_06192842.1	Mg chelatase, subunit ChlI [Serratia odorifera 4Rx13] >gb|EFA148	580	580	29%	0.0	83%	
YP_001480985.1	transcriptional regulator HdfR [Serratia proteamaculans 568] >gb|	411	532	20%	3e-141	93%	

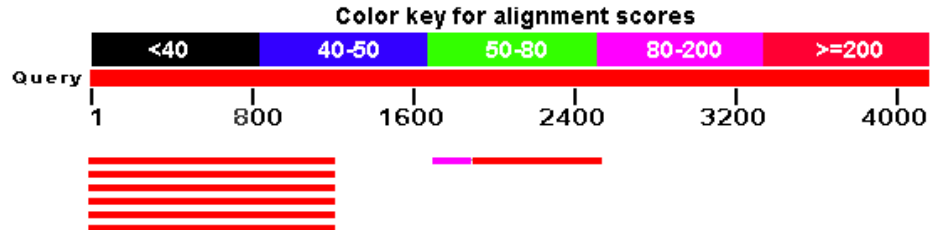


Figure 26: Blast x of the 4,139 bp sequence from clone 19.

Blast x hit that corresponds to a Magnesium Chelatase subunit I from *Serratia.* In this case the LysR transcriptional regulator not only appeared in the putative conserved domains but also in the blast x results.

After we found all this information we decided to search if the genome of *Serratia proteamaculans 568* and *Serratia AS13* was already sequenced and disponible online for study. We found it online and we searched for the genes we have isolated to see if we can get an idea of the rest of the possible genes in the insert that weren't sequenced. The idea was as follows: we have the result of the forward corner of the fosmid (a ribonuclease), we have the result of the reverse corner of the fosmid (a tRNA (uracil-5-)-methyltransferase), we have from the sequencing of the inserted transposons a Mg chelatase subunit I, a unknown gene (DUF 413 - domain of unknown function), a LysR HdfR gene and a AsmA family protein gene, we have the probable direction of the genes by an Orf analysis; so if we have the genome sequenced of the microorganism we must be able to identify the localization of the corners of the insert present in the fosmid and of the other genes in the middle of the insert, the direction of the genes should be the same and we must be able to get all those genes within an approximately 40kb segment of DNA and with that we can identify the other possible genes present in the insert that have not been sequenced. Identyfing those genes yet unsequenced can give a possible explanation of the phenotype observed in the growth of clone 19 in the M9 minimal media in which we grew it. We found searching in the Kegg database (http://www.genome.jp/kegg-bin/show_organism?org=spe) the sequenced genome of *Serratia proteamaculans 568.* In it we found the genes we have identified by sequencing, pointing to the same direction that our Orf analysis indicated us (Figure 25). In the genome the order between the genes Mg chelatase, LysR Hdfr and the domain of unknown function DUF described by the putative domain analysis was identical; supporting with evidence that our analysis was correctly made. The ribonuclease (our fosmid sequencing forward hit) in the upper corner of the fosmid was found to be localized

at the position 5,247,497…5,248,516 of the *Serratia proteamaculans 568* genome and the tRNA uracyl-5-methyl transferase (our fosmid sequencing reverse hit) was localized at the position 5,285,613…5,286,716. By substracting those values; (5,286,716 – 5,247,497) we get to an approximation of the size of the insert: 39,219 bp's which is in accordance with the expected approximated value for the inserts in each clone of the metagenomic library; 40 Kb. We present next an image that confine in red brackets the genes that possibly are present in the insert of clone 19:

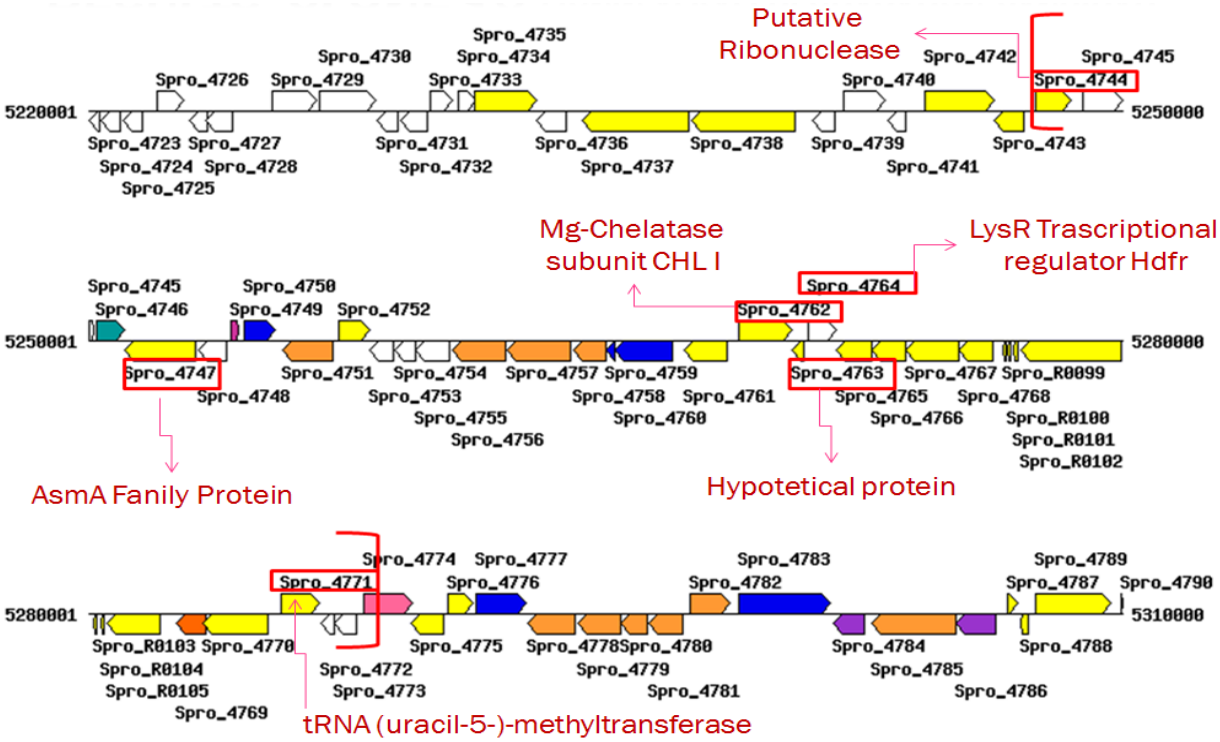


Figure 27: Segment of the Serratia proteamaculans 568 genome with our hits marked by red squares.

	Carbohydrate Metabolism		Metabolism of Terpenoids and Polyketides
	Energy Metabolism		Biosynthesis of Other Secondary Metabolites
	Lipid Metabolism		Xenobiotics Biodegradation and Metabolism
	Nucleotide Metabolism		Genetic Information Processing
	Amino Acid Metabolism		Environmental Information Processing
	Metabolism of Other Amino Acids		Cellular Processes
	Glycan Biosynthesis and Metabolism		Organismal Systems
	Metabolism of Cofactors and Vitamins		Unassigned

Figure 28: Color Codes for KEGG Pathway Categories

Segment from *Serratia proteamaculans* genome. The red squares are the genes we found by sequencing our transposons and the corners of the insert. The rest of the genes are genes that can possibly be found in the rest of the insert we haven't sequenced yet. The genes we sequenced are: Spro_4747 is an AsmA family protein; Spro_4762 is an Mg chelatase subunit Chl I; Spro_4763 is a domain of unknown function; Spro_4764 is a LysR transcriptional regulator HdfR; Spro_4744 is aputative ribonuclease and Spro_4771 is a tRNA (uracil-5-) -methyltransferase.

The genes that are possibly present in the insert that we just observed enclosed by red brackets in the figure 25 are listed in the next table (table 13).

Table 13: possible genes present in the fosmid of clone 19

Accession number	Name	Gene direction	Description
Spro_4744	putative ribonuclease; K07058 membrane protein	+	Protein family: Virul_fac_BrkB. Once predicted to be a tRNA processing enzyme RNase BN but found to be an incorrect prediction.
Spro_4745	major facilitator superfamily metabolite/H(+) symporter	+	Sugar transport protein
Spro_4746	K01521 CDP-diacylglycerol pyrophosphatase [EC:3.6.1.26]	+	Metabolism; Lipid Metabolism; Glycerophospholipid metabolism
Spro_4747	AsmA family protein; K07290 hypothetical protein	-	Involved in the assembly of outer membrane proteins in *E. coli.* May have a role in LPS biogenesis.
Spro_4748	EAL domain-containing protein	-	May have diguanylate phosphodiesterase function and is found in diverse bacterial signaling proteins.
Spro_4749	K01821 4-oxalocrotonate tautomerase [EC:5.3.2.-]	+	Present in diverse pathways: Benzoate degradation, Dioxin degradation, Xylene degradation. This enzyme catalyzes the conversion of 2-hydroxymuconate to 2-oxo-3-hexenedioate.
Spro_4750	ribokinase-like domain-containing protein; K00874 2-dehydro-3-deoxygluconokinase [EC:2.7.1.45]	+	Present in diverse pathways: Pentose phosphate pathway, pentose and glucuronate interconversions.
Spro_4751	K00053 ketol-acid reductoisomerase [EC:1.1.1.86]	-	Present in diverse pathways: Valine, leucine and isoleucine biosynthesis; Pantothenate and CoA biosynthesis. Related to amino acid biosynthesis.
Spro_4752	DNA-binding transcriptional regulator IlvY; K02521 LysR family transcriptional regulator, positive regulator for ilvC	+	LysR family transcriptional regulator, positive regulator for ilvC. "Another LTTR regulating amino acid biosynthetic genes is IlvY, which controls the divergently transcribed ilvC gene encoding acetohydroxy acid isomeroreductase, responsible for the second step in the common pathway for biosynthesis of isoleucine, valine, and leucine in *E. coli*." (96)
Spro_4753	hypothetical protein	-	Unknown function
Spro_4754	hypothetical protein	-	Unknown function
Spro_4755	hypothetical protein	-	Unknown function
Spro_4756	K01754 threonine dehydratase [EC:4.3.1.19]	-	Involved in glycine, serine and threonine metabolism and in valine, leucine and isoleucine biosynthesis.
Spro_4757	K01687 dihydroxy-acid dehydratase [EC:4.2.1.9]	-	Involved in valine, Leucine, isoleucine, phantotenate and CoA biosynthesis.
Spro_4758	K00826 branched-chain amino acid aminotransferase [EC:2.6.1.42]	-	Involved in valine, Leucine, isoleucine, phantotenate and CoA biosynthesis. Also involved in valine, Leucine, isoleucine degradation.

Spro_4759	ilvM; acetolactate synthase 2 regulatory subunit; K11258 acetolactate synthase II small subunit [EC:2.2.1.6]	-	Involved in panthothenate, CoA, valine, leucine and isoleucine, biosynthesis. Involved in the butanoate metabolism, in the C5-branched dibasic acid metabolism.
Spro_4760	acetolactate synthase 2 catalytic subunit; K01652 acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	-	Involved in panthothenate, CoA, valine, leucine and isoleucine, biosynthesis. Involved in the butanoate metabolism, in the C5-branched dibasic acid metabolism.
Spro_4761	major facilitator transporter; K08178 MFS transporter	-	Belongs to the SHS family, lactate transporter.
Spro_4762	Mg chelatase subunit ChlI; K07391 magnesium chelatase family protein	+	Subunit I of the Magnesium chelatase complex composed by subunit I, D, H.
Spro_4763	hypothetical protein; K09897 hypothetical protein	-	Unknown function
Spro_4764	transcriptional regulator HdfR	+	Function not completely known. Related to multiple activities: regulation of aminoacid metabolism, pigment production and flagellum development.
Spro_4765	inner membrane ABC transporter permease YjfF (EC:3.6.3.17); K02057 simple sugar transport system permease protein	-	Transport of simple sugars
Spro_4766	monosaccharide-transporting ATPase (EC:3.6.3.17); K02057 simple sugar transport system permease protein	-	Transport of simple sugars
Spro_4767	ABC transporter-like protein; K02056 simple sugar transport system ATP-binding protein [EC:3.6.3.17]	-	Transport of simple sugars
Spro_4768	periplasmic binding protein/LacI transcriptional regulator; K02058 simple sugar transport system substrate-binding protein	-	Regulator of the transport system
Spro_R0099	tRNA-Trp; K14235 tRNA Trp	-	Involved in aminoacyl-tRNA biosynthesis
Spro_R0100	tRNA-Asp; K14221 tRNA Asp	-	Involved in aminoacyl-tRNA biosynthesis
Spro_R0101	5S ribosomal RNA; K01985 5S ribosomal RNA	-	Involved in genetic information processing and translation. From the RNA family; non coding RNAs
Spro_R0102	23S ribosomal RNA; K01980 23S ribosomal RNA	-	Involved in genetic information processing and translation. From the RNA family; non coding RNAs
Spro_R0103	tRNA-Ala; K14218 tRNA Ala	-	Involved in genetic information processing; translation and in aminoacyl-tRNA biosynthesis
Spro_R0104	tRNA-Ile; K14227 tRNA Ile	-	Involved in genetic information processing; translation and in aminoacyl-tRNA biosynthesis
Spro_R0105	16S ribosomal RNA; K01977 16S ribosomal RNA	-	Involved in genetic information processing and translation. From the RNA family; non coding RNAs
Spro_4769	glutamate racemase (EC:5.1.1.3); K01776 glutamate racemase [EC:5.1.1.3]	-	Involved in aminoacid metabolism. D-glutamine and D-glutamate metabolism
Spro_4770	btuB; vitamin B12/cobalamin outer membrane transporter; K16092 vitamin B12 transporter	-	Involved in the translocation of vitamin B12 across the outer membrane to the periplasmic space.
Spro_4771	K00557 tRNA (uracil-5-)-methyltransferase [EC:2.1.1.35]	+	Catalyzes the formation of 5 methyl uridine at position 54 in all tRNAs.

