Biocatalytic activity screening for the advanced glycation end product (Nε-Carboxymethyl-Lysine) in metagenomics libraries

Biocatalytic Activity Screening for the Advanced Glycation End Product (N^{ε}-Carboxymethyl-Lysine) in Metagenomics Libraries

By

Edgar F. Ferrer González

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

MASTER OF SCIENCE
IN
BIOLOGY

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2015

Approved by:

_____	_____
Juan C. Martínez Cruzado, Ph. D Member graduate committee	Date
_____	_____
Patricia Ortiz Bermúdez, Ph. D Member graduate committee	Date
_____	_____
Carlos Ríos Velázquez, Ph. D President graduate committee	Date
_____	_____
Martha L. López Moreno, Ph.D. Representative, Office of Graduate Studies	Date
_____	_____
Nannette Diffoot Carlo, Ph. D Chairperson of the Department	Date

ABSTRACT

Throughout the years, our body tends to accumulate a great variety of small molecules that are incapable of being metabolized or eradicated from our system. As the concentration of these molecules increases, they start having ill effects on our health and are often related to a great variety of aging diseases. For instance, the Advanced Glycation End product known as **Nε-(carboxymethyl)-Lysine** (CML) plays an important role in the development of secondary pathologies associated with diabetes. The CML has also been found to induce steatosis in non-alcoholic fatty liver. Even though the production of CML can be inhibited, the small molecule remains present in the affected tissue and crossed-linked proteins. Thus, the development of enzymatic based treatments that could remove substantially the accumulated CML would be highly beneficial to our society. In order to find a biocatalytic agent for CML, we implemented functional metagenomics by screening on minimal medium supplemented with CML. Four metagenomic libraries were screened using this method: one from the sub-tropical rainforest, El Yunque, another one from the dry forest at Guánica, other one from benthic microbial mat and the last one from an ephemeral microbial mat found at Cabo Rojo. A total of 178 clones were isolated from all the libraries. The study was focused on the RBI-3 clone from El Yunque metagenomic library. Transposition mutagenesis was performed with Tn5 transposon in order to generate random mutants that could knock out the activity. Each mutant was sequenced using Sanger sequencing. The clone was fully sequenced using Illumina Mi-Seq platform and assembled by *de novo* assembly and hybrid Sanger-Illumina template assembly. Automated annotation was performed by the RAST pipeline, and manual annotations were made near the Tn5 insertion sites. Four predicted candidate genes were characterized via phylogenetic protein analysis due to the

proximity to the Tn5 insertion site. A predicted sulfite oxidase was used on a *trans* complementation assay, as result the experimental data demonstrated that the activity of this gene was rescued and granting *E. coli* EPI300 the ability to use CML as sole carbon source. These findings could lead to the development of *in vitro* assays and further on being applied to industrial applications to lower the levels of CML on food content or a clinical approach to medical bioremediation.

RESUMEN

A lo largo de los años, nuestro cuerpo tiende a acumular una gran variedad de pequeñas moléculas que son incapaces de ser metabolizadas o erradicadas de nuestro sistema. Al la concentración de estas moléculas aumentar, se comienzan a tener efectos negativos sobre la salud, que a menudo están relacionadas con una gran variedad de enfermedades asociadas al envejecimiento. Por ejemplo, La molécula derivada de los productos de glicación terminal avanzada conocida como **N^{ε}-(carboximetil)-Lisina** (CML) desempeña un papel importante en el desarrollo de patologías secundarias asociadas con la diabetes. CML también se ha encontrado que puede inducir la esteatosis en el hígado graso no alcohólico. A pesar de que la producción de CML se puede inhibir mediante tratamiento, la molécula pequeña permanece presente en el tejido afectado y las proteínas reticuladas. Por lo tanto, el desarrollo de tratamientos enzimáticos que podrían eliminar sustancialmente los niveles de CML acumulado sería de gran beneficio para nuestra sociedad. Con el fin de encontrar un algún agente biocatalítico para CML, se ha implementado la metagenómica funcional, mediante análisis de asimilación en medio mínimo suplementado con CML. Un total de cuatro bibliotecas de metagenómicas fueron monitoreadas utilizando este método. Las mismas provenían de suelo del bosque subtropical lluvioso de El Yunque, del bosque seco de Guánica, y las restantes de los tapetes microbianos bentónicos y los efímeros de las salinas de Cabo Rojo. Un total de 178 clones fueron aislados de todas las bibliotecas. El estudio se centró en el clon RBI-3 clon proveniente de la biblioteca metagenómica de El Yunque. Un análisis de mutagénesis por transposición se realizó con el transposón Tn5 con el fin de generar mutantiones aleatorias que podrían desactivar la actividad. Cada mutante fue secuenciado usando secuenciación de Sanger. El clon por otro lado fue totalmente secuenciado utilizando la plataforma Mi-Seq de Illumina y ensamblado mediante ensamble *de novo* y ensamble