Genes sequenced by us are shaded in red; the rest of the genes are possibly located in the rest of the insert that hasn't been sequenced yet.
The genes shaded in blue are genes related to the degradation and biosynthesys of aminoacids.

With all this information, how can we explain clone 19 growth? The genes shaded with violet in table 13 are the genes we found by sequencing. From those genes we found no gene related to the phenotype we were observing; the growth of clone 19 in the M9 media containing lignin alkali as carbon source. Yet from the genes that we believe to possibly be located in the insert of the clone 19, that appeared by searching in the *Serratia proteamaculans* genome; we see two possibilities. One is that the genes with unknown function produced an enzyme that has activity over the lignin alkali, degrading it and enabled the cell to use it as carbon source. But that can only be known by extracting the enzyme produced by those unknown genes and test its activity over lignin alkali. The second possibility, the one we are adhering to; is that those genes involved in the biosynthesis or degradation of aminoacids present in the unsequenced segment helped the cell to grow unther those harsh conditions. Those aminoacid related genes are shaded with blue in table 13. That can explain why the clone 19 grew. As was explained earlier, EPI 300 T1 resistant host cell line has a mutation that makes it auxotroph to leucine and to thiamine; reason why we had to add those compounds to the minimal media. Is a possible explanation that those genes shaded with blue are using the Leucine as carbon source effectively allowing the clone to grow in the minimal media even without using lignin alkali as carbon source. For example the Spro_4751Ketol-acid reductoisomerase is involved in Leucine biosynthesis. The gene Spro_4758 branched-chain aminoacid aminotransferase is involved in Leucine degradation and biosynthesis. So we possibly have an array of genes in the insert that can utilize effectibly an aminoacid as carbon source for the clone 19.

4.4.2 Caffeine degrading candidate sequenced genes; clone 28.

For the caffeine degrading candidate we sequenced with primers that recognized the corners of the transposons and with primers that recognized the corners for the fosmid itself. We obtained the following hits:

Set of genes from the sequencing of the transposons after primer walking that appear to be close to one another in an unknown place in the approximately 40kb insert were:

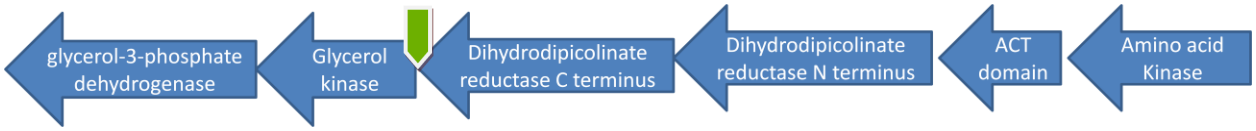


Figure 29: genes in clone 28 from primer walking sequencing

In this image we can see the putative direction of these genes following the ORF analysis. The green insert presents the localization where the transposon was inserted.

An additional set of genes from the sequencing of the transposons after primer walking that appear to be close to one another in an unknown place in the approximately 40kb insert were:

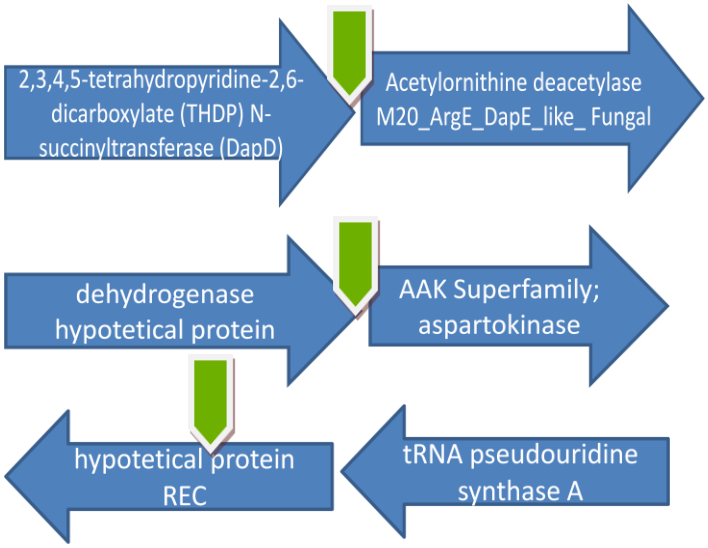


Figure 30: additional genes in clone 28 from primer walking sequencing

In this image we can see the putative direction of this gene following the ORF analysis. The green inserts are presenting the localization where the transposons were inserted.

Genes from the sequencing of the transposons that were in an unknown place in the aproximalely 40kb insert:

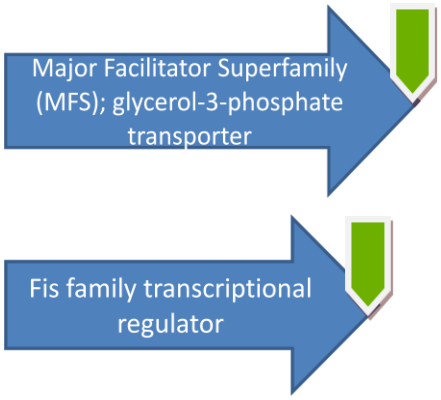


Figure 31: isolated genes in clone 28 from primer walking sequencing

In this image we can see the putative direction of this gene following the ORF analysis. The green inserts are presenting the localization where the transposons were inserted.

Genes from the sequencing of the fosmid from the corners using the forward primer:

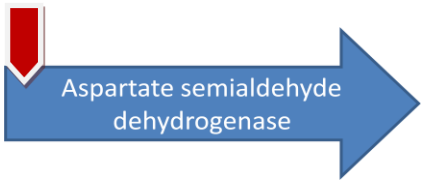


Figure 32: gene sequenced from the corner of the fosmid; forward primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the forward primer.

From all these 15 possible genes 2 genes called our attention; dihydrodipicolinate reductase (*DapB*) and (2,3,4,5)-tetrahydropyridine-(2,6)-carboxylate N-succinyltransferase (*DapD*). Searching in the literature we found that both of this genes are part of the same pathway; the diaminopimelic acid pathway important for the biosynthesis of lysine and of meso-diaminopimelate. Lysine and meso-diaminopimelate (DAP) are two essential metabolites for the development of Gram negative bacteria. Lysine is an essential aminoacid and DAP is a component of the peptidoglycan layer of Gram negative bacteria cell walls. By random transposon mutagenesis in diverse genera of Gram negative bacteria like *Pseudomonas aeruginosa* and other genera of bacteria like *Mycobacterium tuberculosis* (an acid resistant Gram positive bacteria) and *Salmonella typhymurium* (a Gram negative bacteria) researchers like Sreelatha G. Reddy had proven that by the inhibition of genes involved in the DAP pathway, those microbes mentioned died probably because of the instability of the peptidoglycan cell wall created by the absence of DAP biosynthesis (100). DAP is not required nor is produced by humans. For that reason some researchers have considered these two enzymes as possible targets for the development of an antibiotic or an antimycobacterial agent that can act as inhibitor of one or multiple steps in the biosyntetic pathway of DAP (101), (102). More interesting is that some of the other genes that we found are involved also in the same diaminopimelic acid pathway. The pathway for Lysine and DAP is described in the Figure 36. There we can see that some of the enzymes we have sequenced appear in the pathway: aspartokinase, aspartate semialdehyde dehydrogenase, dihydrodipicolinate reductase (*DapB*) and (2,3,4,5)-tetrahydropyridine-(2,6)-carboxylate N-succinyltransferase (*DapD*).

Is important to mention an additional detail; when we analyzed by blast n the sequence we have from the primer walking for the clone 28 (19), we received as result a negative hit. We sequenced 3,178 bp's for this clone 28 that came from the D2-1 Metagenomic Library made with soil from Guanica Dry Forest and we obtained in blast n a negative hit, in other words the

system didn't recognized any known microbe with a similar sequence. Yet for the two principal genes (*DapD* and *DapB*), the blast x hits related them to genes from *Candidatus solibacter*. The sequence from transposon 28(6) analysed by blast x, a querry of 2,257 bp's, gave us a result that resembles the DapD gene from *Candidatus solibacter* with a 32% of querry coverage and an e value of 7e-96 (Figure 31).

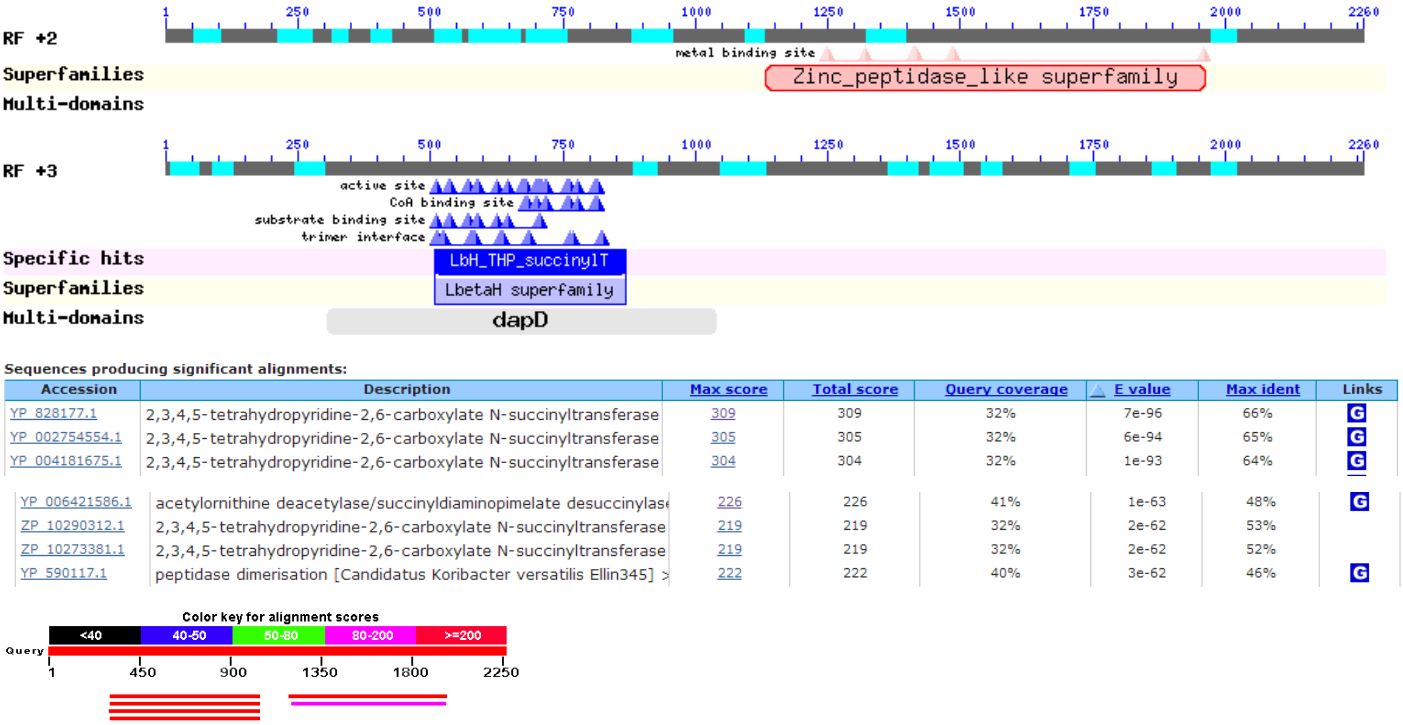


Figure 33: putative domain and blast x results from transposion 28(6); DapD gene

In this image we can see the putative conserved domains and the blast x hits where both genes appeared the dapD gene and the zinc peptidase.

The sequence from transposon 28(19) that resembling a *DapB* gene appeared only in the putative conserved domain when we analyse with blast x the sequence of 3,178 bp's from transposon 28(19) (Figure 32 and Figure 33).

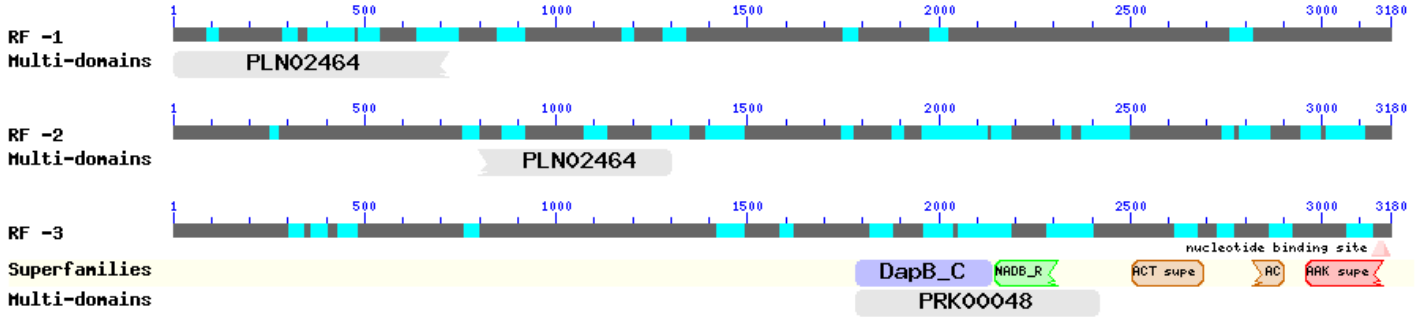
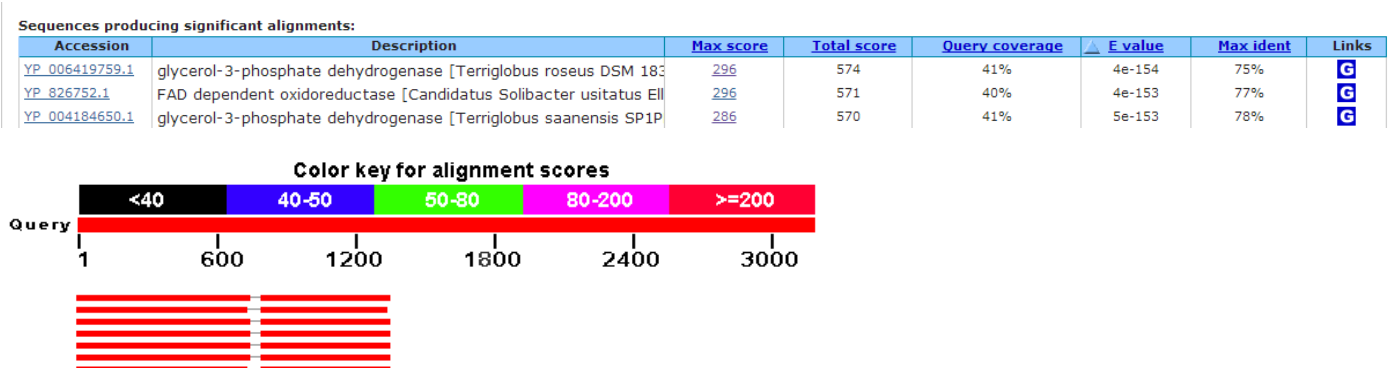


Figure 34: putative conserved domains where DapB gene appears.

In this image we can see the putative conserved domains where the glycerol 3 phosphate dehydrogenase (in grey) appears like PLNO2464 and the DapB gene appears in blue and green.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
YP_006419759.1	glycerol-3-phosphate dehydrogenase [Terriglobus roseus DSM 183	296	574	41%	4e-154	75%	G
YP_826752.1	FAD dependent oxidoreductase [Candidatus Solibacter usitatus Ell	296	571	40%	4e-153	77%	G
YP_004184650.1	glycerol-3-phosphate dehydrogenase [Terriglobus saanensis SP1P	286	570	41%	5e-153	78%	G

Figure 35: blast x of sequence from clone 28(19) where the Dap gene didn't appear in the right part as in the case of the putative domain.

In this image we can see that the Dap B gene that appears in the putative conserved domains doesn't appears in the blast x by sending to analyze the 3,178 bp's from transposon 28(19). Only appears in the blast x a glycerol 3 phosphate dehydrogenase.

By using the same analysis that we performed with the magnesium chelatase case and the lysR Hdfr that didn't appeared in the sequence, we cut approximately 1,500 base pairs (bp's) of the beginning of the sequence and blasted the rest of it, where the dihydrodipicolinate reductase was supposed to be by following the putative conserved domain diagram, and we obtained a hit that resembles a *DapB* gene from *Candidatus solibacter* by using a querry of 1,599 bp's with 40% coverage and an e value of 4 e-54 (Figure 34 and Figure 35).

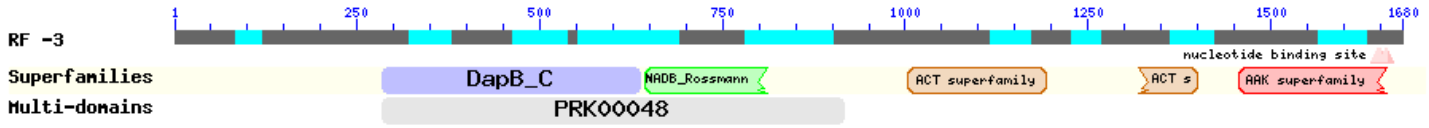


Figure 36: putative conserved domain of the 1,599 bp's from the right side of the sequence of 3,178 bp's from transposon 28(19)

In this image we can see that by cutting the 3,178 bp's sequence and analyzing by blast the segment that is supposed to contain the Dap B gene, this gene appears in the putative conserved domain.

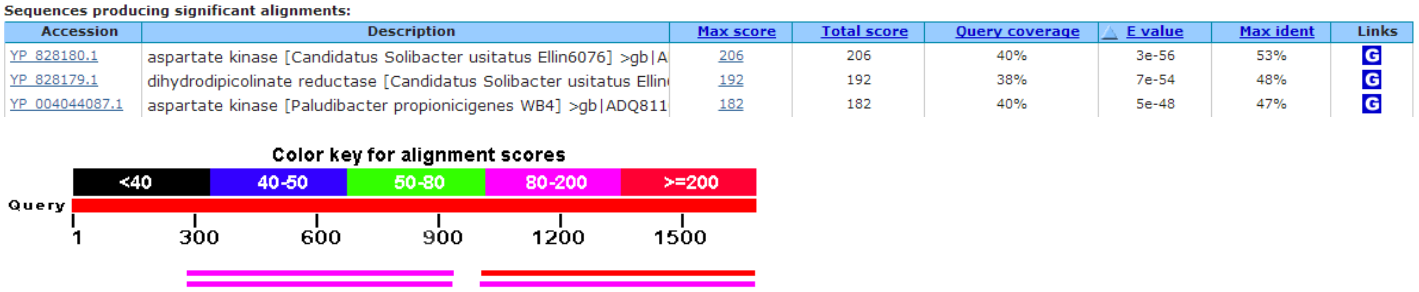


Figure 37: blast x of the 1,599 bp's sequence from the right side of the sequence of 3,178 bp's from transposon 28(19)

In this image we can see that by cutting the 3,178 bp's sequence and analyzing by blast x the segment that is supposed to contain the Dap B gene, also appears as a good hit.

Even when these results are very interesting they cannot give evidence to explain that the clone grew because of a gene that enabled the cell to degrade caffeine as carbon source. Instead of that the results appear to be similar to the ones found for clone 19, the leucine present in the media can be the reason for the growth of clone 28 who is using the aminoacid biosynthesis and degradation genes present in the insert contained in the fosmid to break down the Leucine for catabolism and anabolism. We compared the genome of *Candidatus solibacter usitatus* with the genes we found following as example what we did for the clone 19 and we found all the genes related to the DAP pathway pointing in the same direction confined in a space of approximately 30,000 bp's. Yet we didn't saw the rest of the genes we sequenced like glycerol-3-phosphate dehydrogenase or tRNA pseudouridine synthase. That confirmed to us that the result obtained from the blast n analysis was a correct one, this segment of DNA is possibly from an unknown microbe of soil.

We didn't find an explanation that correlated the degradation of caffeine to the growth of the clone with caffeine degradation but with aminoacid degradation. It seems like the addition of Leucine and thiamine to the minimal media affected the screening by including in the media alternative carbon sources that even in very small concentrations, can be used for the growth of the clone. And if the clone has a mutation that inhibits its leucine biosynthetic pathway but if the

insert contains genes that can substitute steps blocked by the mutation in that pathway, the clone can find a new way of generating energy from Leucine as carbon source.

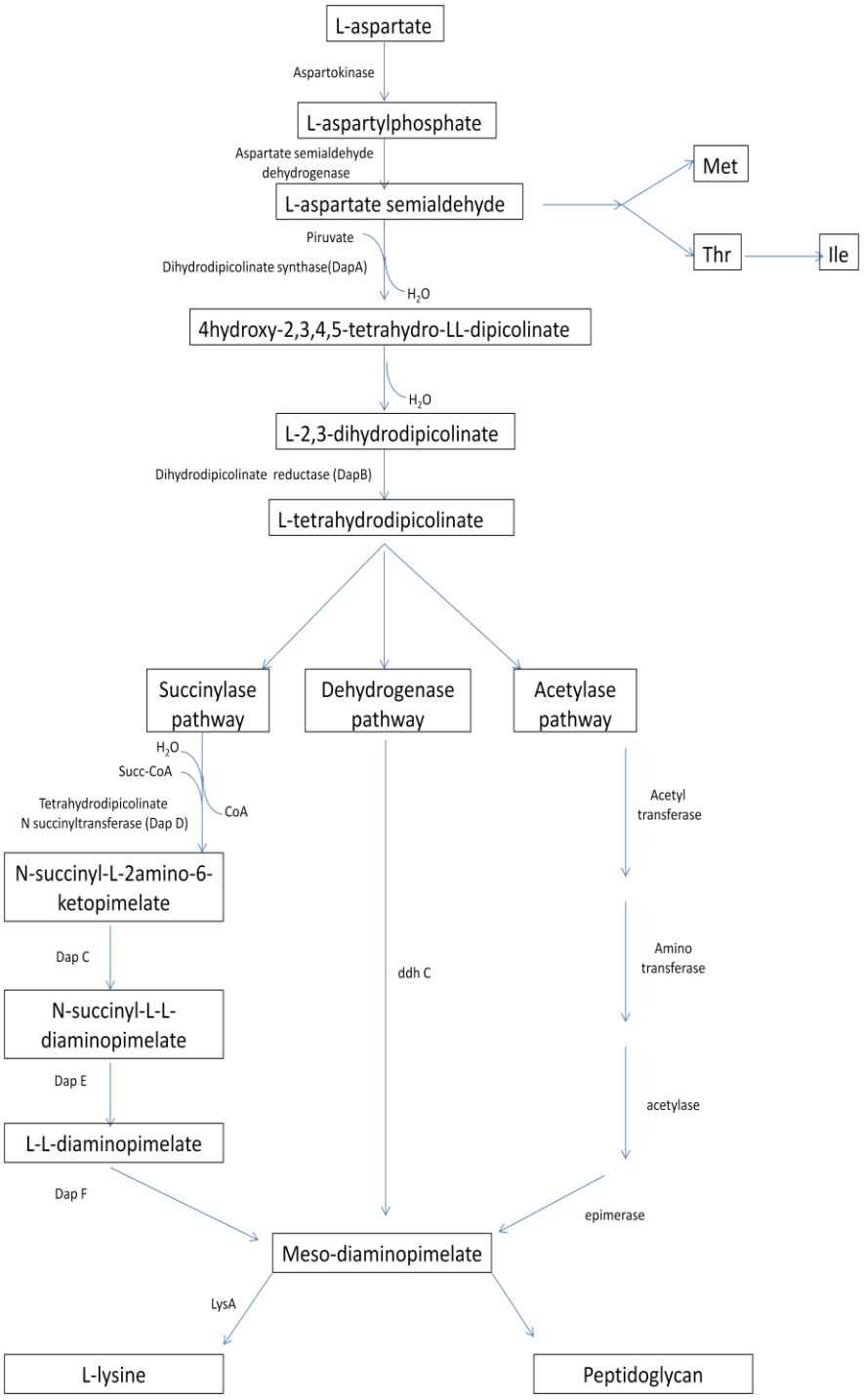


Figure 38: diaminopimelic acid pathway of lysine biosynthesis (90), (91).

In this image we can see the diaminopimelic acid pathway of lysine biosynthesis where appear the genes Aspartokinase, DapB and DapD that we sequenced.