híbrido entre Sanger e Illumina por ensamblaje de molde. Además se realizó una anotación automatizada por el sistema de anotación automático, RAST, mientras que las anotaciones manuales se realizaron para aquellos genes de interés cerca de los sitios de inserción del Tn5. Un total de cuatro genes candidatos se predijeron y fueron caracterizados mediante análisis filogenético de proteínas. El gen anotado como oxidasa de sulfito fue utilizado para una prueba de complementación genética *in trans*. Como resultado, los datos experimentales demostraron que este gen está involucrado con la capacidad de *E. coli* EPI300 de utilizar CML como única fuente de carbono, restableciendo la habilidad perdida en el mutante. Estos hallazgos podrían conducir al desarrollo de ensayos *in vitro* y al desarrollo de aplicaciones industriales para reducir los niveles de CML en los alimentos tanto como un enfoque clínico de bioremediación médica.

DEDICATION

First, I would like to dedicate this work to a very special person in my life, my aunt Marta González Santiago (tití Marta). When everything seems to fall apart, you were there for me. You always cared for me and my brother no matter how hard our situation was. You are my inspiration to move on forward with my life. Thank you for all your support and for always believing in your nephew.

To my mother, Sibia F. González Santiago, I wouldn't be in this world in the first place if it weren't for you. We been through very hard moments together, and yet you were always at home awaiting my arrival. You had sacrificed a lot, and I hope that one day I would be able to pay you back. Thank you for all your unconditional love.

To my dear older brother, Frank X. Ferrer González (Franxo), I love you and I wish the best for you.

To my late father, Frank Ferrer Almodovar, even though at the end we didn't had a good relationship, you always told me to persuade my dreams and overcome them no matter what. Rest in Peace.

Lastly I would like to dedicate this work to all of those people that one way or another had suffered from diabetes or ever taken care of a person who does.

In 1871, Louis Pasteur said: *"Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world."*

Thanks!

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Carlos Ríos Velázquez, for granting me the opportunity of working under your mentorship. I'm very grateful to have you as my mentor, and for believing on my crazy ideas that made this whole chapter of my life possible. Thank you all for your guidance, you had become like a father to me.

My members of my graduate committee, Dr. Juan C. Martínez Cruzado and Dr. Patricia Ortiz Bermúdez. For your inspiration and guidance, giving me the tools and understanding required to accomplish my goals. You guys are the best!

To my lab mates, Moises, Wilmer, Albin, Karlen, Jesie, Laura and Robert for treating me as family and supporting me all the way until the end. My brother, Frank X. Ferrer González, at first I didn't expect you to last so long on the lab, but I'm grateful you did. It was nice having you as my lab mate and seeing that you finally found your path in life, we been trough much this past few years and yet you stayed firmed on your path, you will always be my best friend.

To Goro, Irimar, Kristina and Josue, I will never forget all of your valuable lessons, thank you all sparing your time and instructing on the lab. I really appreciate all your help and lessons.

My non-blood brother, Arnaldo J. Rodriguez Rivera, for being there always when I needed it the most. Thank you for accompanying me on this journey. And to my non-blood little sister, Stephanie Silva L Del Toro, where would I be standing right now if it weren't for you, thank you.

All of my department colleagues, Veronica, Emily, Christian, Jennifer, Eduardo, José, Yashira, Priscila, Miguel, Eduardo, Ingrid, Edna, Stephanie, Rubén, Coralis, Panda, Gordo, Steven, Mel, Derick and Dianiris. Thank you all for the support! It has been the most amazing journey so far.

To the department of biology: Dr. Nanette Diffoot, Dr. Jaime Acosta, Dr. Luis Ríos, Dr. Mildred Chaparro, Dr. Dimuth Siritunga, Dr. Alejandro Ruiz, Dr. Dimaris Acosta, Dr. Vivian Navas, Dr. Rafael Montalvo and Dr. Carlos Rodríguez Mingüela. You were all part of my integral academic life. Thank you for all your teachings and for your friendship.

My students, Ruth Melany Perez, Lidimarie Trujillo and William Rodriguez, I learned so much from you three, by being your mentor. I wish the best for you all.

To the different federal agencies that one way or another funded my research, The Department of Energy (DE-FG09-07SR22571); The United States Department of Agriculture (USDA-HSI-CSREES 2007-02386); The National Science Foundation (NSF-RUI 0455620); The National Institutes of Health training grant, RISE2BEST (NIH-R25GM088023); The National Institutes of Health training grant, BRIDGE TO THE DOCTORATE (NIH-R25GM058389); and to the non-profit organization Howard Hughes Medical Institute (HHMI-52006318).