4.4.3 Cellulose degrading candidate sequenced genes; clone 13

For the cellulose degrading candidate we only sequenced with primers that recognized the corners of the fosmid. We obtained the following hits:

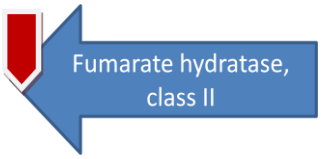


Figure 39: sequenced from the corner of the fosmid; forward primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert indicate the localization where the gene was sequenced; sequenced with the forward primer.

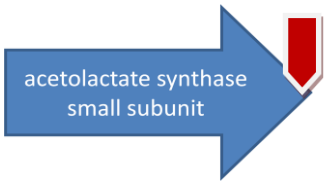


Figure 40: gene sequenced from the corner of the fosmid; reverse primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert indicate the localization where the gene was sequenced; sequenced with the reverse primer.

With this small amount of information we cannot say that this clone doesn't have genes that confer the clone the ability to use carboxymethyl cellulose as carbon source. Yet again we cannot either affirm that the reason why the clone grew in the presence of carboxymethyl cellulose in the minimal media M9 was because it was able to produce a cellulose degrading enzyme. Even with that we can see something that has been a constant trought all the results presente to this point; one more gene related to aminoacid metabolism. Acetohydroxyacid synthases (AHASs) are found in plants, fungi and bacteria and are cabaple of synthetizing de novo branched chain aminoacids (BCAAs) (103). As other researcher we have cited J.A. McCourt explains that enzymes present on the biosynthetic pathway of BCAAs like AHASs, dihydroxyacid dehydratase and ketol acid reductoisomerases are potential targets to develop herbicides, fungicides and antibiotics because these enzymes are not found in animals nor in humans (104). For the case of AHAs; they are the first common enzyme in the biosynthetic pathway of BCAAs and is capable of catalyzing the synthesis of 2-acetolactate or of 2 aceto 2 hydroxybutyrate. For the case of fumarate hydratase class II we found reading that it is involved in the crebs cycle in the reversible hydration-dehydration of fumarate to L malate.

4.4.4 Tween degrading candidate sequenced genes

4.4.4.1 clone A

For this tween degrading candidate we only sequenced with primers that recognized the corners of the fosmid. We obtained the following hits:

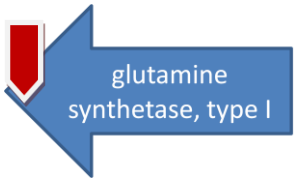


Figure 41: gene from clone A sequenced from the corner of the fosmid; forward primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the forward primer.

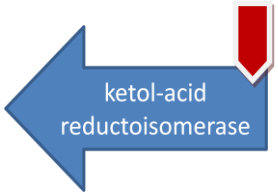


Figure 42: gene from clone A sequenced from the corner of the fosmid; reverse primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the reverse primer.

With this genetic information we cannot say that this clone doesn't have genes that confer the clone the ability to use tween as carbon source. Also we cannot either affirm that the reason why the clone grew in the presence of tween in the minimal media M9 with greater strength that the control EPI 300 was because it was able to produce a tween degrading enzyme. Even with that we can see again some genes related to aminoacid metabolism. We have already mentioned ketol-acid reductoisomerase as one of the possible enzymes contained in the insert of clone 19. Ketol-acid reductoisomerase is present in diverse pathways: valine, leucine and isoleucine biosynthesis; pantothenate and CoA biosynthesis. For Glutamine synthetase we found that it catalyzes the first step at which nitrogen is brought into cellular metabolism: glutamate + $NH4^+$ + ATP => glutamine + ADP + Pi. The product glutaminie, an aminoacid; is a source of nitrogen in the biosynthesis of many other metabolites. (79)

4.4.4.2 Clone C

For this tween degrading candidate we only sequenced with primers that recognized the corners of the fosmid. We obtained the following hits:

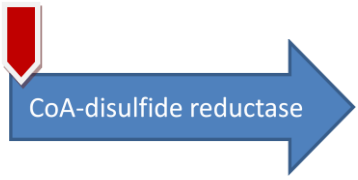


Figure 43: gene from clone C sequenced from the corner of the fosmid; forward primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the forward primer.

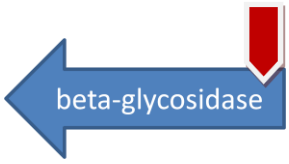


Figure 44: gene from clone C sequenced from the corner of the fosmid; reverse primer.

In this image we can see the putative direction of this gene following the ORF analysis. The red insert is the localization where the gene was sequenced; sequenced with the reverse primer.

The hit beta-glycosidase is a very interesting and important hit for us. Even when in blast x the amount of bases recognized for this hit were low; if future sequencing from primer walking confirms to be a beta-glycosidase; this hit will correspond to what we were looking for. For example in 2011 Cheng-Jian Jiang reports the finding of a β-glucosidase with lipolytic activity from a soil metagenomic libray expressed in *E. coli* (105). But more research has to be made to see if this enzyme is really a beta-glycosidase and if it has a lipase like activity. We obtained this hit by blast x and our hit resembles a beta-glycosidase from the microbe *Terrabacter ginsenosidimutans* by analyzing a querry of 604 bp's with coverage of 44% and an e value of 4 e-09 (Figure 43).

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
ACZ66247.2	beta-glycosidase [Terrabacter ginsenosidimutans]	63.9	63.9	44%	4e-09	45%
CBL80578.1	transglutaminase-like domain protein [uncultured Leeuwenhoekiell	58.9	58.9	14%	1e-08	93%
ACQ84163.1	truncated LacZ [BAC cloning vector attB-P[acman]-CmR-BW-F-2-	56.6	56.6	13%	6e-08	89%
AAD31805.1	LacZ [Cloning vector pHIND2.2]	56.6	56.6	13%	6e-08	89%
CAD50590.1	LacZ alpha protein [Cloning vector pUvBBAC]	56.2	56.2	13%	1e-07	89%
CBL87506.1	ABC transporter ATP-binding protein [uncultured Flavobacteriia ba	57.8	57.8	13%	1e-07	93%
ADJ00056.1	red fluorescent protein [Mariner mini-transposon delivery vector p	56.2	56.2	13%	6e-07	89%

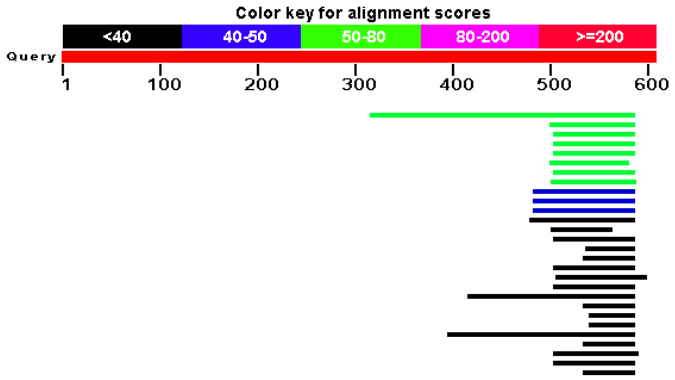


Figure 45: beta glycosidase hit by blast x of 604 pb's of sequence from clone C

In this image we can see that the first hit and the longest one corresponds to a beta glycosidase.

Blast x description of this enzyme is that it resembles to a beta glycosidase with a hydrolyzing function over ginseng saponin. Also blast x description of this enzyme says that it is a beta-glucosidase-related glycosidase involved in carbohydrate transport and metabolism and that it possibly belongs to the glycosyl hydrolase family 3. Glycoside hydrolase members of family 3 are classified as β-D-glucosidases, α-L-arabinofuranosidases, β-D-xylopyranosidases and N-acetyl-β-D-glucosaminidases (106). Saponin molecules are found in plants and lower marine animals. Figure 46 presents some examples:

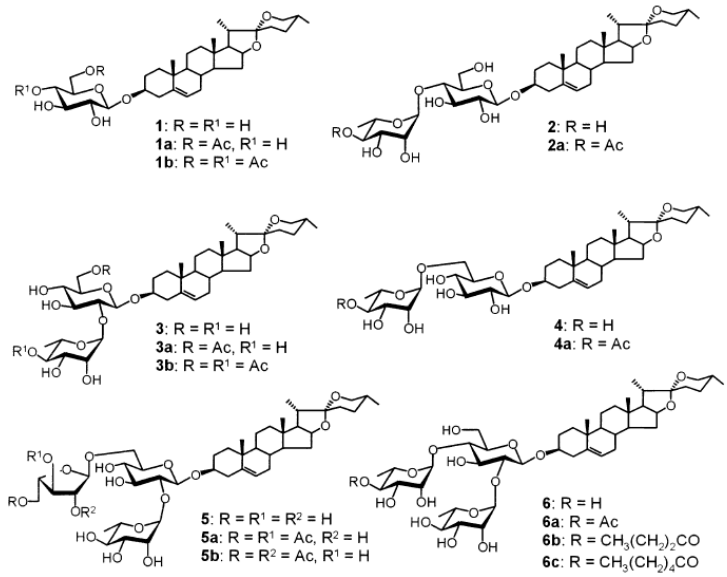


Figure 46: Diosgenyl saponins and their acylated derivatives.

This image shows the most common structures of saponins that can be found in plants and in some lower marine animals (107).

Biao Yu et al., 2001 employed diogenyl β-D-glucopyranoside the simplest saponin structurally speaking to screen for lipase activity over this substrate and found lipases to be able to act over it (107). It has been proven that lipases have activity over saponins and our hit tells us that we possibly have a beta glycosidase with a hydrolyzing function over ginseng saponin. This is our only hit that can possibly explain the growth of the clone in the minimal media by degrading the carbon source intended to (tween) and not by the use of the aminoacid Leucine and the thiamine present in the media.

4.4.5 Genes related to biofuel production

From the list of genes isolated from the screening of our three environmental metagenomic libraries we can extract three genes as possible useful gene producing enzymes for the production of biofuels: 1.) Beta glycosidase from clone C, 2.) Glycerol 3 phosphate dehydrogenase from clone 28; and 3.) Ketol acid reductoisomerase from clone A.

4.4.5.1 Beta glycosidase from clone C

As we discussed earlier this result has the potential to be a lipase. Lipases had exhibited a wide substrate specificity degrading compounds like p-nitrophenyl esters, tweens, phospholipids and even caffeine. Further analysis has to be done to assure with certainty that we have a lipase/glycosidase that has activity over saponins. Is interesting that uniprot describes beta-glycosidases as hydrolases that are not highly specific (like the lipases) and that attack glycosidic bonds in carbohydrates, glycoproteins and glycolipids. We believe that there is a possibility that this enzyme has activity over a variety of substrates like saponins and lipids (like tween) because of uniprot description of the enzyme. This putative Beta glycosidase came from clone C that was isolated from the hypersaline Microbial Mat metagenomic library CH2. If it results to be a lipase; it will be interesting to perform kinetic studies to it to see how efficient this enzyme is degrading lipids like tween or saponins or both. Also additional studies are necessary to see how tolerant to organic solvents is, how tolerant to pH is, how tolerant to salinity or to alkalinity is; because clone C could contain enzymes with special characteristics like salinity tolerance or organic solvent tolerance that can improve an actual industrial process because its insert came from a extreme environment with high salinity. The querry analized on blast x for this insert was 604 bp's with a coverage of 44% and an e value of 4e-09 and the hit resembles the beta-glycosidase from *Terrabacter ginsenosidimutans*.

4.4.5.2 Glycerol 3 phosphate dehydrogenase from clone 28

Glycerol is an abundant carbon source. Actually it is a byproduct of the production of biodiesel and diesel. Gervásio Paulo da Silva et al., 2009 explains that the production of biodiesel from animal fats and oils extracted from vegetable sources generates as by product almost 10% (w/w) glycerol (108). He also points that glycerol can be used to produce fine value chemicals like 1,3-propanediol, dihydroxyacetone, ethanol, butanol, 2,3-butanediol, citric acids, pigments and biosurfactants. Glycerol can be metabolized to dihydroxyacetone phosphate by phosphorilation by glycerol kinase and subsequent conversion of sn-glycerol-3-phosphate (G3P) into dihydroxyacetone phosphate (DHPA) through the action of glycerol-3-phosphate dehydrogenase (109). And from dihydroxyacetone phosphate by additional enzymatic action an array of fine value chemicals can be produced (figure 46). Even when this type of enzyme had been sequenced and analyzed before; kinetic studies must be done to ours to see how efficient this enzyme is, how tolerant to organic solvents, pH, salinity, alkalinity is because as Pamela P. Peralta explains “in DNA, genes involved in the synthesis of fuel can be introduced at the desired numbers into a host genome (110);” and if ours is superior to others in these aspects it can be a substitute for an uneficient glycerol 3 phosphate dehydrogenase. Below is the glycerol 3 phosphate dehydrogenase blast x hit from a sequence of 3,178 bp’s (querry) obtained after sequencing transposon 28(19) that came from the library D2-1 of our dry forest metagenomic library. Our gene resembles a glycerol 3 phosphate dehydrogenase from *Terriglobus roseus* DSM 18391 with query coverage of 41% and an e value of 4e-154 (Figure 45).