To Mr. Donato Seguí, for being a good friend and teaching me the importance of professionalism during my term as your TA and how to impact my students. And Mrs. Gladys Toro, for allowing myself to grow as a professional and for sharing your knowledge with me. I couldn't had ask for better coordinators to work with, you both gave me valuable life lessons.

To Mrs. Magaly Zapata, not only you provided me support as a lab technician, but you also became a great friend, thank you for all the conversations we had and the moments we shared.

Lastly, to all of my friends and family. Thank you for believing in me and for understanding my absence while I engaged myself in this journey.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Metabolism is an essential process for life, where molecules break down or form new structures. However, when our biocatalyst system fails to manage these molecules properly, a series of diseases can manifest. Some toxic molecules form as part of our natural glutathione pathway thus faces eventually biocatalytic insufficiency, this is the case of Advance Glycation End Products (AGEs) (Kislinger et al., 1999). This group has been referred to as aging factors due to the pathology they cause, thus earning the name of age-related molecules. Since our body lacks the biocatalytic activity to resolve the accumulation of AGEs, the use of microbial origin enzymes has been proposed -a process known as medical bioremediation (Schloendorn et al., 2009). Although promising current limitations to medical bioremediation includes the fact that less than one percent of the microorganisms present in the environment are considered cultivable under laboratory conditions (Riesenfeld, et al., 2004; Schloss & Handelsman, 2005). Thus, there is an imperative need to address this problem, to explore microbial enzymes that could be elaborated into a plausible treatment. Culture independent techniques such as functional metagenomics could be employed to provide alternative protocols of biotechnological drug therapeutics development without the need for isolating pure cell cultures. As a result, the process will elaborate innovating processes that will impact how we solve current medical challenges.

The usage of metagenomic libraries as resources for novel drug discoveries from uncultivable microorganisms has proven to be successful for high throughput screenings of diverse small molecules capable of being of used for cancer treatment in contrast with pure cell culture (Singh & Macdonald, 2010). Even though there are certain limitations involved regarding this technique such as gene expression profile and physiological adaptability to store, export or resist the expressed product, the study of these communities of genomic material has been proven

to serve as a biocatalytic reservoir of previously unknown functions (Allen et al., 2009; Schloss & Handelsman, 2003), thus increasing the chances of finding novel bioprospects with therapeutics ends in contrast with pure culture of microorganisms.

A possible application is the finding of a biocatalytic agent capable of metabolizing an AGE byproduct, particularly N^{ε}-(carboxymethyl)-Lysine (CML). This molecule has been associated with the development of diseases ranging from blindness, vascular stiffness, kidney failure and non-alcoholic fatty liver (Gaens et al., 2012; Kamata et al., 2009; Pollreisz & Schmidt-Erfurth, 2010). The common aspects of these pathologies are that they progress through time. Diseases that follow this pattern are known as age-related diseases (Sell & Monnier, 2011). Making them an attractive target for the development of enzymatic therapies capable of reducing the amount of CML in the system and its adverse effects.

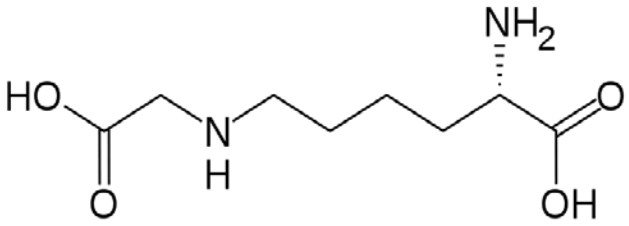

Figure 1. *Representation of the chemical structure of N^{ε}-(carboxymethyl)-Lysine*

Here we explored the microbial community present at the subtropical rainforest El Yunque – Dwarf forest (also known as Cloud Forest or Elvin Woods), an extreme environment with precipitation rate of over 250 inches per year (U.S. Department of Agriculture, 2013). Given the forest conditions and the highly functional diversity from the microbial communities present, the probabilities of capturing novel pathways will be greater than other places of interest. Eventually, this would add up to a biocatalytic reservoir against CML, which could be employed as potential enzyme therapy by remediating the toxic AGE product from a cellular level. Eventually,

contributing to the development of the medical bioremediation therapy for age-related diseases (Mathieu et al., 2009).