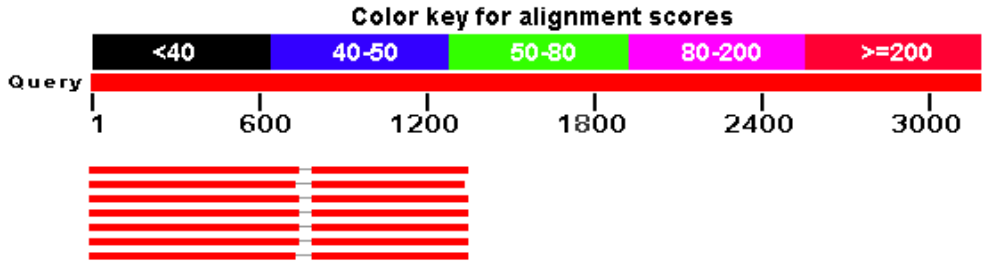


Figure 47: glycerol 3 phosphate dehydrogenase blast x hit from clone 28.

Blast x hit for the glycerol 3 phosphate dehydrogenase sequenced from the clone 28.

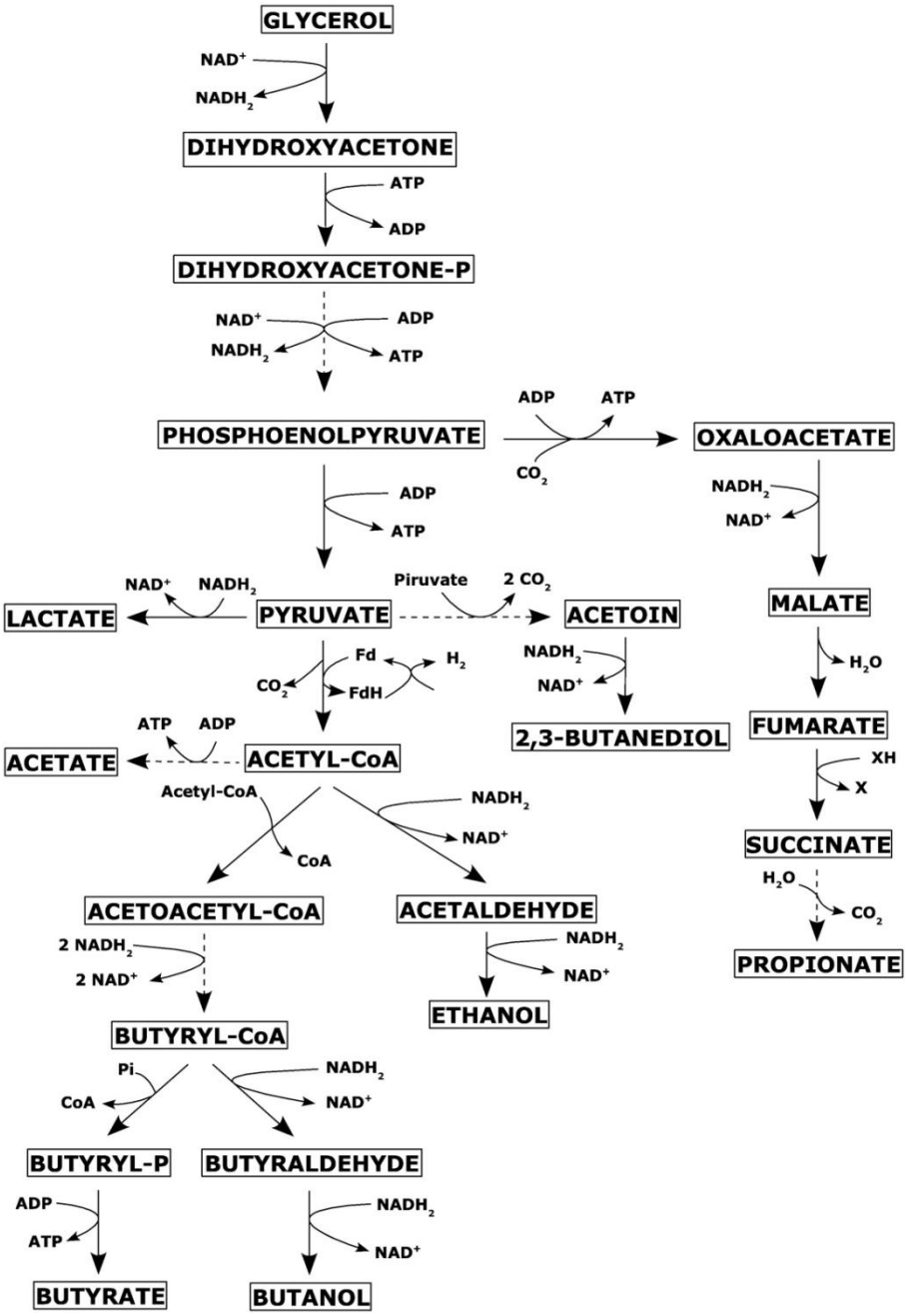


Figure 48: Overview of some possible end products from different microorganism during glycerol degradation.

Multiple routes for the biotransformation of glycerol into value added chemicals (108).

4.4.5.3 Ketol acid reductoisomerase from clone A

We mentioned that ketol acid reductoisomerase (KARI) is present in diverse pathways related to aminoacid biosynthesis. We also mentioned that it is an enzyme studied to find inhibitors to develop new biocidal agents. But in addition to that, KARI is related to the production of an alcohol based biofuel; to isobutanol production. Isobutanol is a potential substitute of gasoline. We can number some reasons why Isobutanol is a good candidate as gasoline substitute: 1.) is an excellent gasoline blend chemical, 2.) is a precursor to C4 petrochemical building blokcs, 3.) it has a high yield production by microbial fermentation, 4.) it has already reached later phases of comerzialization (Gevo based in Englewood-Colorado and Butalco in Fuerigen-Switzerland are producing it by microbial fermentation), 5.) it is an energy dense, low vapor pressure, high octane hydrocarbon that burns in a combustion engine like conventional gasoline without loosing performance, 6.) it can be converted to butenes and other commodity chemicals (110), (111). Again even when this enzyme have been isolated from other organisms and even when is in used already, is important to perform kinetic studies to our KARI to see how efficient this enzyme is, also additional studies are necessary to see how tolerant to organic solvents is, how tolerant to pH is, how tolerant to salinity or to alkalinity is; because clone A came from a hypersaline Microbial Mat soil sample and this enzyme could have special characteristics; for example there is a chance that this enzyme is resistant to high concentrations of salt (halotolerant). To have an efficient and tolerant enzyme could help to improve the 2 ketol acid pathway for the production of Isobutanol in industrially usefull microbes. For example Pamela P. Peralta et al., 2012 tells us that there are reports of researchers that have engineered the 2 ketol acid pathway in *Corynebacterium glutamicum* for the production of isobutanol, knowing that *Corynebacterium glutamicum* is a high aminoacid productor bacterium (110). Pamela explains us that to use advance biofuels as substitutes for the cheaper fosil fuels, we must find efficient processes that are economically competitive with existing products production processes. That's why is important to increase our knowledge of enzymes to be able to find the more efficient ones that can act as substitutes for less efficient enzymes improving known processes that can substitute fosil based fuels with biobased biofuels. An example of an engineered pathway for the production of isobutanol is depicted in figure 46. Our ketol acid reductase from clone A came from the library FS6 from the hypersaline Microbial Mats from Cabo Rojo Puerto Rico. We analyzed a 505 bp's sequence and obtained a coverage of 74% and an e vaule of 4e-50 for our Ketol acid reductoisomerase that resembles a ketol-acid reductoisomerase from *Pyrococcus sp. NA2*.

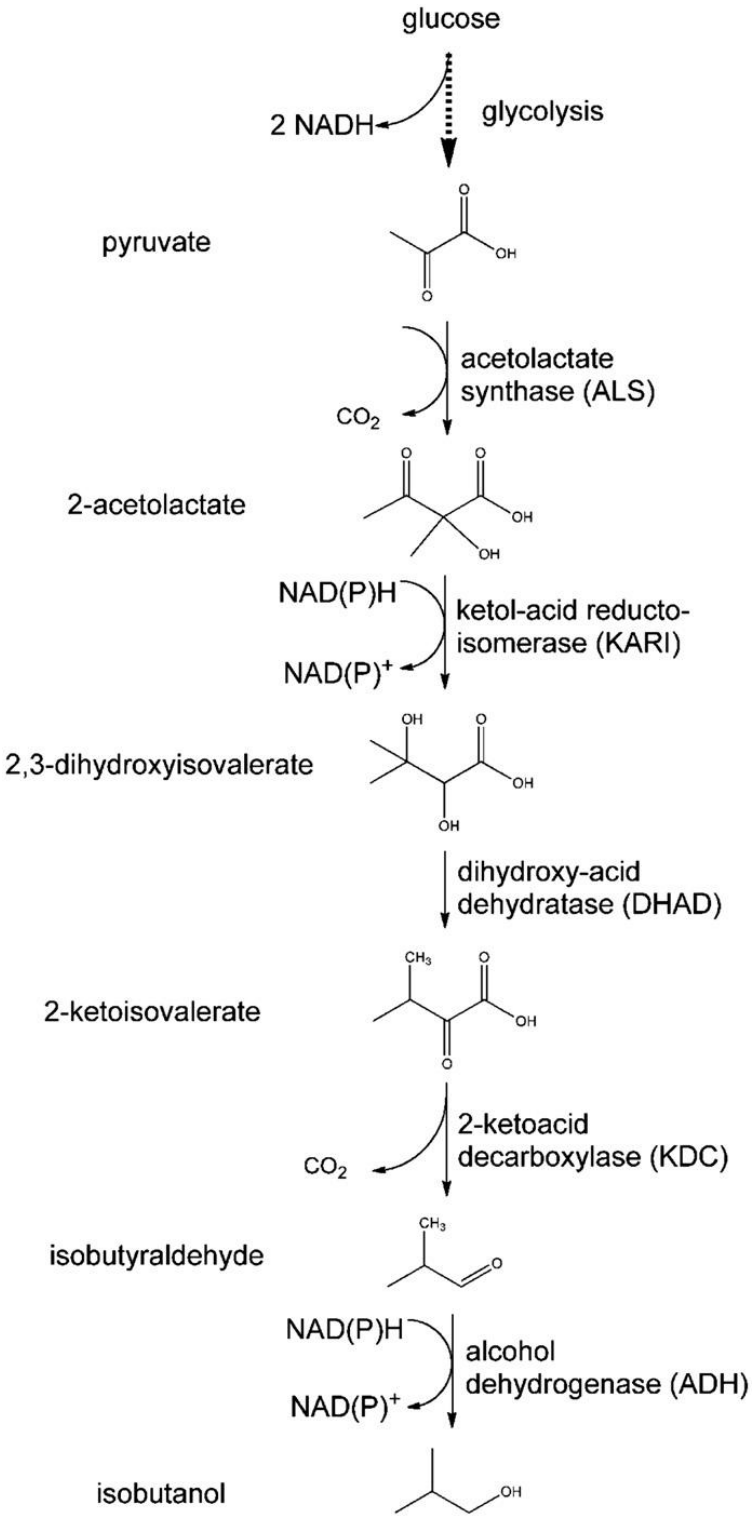


Figure 49: Engineered pathway for isobutanol production.

Engineered metabolism for the production of isobutanol (110). Here we can see the enzyme KARI as part of the metabolism a similar enzyme to the one we isolated from clone A.

5. Conclusions

We proved that the metagenomic approach provide a usefull method to get genes from unknown microbes. An example of that is the result of blast n for the sequence of the clone 28 which didn't resemble to any known genome on the database and the hits for the individual sequences appeared to resemble to genes from different microbes not to the same microbe. We prove that the metagenomic approach provide a usefull method to get genes related to known microbes like for the example clone 19 hits related to *Serratia* type genes. We also found additional clones with the ability to grow on lignin alkali, caffeine, tween and cellulose. We found interesting genes related to the degradation and biosynthesis of aminoacids compounds and interesting genes usefull for the bioengineering of biofuel producing metabolic routes proving that the metagenomic research can be a tool to unveil genes encoding for enzymes usefull for a variety of applications. As for the putative genes we have isolated we propose as responsible candidates for the growth of the clones in their respective media: the gene LysR Hdfr as responsible of the growth of clone 19 in the M9 minimal media with lignin alkali as test compound; the genes related to the diaminopimelate pathway as responsable for the growth of the clone 28 in the M9 minimal media with caffeine as test compound; the gene Ketol acid reductoisomerase as responsible of the growth of clone A in the M9 minimal media with tween as test compound and the putative Beta Glycosidase as responsible of the growth of clone C in the M9 minimal media with tween as test compound.

We think that most of the growth from the mayority of the clones was due to the presence of Leucine in the minimal media. If the host cell EPI 300 didn't had the mutations that make it auxotroph for leucine and thiamine we could have narrowed the research to only the carbon sources we wanted to test. Nevertheless the leucine apparently added an additional carbon source for which we tested also the libraries. Even with the presence of leucine in the media; why we didn't obtain more specific results? We think that Don Cowan et al., 2005 expresses the reason better that anybody: "The limitation of *E. coli* as a host for comprehensive mining of metagenomic samples is highlighted by the low number of positive clones obtained during a single round of screening (typically less than 0.01%)…This suggests that without sample enrichment the discovery of specific genes in a complex metagenome is technically challenging (28)." Following this thought of Cowan we can understand that finding genes that can produce specific enzymes that can degrade specific compounds needs a more specific approach in the building of the metagenomic library. If we construct a metagenomic library from

a environmental sample without an enrichment process and screen it for the compound we are interested in finding degrading enzymes for; we are searching for a specific gene in a environmental sample where “one gram of soil may harbor up to 10 billion microorganisms of possibly thousands of different species” (8). The population that may contain a specific gene we are searching for can be just a segment of the total population and enrichment, even when it is a process that will alter the representativity of the population on site; will increase the biomass of the microbes with the genetic material to use the enriching substrate giving us a better chance to obtain the genes we are searching for. Not only the population that may contain a “specific gene” is just a segment of the total population; it also can be from different species which brings in another problem.

Another strong reason we found to describe the difficulties we have in obtaining more specific hits for the chemicals we were testing can be adjudicated to the host cell. In the environment we have Gram-positive, Gram-negative, yeast, algae and some microscopic animals that contain different and diverse genetic material with diverse GC content. Matthew D. McMahon et al., 2012 explains that *E. coli* can express only 40% of the environmental genes and that value drops to 33% for high GC% actinomycete DNA (112). So to increase our chance of obtaining better results from mining the metagenome we must use additional host cells; others that *E.coli*. We must also use Gram + bacteria and Yeasts. Gram + bacteria as host may get to that other percent of DNA material (approximately 60%) that *E. coli* cannot express and Yeast could be the most suitable host for expressing eukaryotic material like the one found in fungi. Yeast has the versatility of a similar growth to bacteria but with the genetic machinery of an eukaryotic organism. In addition to Gram positive, Gram negative and yeast some researchers think that if we are searching for enzymes in extreme environments, like archae microbes; we must use as host cell a microorganism with the genetic material machinery to survive in those exteme conditions because it will be more suited to express the genes we are trying to obtain from the metagenomic approach.

For these reasons we propose a different approach to mine the treasure chest of genetic material found in the environment. First we must not forget the 1% of the microbes that fall in the category of culturable microorganism. Don Cowan et al., 2005 expresses that the term “unculturable” is inappropriate and that we just have to discover the correct culture conditions to obtain more culturable results (28). Even think about it, how much is 1% of “10 billion microorganisms of possibly thousands of different species”. After taking into account the

culturable portion of the environmental sample we want to test; the next part is to make an intelligent selection of the chemicals to be used in the media preparation. If we know anticipatedly the substrate and the product we want to obtain, even when we don't know the intermediates of the pathway, we can make better decisions choosing the subrstrate for enrichment. For example if we want to produce fine value chemicals from cellulose we can search for known microbes that use cellulose and see the possible products they produce. Then we can enrich with cellulose the environemtal sample we want to use to construct a metagenomic library, and search for known degradation products and compare them with new degradation products. The next step will be to select a vector to carry the environmental insert. It has to contain the fosmids characteristics but it has to contain also a gene that enables the overexpression of the genes contained in the insert. Sometimes the clones of the metagenomic libraries contain a gene with the characteristics we are searching for but because the expression of the enzyme the gene codifies for is low, the clone didn't show the phenotype that is supposed to demonstrate because it contains the genotype. Also the vector has to be in direct correspondence with the host cell because not all vectors works on all types of hosts.

The next step is to select the host cells to be used for the development of clones, of the metagenomic library. For Gram negative host organism we can choose *E.coli.* For Gram-positive bacteria we can choose as good option a developed Streptomyces lividans TK24 (113). For an eukaryotic host we can choose the well known yeast *Saccharomyces cerevisiae* because is an actual industrial yeast strain used in industrial production of bread, food ingredients and alcohol (114); and because there is evidence of plasmid shuttle vectors that works on both *E. coli* and *Saccharomyces cerevisiae* for the insertion of exogenoud DNA into the host (115). And if the researcher wants to make a metagenomic library from a harsh environment, they should search for a culturable extremophyle that can serve as host cell. For example if the environmental area of interest from where the DNA sample is to be extracted is a hypersaline soil sample, an halotolerant archaea can be used as host cell for the preparation of the metagenomic library.

After the chemical to be tested is known, the vectors and the corresponding number of host microbes to be used, the next step is the enrichment. Here we can measure the difference in population by T-RFLPbefore and after enrichment (116), to identify with evidence which genera of the microbes present in the population of the environmental sample are rising its biomass after the enrichment. The enrichment can be done on site if the chemical to be tested

is not toxic to the environment or of site preparing a mesocosm with all the same environmental conditions to be able to enrich with a toxic or environmentally dangerous material without affecting the environment. And finally after the enrichment is done the final step is to extract the DNA from the environmental sample, make a metagenomic library and screen with minimal media and with colorimetric assays searching for the degradation of the compound elected.

The possibilities by doing a metagenomic research this way, of finding expected genes with a higher hit ratio is greater than by just not inducing the biomass growth of the organism that has the genes we are interested in. Following this procedure cannot be a work of one or two persons but of a group of people working under the supervision of the counselor. One team will approach the cultivable portion. Another team will work with a Gram positive host, another with a Gram-negative host, another with yeast as host and in the case of a harsh environment another with an extremophile host. By following this procedure a greater success is highly possible and we can extract more treasures from that treasure chest called metagenome.

6. References

Article I. Works Cited

1. *Summer Workshop in Metagenomics: One Week Plus Eight Students Equals Gigabases of Cloned DNA.* **Rios-Velazquez, Carlos, et al.** Wisconsin, Madison : JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, 2011.
2. *Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities.* **Kevin Chen, Lior Pachter.** 2, Berkeley, California : PLoS Computational Biology, 2005, Vol. 1 .
3. *Community Structure, Geochemical Characteristics and Mineralogy of a Hypersaline Microbial Mat, Cabo Rojo, PR.* **Casillas-Martinez, Lilliam, Gonzalez, and Millie L. and al, et.** Humacao, Puerto Rico : Geomicrobiology Journal, 2005, Vol. 22.
4. *Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms.* **Michelle R. Rondon, Paul R. August, Alan D. Bettermann.** 6, Madison, Wisconsin : Applied and Environmental Microbiology, 2000, Vol. 66.
5. http://www.expertglossary.com. *ExpertGlossary.* [Online] [Cited: December 10, 2012.] http://www.expertglossary.com/biotech-genetics/definition/primer-walking.
6. *Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass.* **Klinke HB, Thomsen AB, Ahring BK.** 1, Roskilde, Denmark. : Applied Microbiology and Biotechnology, 2004, Vol. 66.
7. *Exploring the microbial biodegradation and biotransformation gene pool.* **Teca Calcagno Galvao, William W. Mohn, Victor de Lorenzo.** 10, Madrid, Spain and Vancouver, Canada : Trends in Biotechnology, 2005, Vol. 23.
8. *Microbial diversity and function in soil: from genes to ecosystems.* **Ovreas, Vigdis Torsvik and Lise.** Bergen, Norway : Current Opinion in Microbiology, 2002.
9. *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* **Jo Handelsman, Michelle R Rondon, Sean F Brady.** Wisconsin, Madison : Chemistry and Biology, 1998.
10. *Prospecting for novel biocatalysts in a soil metagenome.* **Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR.** 10, Julich, Germany : Applied and Environmental Microbiology, 2003, Vol. 69.
11. *Size does matter: Application-driven approaches for soil metagenomics.* **Kavita S. Kakirde, Larissa C. Parsley, Mark R. Liles.** Elsevier, Auburn, AL 36849, USA : Soil Biology and Biochemistry, 2010.
12. *Unraveling activities by functional based approaches using metagenomic libraries from dry and rain forest soils in Puerto Rico.* **José M. Cruz, Manuel A. Ortega, Jean C. Cruz, Pedro Ondina, Rossivette Santiago, Carlos Ríos Velázquez.** Mayaguez, Puerto Rico : Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology, 2010.
13. *Recovery, Purification, and Cloning of High-Molecular-Weight DNA from Soil Microorganisms.* **Mark R. Liles, Lynn L. Williamson, Jitsupang Rodbumer, Vigdis Torsvik, Robert M. Goodman, and Jo Handelsman.** 10, New Brunswick, New Jersey : Applied and Environmental Microbiology, 2008, Vol. 74.
14. *Strategies for accesing soil metagenome for desired applications.* **J. Rajendhran, P. Gunasekaran.** Elsevier, Madurai, India : Biotechnology Advances, 2008, Vol. 26.
15. *The art and design of functional metagenomis screens.* **Marcus Taupp, Keith Mewis and Steven J Hallam.** Elsevier, British Columbia, Canada : Current Opinion in Biotechnology, 2011, Vols. 22: 465-472.
16. *Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors.* **Unden, G. and Bongaerts, J.** Mainz, Germany : Biochimica ET Biophysica Acta, 1997, Vol. 1320.

17. *Methods for identification of recombinants of phage λ.* **Brigitte Sanzey, Odile Mercereau, Therese Ternynck, Philippe Kourilsky.** 10, pp. 3394-397, Department of Molecular Biology, Institut Pasteur, 75015 Paris, France : Proc. Nati. Acad. Sci. USA, October 1976, Vol. 73.
18. *Plasmids of Escherichia coli as Cloning Vectors.* **Keith-Backman, Francisco-Bolivar.** s.l. : Methods in Enzymology, 1979, Vol. 68. ISBN 0-12-181968-X.
19. Molecular Cell Biology. 4th edition. [book auth.] Berk A, Zipursky SL, et al. Lodish H. *Molecular Cell Biology. 4th edition.* New York : W.H. Freeman and Company, 2000.
20. *Structure and Function of the F Factor and Mechanism of Conjugation.* **Neville Firth, Karin Ippen-Ihler, Ronald A. Skurray.**
21. *Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector.* **Hiroaki Shizuya, Bruce Birren, Ung-Jin Kim, Valeria Mancino, Tatiana Slepak,Yoshiaki Tachiiri, Melvin Simont.** pp. 8794-8797,, California Institute of Technology, Pasadena, CA 91125 : PNAS, September 1992, Vol. 89. PMCID: PMC50007.
22. *In Vitro Packaging of λ and Cosmid DNA.* **Hohn, Barbara.** s.l. : Methods in Enzymology, 1979, Vol. 68. ISBN 0-12-181968-X.
23. *Escherichia coli Plasmids Packageable in Vitro in λ Bacteriophage particles.* **Collins, John.** s.l. : Methods in Enzymology, Vol. 68. ISBN 0-12-181968-X.
24. *Construction of a gorilla fosmid library and its PCR screening system.* **Kim, Choong-Gon, Fujiyama, Asao and Saitou, Naruya.** 5, Mishima and Tokyo; Japan : Genomics, 2003, Vol. 82.
25. *Stable propagation of cosmid sized human DNA inserts in an F factor based vector.* **Ung-Jin Kim, Hiroaki Shizuya, Pieter J.de Jong', Bruce Birren and Melvin l.Simon*.** No. 5 1083-1085, Pasadena and Livermore, CA, USA : Nucleic Acids Research, 1992, Vol. 20.
26. *To BAC or not to BAC: marine ecogenomics.* **Béjá, Oded.** Elsevier, Haifa, Israel : Current opinion in Biotechnology, 2004, Vols. 15: 187-190.
27. *Metagenome microarray for screening of fosmid clones containing specific genes.* **Park, Soo-Je, Kang, Cheol-Hee and Rhee, Jong-Chan Chae & Sung-Keun.** New Brunswick, NJ, USA : Blackwell Publishing Ltd, 2008.
28. *Metagenomic gene discovery: past, present and future.* **Don Cowan, Quinton Meyer, William Stafford, Samson Muyanga, Roy Cameron and Pia Wittwer.** 6, Cape Town, South Africa : Trends in Biotechnology, 2005, Vol. 23.
29. *Functional Genomics: Expression Analysis of Escherichia coli Growing on Minimal and Rich Media.* **Tao, Han and Christoph Bausch, Craig Richmond, Frederick R. Blattner and Tyrrell Conway.** 20, Madison, Wisconsin : Journal Of Bacteriology, 1999, Vol. 181.
30. *Metagenomics and industrial applications.* **Patrick Lorenz, and Jurgen-Eck.** s.l. : Nature Reviews/Microbiology, 2005, Vol. 3.
31. *Biotechnological potential of coffee pulp and coffee husk for bioprocesses.* **Ashok Pandey, Carlos R. Soccol, Poonam Nigam, Debora Brand, Radjiskumar Mohan, Sevastianos Roussos.** s.l. : Biochemical Engineering Journal, 2000, Vol. 6.
32. *Phenolic compounds in coffee pulp: Quantitative determination by HPLC.* **Ramirez-Martinez, Jose R.** 2, San Crist6ba1, Venezuela : Journal of the Science of Food and Agriculture, 1988, Vol. 43.
33. *Outlook for cellulase improvement: Screening and selection strategies.* **Y.-H. Percival Zhang, Michael E. Himmel, Jonathan R. Mielenz.** Blacksburg, VA, USA : Biotechnology Advances, 2006, Vol. 24.
34. **Nele Ilmberger, Wolfgang R. Streit.** Screening for cellulase-encoding clones in metagenomic libraries. [book auth.] Rolf Daniel Wolfgang R. Streit. *Metagenomics: Methods and Protocols.* s.l. : Methods in Molecular Biology, 2010.
35. *Three Microbial Strategies for Plant Cell Wall Degradation.* **Wilson, David B.** New York, USA : Annals of the New York Academy of Sciences, 2008.