1.2 LITERATURE REVIEW

1.2.1 Aging and oxidative stress theory

Aging is characterized by the degeneration of anatomical integrity and function across multi-organ systems susceptible to stress. As the systems degenerate, we see increases in pathologies, diseases commonly known as age-related diseases which eventually induce a higher risk of death (Semba et al., 2010). However, the exact reasons why does this processes occur are still not well known. One of the most well-accepted theories from the scientific community is known as the oxidative stress theory. Reactive oxygen species (ROS) are constantly being formed as the result of cellular respiration held within the mitochondria. As the mitochondria fail through time, ROS are then released into the cytosol prompting a cellular senescence response that could lead to cell death (Romano et al., 2010; Vicencio et al., 2008). This biological process tends to cause an augmentation systematic failure leading to the aging phenotype, and different strategies can be taken to reduce the amount of cellular damage.

However, not all sources of ROS are derived directly from the mitochondria. Partial reduction of O_2 catalyzed by different metabolic pathways. For example, fatty acid oxidation that occurs in the peroxisomes, yields production of hydrogen peroxide (H_2O_2) catalyzed by acyl-CoA oxidase. Other ROS producing pathways can be found in the endoplasmic reticulum during the protein folding process and cytosolic pathways that involve enzymes such as amino acid oxidases, cyclo-oxygenase, lipid oxygenase, and xanthine oxidase that eventually lead to the production of H_2O_2 (Hernández-García et al., 2010). All of these previously mentioned pathways usually occur endogenously in the cell.

1.2.2 Advanced Glycation End Products

When an external variable enters in the equation, we find that different physiological reactions are capable of inducing a ROS reaction. For example, prolonged periods of inflammation can cause permanent cellular damage by denaturing the intracellular proteins, rendering the cell ability to withstand a stress response and eventually triggering cell death. Within these external factors, we find that a group known as Advanced Glycation End (AGE) products are responsible for a series of diseases highly related with oxidative stress. We associate the pathologies expressed by this family of molecules to an increment of vascular permeability, increment of vascular stiffness, inhibition of vascular dilatation by interfering with the nitric oxide cycle, oxidized LDL, binding of cells including macrophage, endothelial and mesangial cells resulting in the liberation of cytokines and enhancement of the oxidative stress reaction (Yan et al., 2007). As result, many diseases known today had been found to contain a direct relationship associated with the presence of AGE's. Such as atherosclerosis, neuropathy, nephropathy, non-alcoholic fatty liver, rheumatoid arthritis and others (Castellani et al., 2001; Goh & Cooper, 2008; van Dam et al., 2013; Yan et al., 2007).

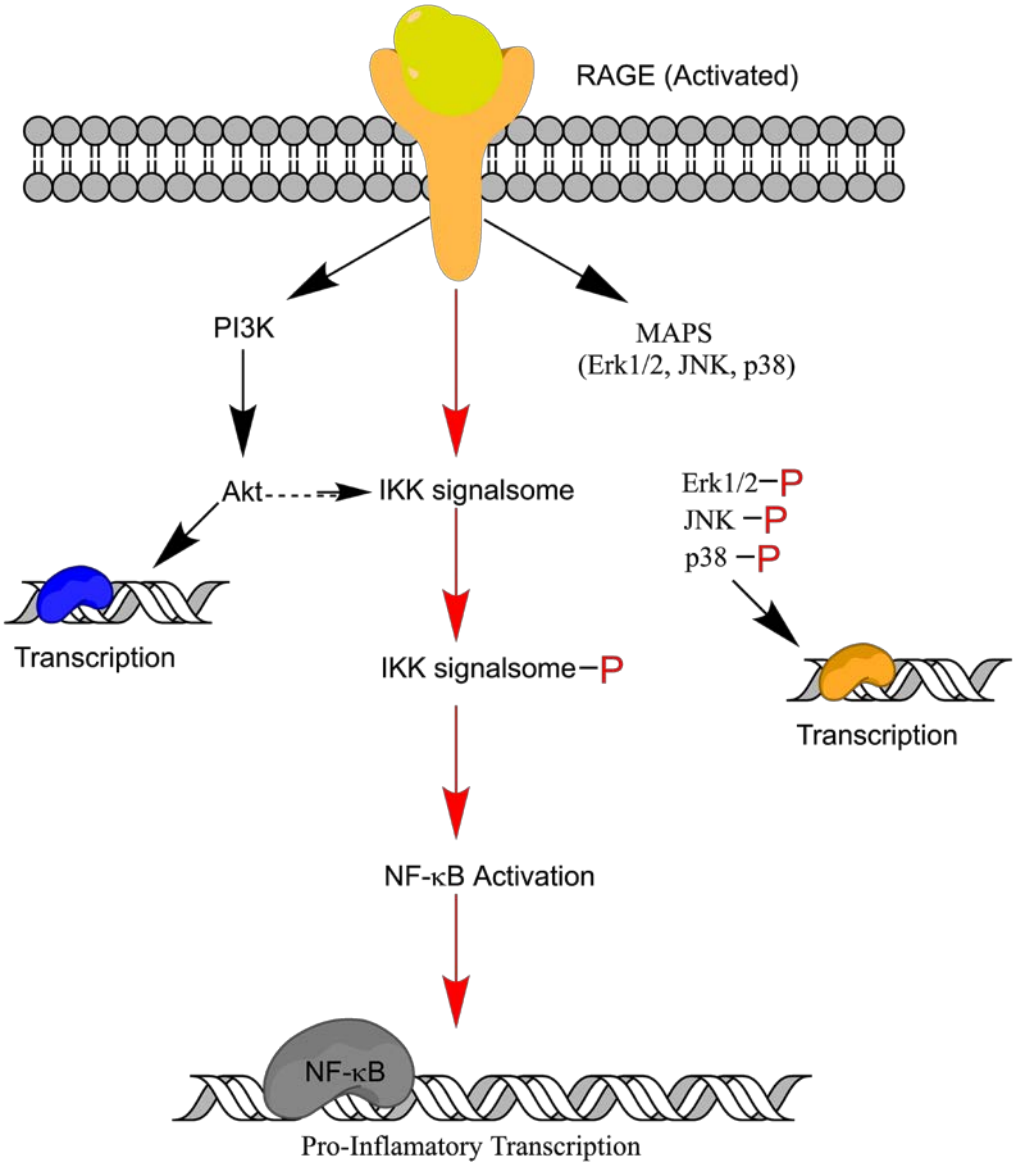

Figure 2. Proposed pathways of RAGE receptor activation, when associated with diseases. According to the ligand origin and tissue location, it would activate one of the three proposed pathways shown in this figure. The PI3K pathway is known to be an initiator of cellular death by apoptosis. MAPS pathway plays an important role in intracellular communication that would eventually regulate the activation or deactivation of several kinases. As a result, transcriptional information can be related to the production of cytokines, apoptosis, and others. Lastly, the NF-κB is one of the most studied pathways in RAGE activation due to its pro-inflammatory response (Lin et al., 2009).

Many AGE products have been studied in the pathologies mention above, and for years they were considered only biomarkers for progression of diseases, indicating that oxidative stress

damage had been done (Roy et al., 2012; Schleicher et al., 1997). In recent years more sensitive techniques and equipment became available, since then it has been found that these molecules also play an important role in the development of certain pathologies instead of just reporting the damage as it was previously considered to be its function. The mechanism of action currently known for this family of compounds varies with tissue and source of AGE. In Figure 2, we can see the three principal and most studied pathways that the cells carry on when the receptor for AGE (RAGE) gets activated (Lin et al., 2009). Out of the possible ligands associated with RAGE, N^{ε}-Carboxymethyl-L-Lysine (CML) has been one of the most studied molecules from this family, due to the diversity of pathologies on where it has been be found and associated.

CML is a unique molecule since it can affect the system by various mechanisms such as protein crosslinking, activation of RAGE and protein denaturalization (Mizutari et al., 1997). However, its effect is not noticeable until a high concentration of CML is present and is hence denominated an aging molecule (Knecht et al., 1991; Schleicher et al., 1997). Another factor that grants it a position within the aging molecules family is the way it is produced within the body. Endogenous production of CML can contain several initiating pathways as shown in figure 3, with the Glycation pathway being one of the most important routes since it is a non-enzymatic reaction that requires only the presence of the sugar and protein in order to form CML trough a reaction known as the Maillard Reaction (Li et al., 2012; Monnier et al., 2005). The Maillard Reaction consists of the production of Schiff Bases and Amadori products in the presence of a sugar and a protein.