36. *Characterization of a gene encoding cellulase from uncultured soil bacteria.* **Soo-Jin Kim, Chang-Muk Lee, Bo-Ram Han, Min-Young Kim, Yun Soo Yeo, Sang Hong Yoon, Bon-Sung Koo, Hong-Ki Jun.** Suwon and Busan, Korea : Blackwell Publishing Ltd., 2008.
37. *Characterizattion of a metagenome-derived halotolerant cellulase.* **S. Voget, H.L. Steele, W.R. Streit.** Duisburg, Germany : Journal of Biotechnology, 2006, Vol. 126.
38. *Molecular cloning of glycoside hydrolase family 9 cellulase gene from buffalo rumen.* **Rungrattanakasin B., K. Jirajaroenrat, K. Maneewan and Chaowarat, M.** Bangkok, Thailand : @011 International Conference on Bioscience, Biochemistry and Bioinformatics, 2011, Vol. 5.
39. *Molecular and biochemical characterization of two xylanase-encoding genes from Cellulomonas pachnodae.* **Cazemier AE, Verdoes JC, van Ooyen AJ, and Op den Camp HJ.** Nijmegen, The Netherlands. : Applied Environmental Microbiology, 1999, Vol. 65.
40. *Profile of native cellulosomal proteins of Clostridium cellulovorans adapted to various carbon sources.* **Morisaka H, Matsui K, Tatsukami Y, Kuroda K, Miyake H, Tamaru Y, Ueda M.** Kyoto, Japan. : AMB Express, 2012, Vol. 2.
41. *High production of cellulose degrading endo-1,4-β-d-glucanase using bagasse as a substrate from Bacillus subtilis KIBGE HAS. .* **Bano S, Qader SA, Aman A, Syed MN, Durrani K.** Karachi, Pakistan. : Carbohydrate polimers, 2013, Vol. 91.
42. *Processive Endoglucanase Active in Cristaline Cellulose Hydrolysis by the Brown Rot Basidiomycete Gloeophyllum trabeum.* **Roni Cohen, Melissa R. Suzuki, Kenneth E. Hammel.** 5, Madison, Wisconsin : Applied and Environmental Microbiology, 2005, Vol. 71.
43. *Bacteria and lignin degradation.* **Jing LI, Hongli YUAN, and Jinshui YANG.** 1, Beijing, China : Front. Biol., 2009, Vol. 4.
44. *Fungal biodegradation and enzymatic modification of lignin.* **Mehdi Dashtban, Heidi Schraft, Tarannum A. Syed, Wensheng Qin.** 1, Ontario, Canada : Int J Biochem Mol Biol, 2010, Vol. 1.
45. *On the bacterial degradation of lignin.* **Janshekar H., Fiechter A.** Zurich, Switzerland : European J Appl Microbiol Biotechnol, 1982.
46. **Cullen D., Kersten P. J.** Enzymology and Molecular Biology of Lignin Degradation. [book auth.] R. Brambl & G.A. Marzluf. *Biochemistry and Molecular Biology, 2nd Edition.* Heidelberg, Berlin : Springer-Verlag, 2004.
47. *Specific and global regulation of genes associated with the degradation of aromatic compounds in bacteria.* **Gerischer, Ulrike.** 2, Ulm, Germany : J. MOl. Microbiol. Biotechnol. Review, 2002, Vol. 4.
48. *Lignin Biodegradation with Ligninolytic Bacterial Strain and Comparison of Bacillus subtilis and Bacillus sp. Isolated from Egyptian Soil.* **Abd-Elsalam, Hassan E., El-Hanafy, Amr A.** 1, Alexandria, Egypt : American-Eurasian J. Agric. & Environ. Sci., 2009, Vol. 5.
49. *Degradation of caffeine by Pseudomonas alcaligenes CFR 1708.* **Babu V.R. Sarath, Patra S. Thakur M.S., Karanth N.G., Varadaraj M.C.** Mysore, India : Enzyme and Microbial Technology, 2005, Vol. 37.
50. *Biological detoxification of cofee husk by filamentous fungi using a solid state fermentation system.* **Brand Débora, Pandey Ashok, Sevastianos Roussos, Carlos R. Soccol.** Marseille, France : Enzyme and Microbial Technology, 2000, Vol. 27.
51. *Physiology, biochemistry and possible applications of microbial caffeine degradation.* **Gummadi Sathyanarayana N., Bhavya B., Ashok Nandhini.** Chennai, India : Appl Microbiol Biotechnol, 2011.
52. *Microbial and enzymatic methods for the removal of caffeine.* **Gokulakrishnan S., Chandraraj K., Gummadi Sathyanarayana N.** Chennai; India : Enzyme and Microbial Technology, 2005, Vol. 37.
53. *Catabolic pathways and biotechnological applications of microbial caffeine degradation.* **Dash Swati Sucharita, Gummadi Sathyanarayana N.** Chennai; India : Biotechnol Lett, 2006.
54. *Production of Acinetobacter radioresistens lipase using Tween 80 as the carbon source.* **Chen-You Li, Chu-Yuan Cheng, Teh-Liang Chen.** Tainan, Taiwan : Enzyme and Microbial Technology , 2001, Vol. 29.

55. *Use of metagenomic approaches to isolate lipolytic genes from activated sludge.* **Ren-Bao Liaw, Mei-Ping Cheng, Ming-Che Wu, Chia-Yin Lee.** Taipei, Taiwan : Bioresource Technology, 2010, Vol. 101.
56. *Highly soluble expression and molecular characterization of an organic solvent-stable and thermotolerant lipase originating from the metagenome.* **Xinjiong Fan, Xiaolong Liu, Kui Wang, Sidi Wang, Rui Huang, Yuhuan Liu.** Guangzhou, China : Journal of Molecular Catalysis B: Enzymatic, 2011, Vol. 72.
57. *A thermostable esterase from Thermoanaerobacter tengcongensis opening up a new family of bacterial lipolytic enzymes.* **Lang Rao, Yanfen Xue, Cheng Zhou, Jin Tao, Gang Li, Jian R. Lu, Yanhe Ma.** Beijing, China : Biochimica el Biophysica Acta, 2011, Vol. 1814.
58. *Isolation and functional expression of a novel lipase gene isolated directly from oil-contaminated soil.* **Kaijing Zuo, Lida Zhang, Hongyan Yao, Jin Wang.** Sichuan, China : Acta Biochimica Polonica, 2010.
59. *Biodiesel production with special emphasis on lipase-catalyzed transesterification.* **Prakash S. Bisen, Bhagwan S. Sanodiya, Gulab S. Thakur, Rakesh K. Baghel, G. B. K. S. Prasad.** Gwalior, India : Biotechnology Letters, 2010, Vol. 32.
60. *Biodiesel production via esterification reactions catalyzed by lipase.* **A. P. de A. Vieira, M. A. P. da Silva, M. A. P. Langone.** Rio de Janeiro, Brazil : Latin American Applied Research, 2006, Vol. 36.
61. *Novel, Highly Specific N-Demethylases Enable Bacteria To Live on Caffeine and Related Purine Alkaloids.* **Ryan M. Summers, Tai Man Louie, Chi-Li Yu, Lokesh Gakhar, Kailin C. Louie, and Mani Subramanian.** Iowa, USA : Journal of Bacteriology, 2011, Vol. 194.
62. *Discovery and Characterization of Heme Enzymes from Unsequenced Bacteria: Application to Microbial Lignin Degradation.* **Margaret E. Brown, Mark C. Walker, Toshiki G. Nakashige, Anthony T. Ivarone, and Michelle C. Y. Chang.** Berkeley, Califronia : Journal of the American Chemical Society, 2011, Vol. 133.
63. *Metagenomic approach for the isolation of a novel low temperature active lipase from uncultured bacteria of marine sediment.* **Fredrik Hardeman, and Sara Sjoling.** Stockholm, Sweden : FEMS Microbiol Ecol, 2007, Vol. 59.
64. *Molecular cloning and characterization of a novel family VIII alkaline esterase from a compost metagenomic library.* **Yong Ho Kim, Eun Ju Kwon, Sung Kyum Kim, Yu Seok Jeong, Jungho Kim, Han Dae Yun, Hoon Kim.** Chinju, Republic of Korea : Biochemical and Biophysical Research Communications, 2010, Vol. 393.
65. *Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements.* **Bernard Hallet, David J. Sherratt.** 2, Oxford, UK : FEMS Microbiology Reviews, 1997, Vol. 21.
66. *PiggyBac Transposon Mutagenesis: A Tool for Cancer Gene Discovery in Mice.* **Roland Rad, Lena Rad, Wei Wang.** s.l. : Science, 2010, Vol. 330. 1095-9203.
67. *Mechanism and regulation of Mg-chelatase.* **Willows, Caroline J. Walker and Robert D.** Clemson and Providence, U.S.A. : Biochemical Journal, 1997, Vol. 327.
68. *Exoribonuclease and Endoribonuclease Activities of RNase BN/RNase Z Both Function In Vivo.* **Tanmay Dutta, Arun Malhotra and Murray P. Deutscher.** Miami, Florida; USA : The American Society for Biochemistry and Molecular Biology, 2012.
69. *Purification and Characterization of the tRNA-processing Enzyme RNase BN.* **Colleen Callahan, Doris Neri-Cortes, and Murray P. Deutscher.** 2, Miami, Florida; USA : The journal Of Biological Chemistry, 2000, Vol. 275.
70. *Interaction of tRNA (uracil-5-)-methyltransferase with NO2Ura-tRNA.* **Xiangrong Gu, Akira Matsuda, Kathryn M. Ivanetich, and Daniel V. Santi.** 6, California, San Francisco : Nucleic Acids Research, 1996, Vol. 24.