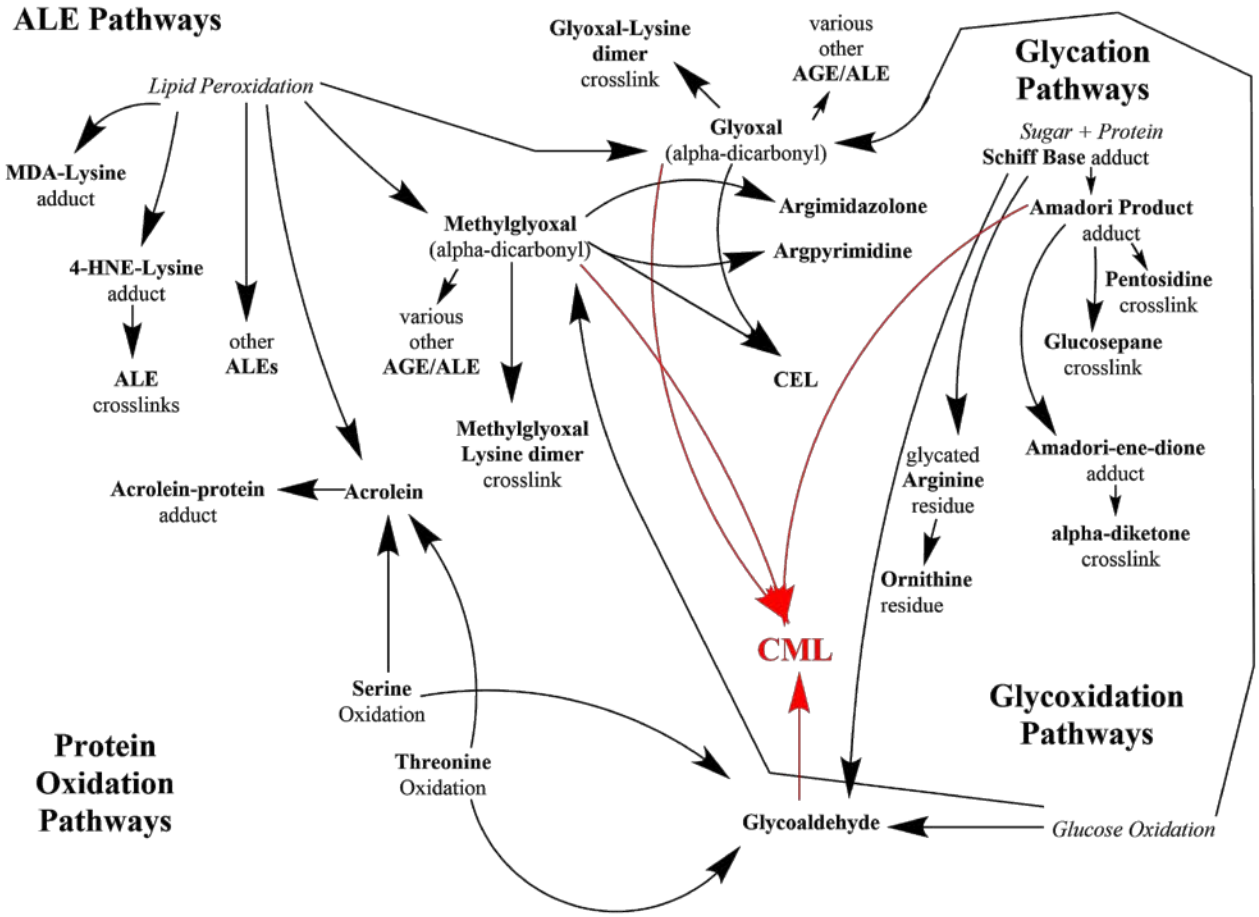

Figure 3. Pathways directly related to the production of Nε-Carboxymethyl-L-Lysine.
Red arrows indicate precursors required for the formation of CML from each different pathway, being glycoxidation and glycation pathways the one most studied (Furber, 2010).

There are other sources of incorporating CML into a human body. Recent studies specialized in the western hemisphere diet which involves principally baking at high temperature or deep frying. Have demonstrated that within this diet the amount of CML produced by a Maillard reaction is considerable high enough to induce a direct effect on human health by consuming for a prolonged period the food elaborated from the processes described above (Hull et al., 2012; Lee et al., 2013).

1.2.3 Pathologies Related to CML

According to the Centers for Disease Control and Prevention (CDC) the leading causes of death in the United States for the 2010 report are found to be age-related non-communicable diseases (NCD). Within the report, we find that leading NCD's are heart diseases, cancer, chronic lower respiratory diseases, cerebrovascular diseases, Alzheimer's and diabetes (Murphy et al., 2013). CML can affect various systems at once, however it's more common to diabetic patients or people who suffer from hyperglycemia, due to the high amount of sugar available in the blood that would eventually be auspicious for interacting by the Maillard reaction (Goh & Cooper, 2008; Randhir Singh et al., 2014). The most common pathologies associated with this population involves severe neuropathy, retinopathy (Randhir Singh et al., 2014; van Dam et al., 2013) and nephropathy (Forbes et al., 2002; Li et al., 2012; Singh et al., 2001; Yan et al., 2007). Both of these conditions involves and prolong inflammatory response associated with the activation of RAGE affecting the NF-κB pathway.

The outcomes are not solely associated with diabetics and hyperglycemia patients. For example, the steatosis in non-alcoholic fatty liver is the result of an uncontrollable inflammation event induced by the activation of the NF-κB pathway. CML has been demonstrated to interact with the RAGE in hepatic cells using model develop to simulate long term exposure of CML in a short period of time. As consequence of the activation of this pathway, the hepatocytes affected eventually will scarified the tissue providing a loss of function in the liver tissue (Gaens et al., 2012; Lin et al., 2009; Toth et al., 2008).

Recently the American Heart Association published a research article indicating that the interaction of the CML ligand with RAGE has a direct impact on obesity associated with insulin resistances due to dysregulation of inflammatory adipokines (Gaens et al., 2014). Other conditions associated with CML is an early development of atherosclerosis plaques (Lin et al., 2009; Schleicher et al., 1997) and vascular arterial stiffness (Goh & Cooper, 2008).

1.2.4 Current treatments

There are no clinical treatments currently available to reduce or stop the production of CML in the body. However, several chemical reactions that inhibit the production of CML *in vitro* have been described. Isoferulic acid, is a compound extracted from *Cimicifuga heracleifolia*, an herbal medicine, commonly used in oriental regions such as China and Japan, known for its anti-inflammatory properties. The mechanism of action attack the formation of CML by preventing the cleavage of amadori products, methylglyoxal, glyoxal and 3-deoxyglucoson in both fructose and glucose glycation pathways (Meeprom et al., 2013). The other currently known mechanisms impacts directly the food industry, by switching from high temperatures or deep frying to microwave heating which have been shown to prevent the formation of Schiff bases and thus the formation of CML.

1.2.5 Medical bioremediation approach

For years, microorganisms have proven to be of endless resources fullness that range from food and drinks and all the way up to pharmaceuticals and environmental tools. Their diverse metabolisms have turned them into the ultimate environmental cleansing machines, breaking down natural and xenobiotic pollutants (Karigar & Rao, 2011). Given that the human body is a unique environment, scientists proposed to treat certain enzyme deficiency rendering the patient incapable

of breaking down toxic molecules that are hard to eliminate from our body or simply not capable of being metabolized by supplementing microbial origin enzymes, and thus the concept of medical bioremediation began (Schloendorn et al., 2009). Eventually the field expanded into a medically related issue and thus experiments began in order to find biocatalytic elements capable of breaking down toxic molecules found in a variety of diseases related to aging (Mathieu et al., 2009). A concurrent result is that a cultivable microorganism from activated sludge was found to metabolize 7-ketocholesterol, a oxysterol known to play an important role in the development of a great variety of cardiac pathologies, especially in the formation of foam cells within an artherosclerosis plaque (Mathieu et al., 2008). Besides this, *Mycoplasma hyorhinis* one of the smallest life forms have been found to bes capable of degrading Amyloid Beta (A-β) aggregates on cell culture plates. When antibiotics were supplied the A-β accumulation was restored (Zhao et al., 2008).

1.2.6 Functional metagenomics

However, the current publications associated with medical bioremediation are all limited by the same factor. Most of their candidates are solely microorganisms capable of being cultivated in a laboratory environment. This limits our available resources since the estimated percent of cultivable microorganisms is <1% (Schloss & Handelsman, 2005). In order to explore the vast resources that microorganisms have to offer, the implementation of culture-independent techniques such as metagenomics has been very favorable, acting as a gateway to the uncultivable microbial population (Handelsman, 2004). Moreover, uncultivable microorganisms have been proven to be a vast reservoir for novel treatments and industrial applications (Couto et al., 2010; Schloss & Handelsman, 2003; Singh & Macdonald, 2010). For example, novel lipases have been discovered from metagenomic libraries derived from a mangrove ecosystem (Couto et al., 2010) and different soils types (Henne et al., 2000). These lipases can play an important role in industrial

applications where different waste lipids are produced and proper handling is needed in order to reduce the environmental damage (Karigar & Rao, 2011). Along with the development of better sequencing tools, exploring new natural resources with current principles of natural drug assembly's machinery such as polyketide synthases (PKS) family have demonstrated that new drug screening from uncultivable microorganisms can be implemented as a fully automated high-throughput process obtaining new chemical modifications and properties without the need of expensive organic chemistry synthesis (Owen et al., 2013).

In contrast, with sequence based metagenomics that relies solely on predictions, functional metagenomics allows us to perform activity based screenings. According to the specific activity you are searching for, the proper screen protocol must be established. Some of these might involve enzymatic reactions that can be measured using a spectrometer (Alcaide et al., 2015) or by colorimetric tests (Sharma et al., 2014). Performing mutations (Uchiyama & Miyazaki, 2009) and transferring the library to other organisms by usage of shuttle vectors (Guazzaroni et al., 2014; Liebl et al., 2014; Troeschel et al., 2010) along with genetic complementation and sequencing have helped other researcher determine functional properties found in these libraries.