71. *Structure and Mechanism of the Glycerol-3-Phosphate Transporter from Escherichia coli.* **Yafei Huang, M. Joanne Lemieux, Jinmei Song, Manfred Auer, Da-Neng Wang.** New York, USA : Science, 2003, Vol. 301.
72. *The Central Enzymes of the Aspartate Family of Amino Acid Biosynthesis.* **Viola, Ronald E.** 5, Toledo, Ohio : Accounts Of Chemical Research, 2001, Vol. 34.
73. *The truA gene of Pseudomonas aeruginosa is required for the expression of type III secretory genes.* **Kyung-Seop Ahn, Unhwan Ha, Jinghua Jia, Donghai Wu and Shouguang Jin.** Gainesville, Florida, USA; Korea and Shanghai, China : Microbiology, 2004, Vol. 150.
74. *Activity and Transcriptional Regulation of bacterial Protein-Like Glycerol-3-Phosphate Dehydrogenase of the Haloarchaea in Haloferax volcanii.* **Katherine S. Rawls, Jonathan H. Martin, Julie A. Maupin-Furlow.** Gainesville, Florida; USA : Journal of Bacteriology, 2011, Vol. 193.
75. *NMR Studies Uncover Alternate Substrates for Dihydrodipicolinate Synthase and Suggest that Dihydrodipicolinate Reductase is also a Dehydratase.* **Sean R. A. Devenish, John W. Blunt, and Juliet A. Gerrard.** Cambridge, United Kingdom : Journal of Medicinal Chemistry, 2010, Vol. 53.
76. *Interaction of Pyridine Nucleotide Substrates with Escherichia coli Dihydrodipicolinate Reductase: Thermodynamics and Structural Analysis of Binary Complexes.* **Sreelatha G. Reddy, Giovanna Scapin and John S. Blanchard.** 41, Bronx, New York : Biochemistry, 1996, Vol. 35.
77. *Functional and Metabolic Effects of Adaptive Glycerol Kinase (GLPK) Mitants in Escherichia coli.* **M. Kenyon Applebee, Andrew R. Joyce, Tom M. Conrad, Donald W. Pettigrew and Bernhard O. Palsson.** California, USA : The journal of Biological Chemistry, 2011, Vol. 286.
78. *Synthesis and evaluation of conformationally restricted inhibitors of aspartate semialdehyde dehydrogenase.* **Cox, Andrew S. Evitt and Russell J.** s.l. : The Royal Society of Chemistry, 2011, Vol. 7.
79. *Discovery of the ammonium substrate site on glutamine synthetase, a third cation binding site.* **Shwu-Huey Liaw, Ichun Kuo, and David Eisenberg.** Taipei, Taiwan : The Protein Society, 1995, Vol. 4.
80. *Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methyl propan-1-ol production at theoretical yield in Escherichia coli.* **Sabine Bastian, XiangLiu,JosephT.Meyerowitz, ChristopherD.Snow, Mike M.Y.Chen, Frances H.Arnold.** Pasadena,CA, USA : Metabolic Engineering, 2011, Vol. 13.
81. *Identification and characterization of a novel fumarase gene by metagenome expression cloning from marine microorganisms.* **Chengjian Jiang, Lan-Lan Wu, Gao-Chao Zhao, Pei-Hong Shen, Ke Jin, Zhen-Yu Hao, Shuang-Xi Li, Ge-Fei Ma.** Guangxi, China : Microbial Cell Factories, 2010.
82. *Conformational changes upon ligand binding in the essential class II fumarase Rv1098c from Mycobacterium tuberculosis.* **Ariel E. Mechaly, Ahmed Haouz, Isabelle Miras, Nathalie Barilone, Patrick Weber, William Shepard , Pedro M. Alzari, Marco Bellinzoni.** Paris cedex 15, France : Febs Letters, 2012, Vol. 586.
83. *Identification of an acetolactate synthase small subunit gene in two eukaryotes.* **Duggleby, Ronald G.** Brisbane, Australia : Gene, 1997, Vol. 190.
84. *Expression, purification and preliminary crystallographic analysis of Rv3002c, the regulatory subunit of acetolactate synthase (IlvH) from Mycobacterium tuberculosis.* **Jiang Yin, Grace Garen, Craig Garen and Michael N. G. James.** Alberta, Canada : Structural Biology and Crystallization Communications, 2011, Vol. F67. ISSN 1744-3091.
85. *GCN5-Related N-Acetyl Transferases: A Structural Overview.* **Fred Dyda, David C. Klein, and Alison Burgess Hickman.** Bethesda, Maryland : Annu. Rev. Biophys. Biomol. Struct., 2000, Vol. 29.
86. *Crystal Structure of Gamma-Glutamyl Phosphate Reductase (TM0293) From Thermotoga maritima at 2.0 Å Resolution.* **Rebecca Page, Michael S. Nelson, Frank von Delft, Marc-André Elsliger, Jaume M. Canaves.** La Jolla, California : PROTEINS: Structure, Function, and Bioinformatics, 2004, Vol. 54.

87. *Ferrous iron transport protein B gene (feoB1) plays an accessory role in magnetosome formation in Magnetospirillum gryphiswaldense strain MSR-1.* **Chengbo Rong, Yijun Huang, Weijia Zhang, Wei Jiang, Ying Li, Jilun Li.** Beijing, China : Research in Microbiology, 2008, Vol. 159.
88. *Biochemical characterisation of aconitase from Corynebacterium glutamicum.* **Bott, Meike Baumgart and Michael.** Jülich, Germany : Journal of Biotechnology, 2011, Vol. 154.
89. *Kinetic Analyses of the Magnesium Chelatase Provide Insights into the Mechanism, Structure, and Formation of the Complex.* **Willows, Artur Sawicki and Robert D.** 46, New South Wales, Australia : The Journal of Biological Chemistry, 2008, Vol. 283.
90. *Chelatases: distort to select?* **Salam Al-Karadaghi, Ricardo Franco, Mats Hansson, John A. Shelnutt, Grazia Isaya, and Gloria C. Ferreira.** 3, Lund, Sweden : Trends Biochem Sci. , 2006, Vol. 31.
91. *Current understanding of the function of magnesium chelatase.* **J. D. Reid, C. N.Hunter.** 4, Sheffield, U.K. : Biochemical Society Transactions, 2002, Vol. 30.
92. *Magnesium chelatase from Rhodobacter sphaeroides: initial characterization of the enzyme using purified subunits and evidence for a Bchl-BchD complex.* **Lucien C. D. Gibson, Poul Erik Jensen and C. Neil Hunter.** Sheffield, UK : Biochem. J., 1999, Vol. 337.
93. *Thioredoxin Redox Regulates ATPase Activity of Magnesium Chelatase CHLI Subunit and Modulates Redox-Mediated Signaling in Tetrapyrrole Biosynthesis and Homeostasis of Reactive Oxygen Species in Pea Plants.* **Tao Luo, Tingting Fan, Yinan Liu, Maxi Rothbart, Jing Yu, Shuaixiang Zhou, Bernhard Grimm and Meizhong Luo.** Berlin, Germany : Plant Physiology, 2012, Vol. 159.
94. *Degradation of indulin, a kraft pine lignin, by Serratia marcescens.* **MOHANRAJ MANAngeeswarani, Vijayanandraj V. Ramalingam, Karthik Kumar.** Chennai, India : Journal of Environmental Science and Health Part B , 2007, Vol. 42.
95. *The structure of full-length LysR-type transcriptional regulators. Modeling of the full-lenght OxyR transcriptional factor dimer.* **Kierzek, Jolanta Zaim and Andrzej M.** Warsaw, Poland : Nucleic Acids Research, 2003, Vol. 31.
96. *Molecular Biology of the LysR Family of transcriptional regulators.* **Schell, Mark A.** Athens, Georgia : Annu. Rev. Microbiol., 1993.
97. *Regulation of the Salmonella enterica std Fimbrial Operon by DNA Adenine Methylation, SeqA, and HdfR.* **Marcello Jakomin, Daniela Chessa, Andreas J. Ba¨umler, and Josep Casadesu´s.** 22, Seville, Spain : Journal of Bacteriology , 2008, Vol. 190.
98. *H-NS-Dependent Regulation of Flagellar Synthesis Is Mediated by a LysR Family Protein.* **Park, Minsu Ko and Chankyu.** 16, Taejon, Republic of Korea : Journal of Bacteriology, 2000, Vol. 182.
99. *HdfR is a regulator in Photorhabdus luminescens that modulates metabolism and symbiosis with the nematode Heterorhabditis.* **Catherine A. Easom, and David J. Clarke.** 4, Cork, Ireland. : Environmental Microbiology, 2012, Vol. 14.
100. *Interaction of Pyridine Nucleotide Substrates with Escherichia coli Dihydrodipicolinate Reductase: Thermodynamic and Structural Analysis of Binary Complexes.* **Sreelatha G. Reddy, Giovanna Scapin, and John S. Blanchard.** Bronx, New York : Biochemistry, 1996, Vol. 35.
101. *Tetrahydrodipicolinate N-Succinyltransferase and Dihydrodipicolinate Synthase from Pseudomonas aeruginosa: Structure Analysis and Gene Deletion.* **Robert Schnell, Wulf Oehlmann, Tatyana Sandalova, Yvonne Braun, Carmen Huck, Marko Maringer, Mahavir Singh, Gunter Schneider.** 2, Braunschweig, Germany : PLoS ONE, 2012, Vol. 7.
102. *Cloning of the DapB Gene, Encoding Dihydrodipicolinate Reductase, from Mycobacterium tuberculosis.* **Martin S. Pavelka JR, Torin R. Weisbrod, and William R. Jacobs JR.** 8, Bronx, New York : Journal of Bacteriology, 1997, Vol. 179.
103. *Mechanisms of acetohydroxyacid synthases.* **David M Chipman, Ronald G Duggleby and Kai Tittmann.** Beer-Sheva, Israel : Current Opinion in Chemical Biology, 2005, Vol. 9.

104. *Acetohydroxyacid synthase and its role in the biosynthetic pathway for branched-chain amino acids.* **J. A. McCourt, and R. G. Duggleby.** Brisbane, Australia : Amino Acids, 2006, Vol. 31.
105. *A novel β-glucosidase with lipolytic activity from a soil metagenome.* **Cheng-Jian Jiang, Gao Chen, Jie Huang, Qin Huang, Ke Jin, Pei-Hong Shen, Jun-Fang Li, and Bo Wu.** Guangxi University, China : Folia Microbiol, 2011, Vol. 56.
106. *Comparative Modeling of the Three-Dimensional Structures of Family 3 Glycoside Hydrolases.* **Andrew J. Harvey, 1 Maria Hrmova, Ross De Gori, Joseph N. Varghese, and Geoffrey B. Fincher.** Glen Osmond, South Australia, Australia : Proteins: Structure, Function, and Genetics, 2000, Vol. 40.
107. *Lipase-catalyzed regioselective acylation of diosgenyl saponins.* **Biao Yu, Guowen Xing, Yongzheng Huia, and Xiuwen Hanb.** Shanghai, China : Tetrahedron Letters, 2001, Vol. 42.
108. *Glycerol: a promising and abundant carbon source for industrial microbiology.* **Gervásio Paulo da Silva, Matthias Mack, Jonas Contiero.** Senhor do Bonfim, Brazil : Biotechnology Advances, 2009, Vol. 27.
109. *Activity and Transcriptional Regulation of Bacterial Protein-Like Glycerol-3-Phosphate Dehydrogenase of the Haloarchaea in Haloferax volcanii.* **Katherine S. Rawls, Jonathan H. Martin, and Julie A. Maupin-Furlow.** 17, Gainesville, Florida : Journal of Bacteriology, 2011, Vol. 193.
110. *Microbial engineering for the production of advanced biofuels.* **Pamela P. Peralta-Yahya, Fuzhong Zhang, Stephen B. del Cardayre, and Jay D. Keasling.** Berkeley, California : Nature, 2012, Vol. 488.
111. *Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli.* **Sabine Bastian, XiangLiu, Joseph T.Meyerowitz, Christopher D. Snow, Mike M.Y. Chen.** Pasadena, Ca : Metabolic Engineering, 2011, Vol. 13.
112. *Metagenomic Analysis of Streptomyces lividans Reveals Host-Dependent Functional Expression.* **Matthew D. McMahon, Changhui Guan, Jo Handelsman, and Michael G. Thomasa.** 10, New Haven, Connecticut : Applied and Environmental Microbiology, 2012, Vol. 78.
113. **Kristof Vrancken, Lieve Van Mellaert, and Jozef Anné.** Cloning and Expression Vectors for a Gram-Positive Host, Streptomyces lividans. [book auth.] and Rolf Daniel Wolfgang R. Streit. *Metagenomics: Methods and Protocols.* s.l. : Humana Press, 2010.
114. *Construction of integrative plasmids suitable for genetic modification of industrial strains of Saccharomyces cerevisiae.* **Fernanda Cristina Bezerra Leite, Rute Salgues Gueiros dos Anjos, Anna Carla Moreira Basilio, Guilherme Felipe Carvalho Leal, Diogo Ardaillon Simões, Marcos A. de Morais Jr.** Recife, PE, Brazil : Plasmid, 20012.
115. *New and Redesigned pRS Plasmid Shuttle Vectors for Genetic Manipulation of Saccharomyces cerevisiae.* **Haase, Mark K. Chee and Steven B.** Durham, North Carolina : Genes/Genomes/Genetics, 2012, Vol. 2.
116. *Microorganism communities and chemical characteristics in sludge-bamboo charcoal composting system.* **L, Hua, et al.** Xi'an, China : Environmental Technology, 2011, Vol. 32.
117. *Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene.* **Stone NE, Fan J-B, Willour V, Pennacchio LA, Warrington JA, Hu A, Chapelle A, Lehesjoki A-E, Cox DR, Myers RM.** 3, s.l. : Genome Research, 1996, Vol. 6. doi:10.1101/gr.6.3.218. PMID 8963899..
118. **Daum, Hank.** *Email answer to a question I made to Hank Daum.* s.l. : Epicentre Biotechnologies, October 22 2009.
119. *Purification and Characterization of Fumarase from Corynebacterium glutamicum.* **Tomoko Genda, Shoji Watabe, and Hachiro Ozaki.** Yamaguchi, Japan : Biosci. Biotechnol. Biochem, 2006, Vol. 70.
120. *Mechanism and regulation of Mg-chelatase.* **Caroline J. Walker, and Robert D. Willows.** Providence, USA : Biochem. J. , 1997, Vol. 327.
121. *Effects on Gram-Negative and Gram-Positive Bacteria Mediated by 5-Aminolevulinic Acid and 5-Aminolevulinic Acid Derivatives.* **Nicolas Fotinos, Maruska Convert, Jean-Claude Piffaretti, Robert**

Gurny, and Norbert Lange. 4, Geneva, Switzerland : Antimicrobial Agents and Chemotherapy, 2008, Vol. 52.