CHAPTER 2: SCREENING METAGENOMIC LIBRARIES FOR BIOCATALYTIC ACTIVITIES RELATED TO N^{E}-(CARBOXYMETHYL)-LYSINE

2.1 INTRODUCTION

Advances in synthetic biology have made it possible to engineer the way organisms function, transforming them into living organic machines. Along with the boom in recent genomics, gene discovery has become an essential task that involves endless possibilities as we explore new emerging fields. These genes could serve as a base for synthetic blocks to build microbial machinery more efficiently and improving many biocatalytic functions in high demand for agricultural, industrial, energy and medical applications (Fernández-Arrojo et al., 2010; Guazzaroni et al., 2014). This positions metagenomics in a very promising field of novel discoveries that help develop advances in the previously mentioned fields.

Screening metagenomics libraries for functional aspects have been an efficient way of discovering novel products in the last couple of years such as DNA Polymerase (Simon et al., 2009), β-lactamase (Riesenfeld, Goodman, et al., 2004), alcohol dehydrogenases (Wexler et al., 2005), nitrate reductase (Bartossek et al., 2012) and others. These discoveries were achieved using two distinctive screening mechanisms, one of the most widely used was high throughput DNA sequencing, thanks to the advances in next-generation-sequencing technology. By the posterior use of comparative genomics, one could identify plausible coding proteins based on conserved motifs. The other widely used screening approach was functional based.

Functional metagenomics is a cultured-independent approach that allows us to study genomic fragments from the environment and uncultured microorganism without relying on previous knowledge, unlike the sequence-driven analysis. Here we can find colorimetric assays that can help us indicate a specific activity, yet sometimes it can induce false positives. Then we got heterologous gene expression, which is one of the most powerful traits offered by functional metagenomics studies. This allows us to discover novel biological functions based on survivability

of the clone in the presented conditions reducing false positives (Riesenfeld, Schloss, et al., 2004; Wenzel & Müller, 2005). This opens a new way for drug discovery and therapeutic proteins by allowing us to screen clones for activity instead of relying on pure cultures, given that the conditions for expression can be met. In this case, we perform a screening selected for any activity related to the biocatalysis of N^{ε}-Carboxymethyl-L-Lysine, an oxidase amino acid linked to the development of many disease and long known biomarker for the progression long term diseases.

2.2 METHODOLOGY

Metagenomic Libraries Screening

Previously generated metagenomics libraries from the dwarf forest of El Yunque subtropical rain forest, Guánica dry forest (Cruz et al., 2010) and Cabo Rojo microbial mats (Torres-Zapata, 2012) from the Microbial Biotechnology and Bioprospecting Laboratory at UPRM, were employed in a carbon source assimilation test using M9 minimal medium, specifically design for recombinant strains of *E. coli*. The medium contained various variations of the recipe in order to provide proper controls since our host *E. coli* EPI300 strain requires the supplementation of leucine and thiamine due to deletions of these genes. The screening recipe was formulated on 1L of solid M9 minimal medium plates, and was followed as directed by Figueroa Matías, 2012. In a 1.5L container the following was incorporated: 200mL of M9 minimal medium salts 5X (BD Scientific), 2mL MgSO4 1M, 0.1mL $CaCl_2$ 1M, 10mL Leucine 10,000µg/mL, 20mL Thiamine 500µg/mL 10mL Chloramphenicol 1,500µg/mL, 200ml 2% Carbon Source, 557.9mL dH_2O and 15g of bacteriological grade agar.

For the no carbon source control group, the 2% carbon source solution was substituted with dH_2O. In order to rule out utilization of thiamine or leucine, the groups were divided into no carbon source added with either thiamine or leucine. Another control group contained a solution of 2%

glucose as carbon source along with the original recipe. Lastly, the experimental group contained a 200mL solution of CML at 2% for a final working concentration of 0.4% of CML per plate.

In order to screen the libraries, each library was grown overnight with the induction solution in Luria-Bertani (LB) Broth with 15μg/mL of chloramphenicol in an orbital shaker at 37°C, then washed using a physiological saline solution (0.85% NaCl) in order to remove traces of the medium that could generate false positives in the carbon source assimilation assay. After washing the cells, 100μL of the cell suspension was added to each plate and spread using 0.6mm glass beads. Initial incubation was held at 37°C for no more than 1 week. The same procedure was employed with *E. coli* EPI300 strain (Epicentre) containing the library vector (pCC1FOS) without any insert cloned into it. Clones of interest were stored using a 15% glycerol solution in a 1-2mL solution in an ultra-low temperature freezer (-80°C). For this, each clone was grown individually in LB Broth with 12.5μg/mL of chloramphenicol and cryogenized at log phase.

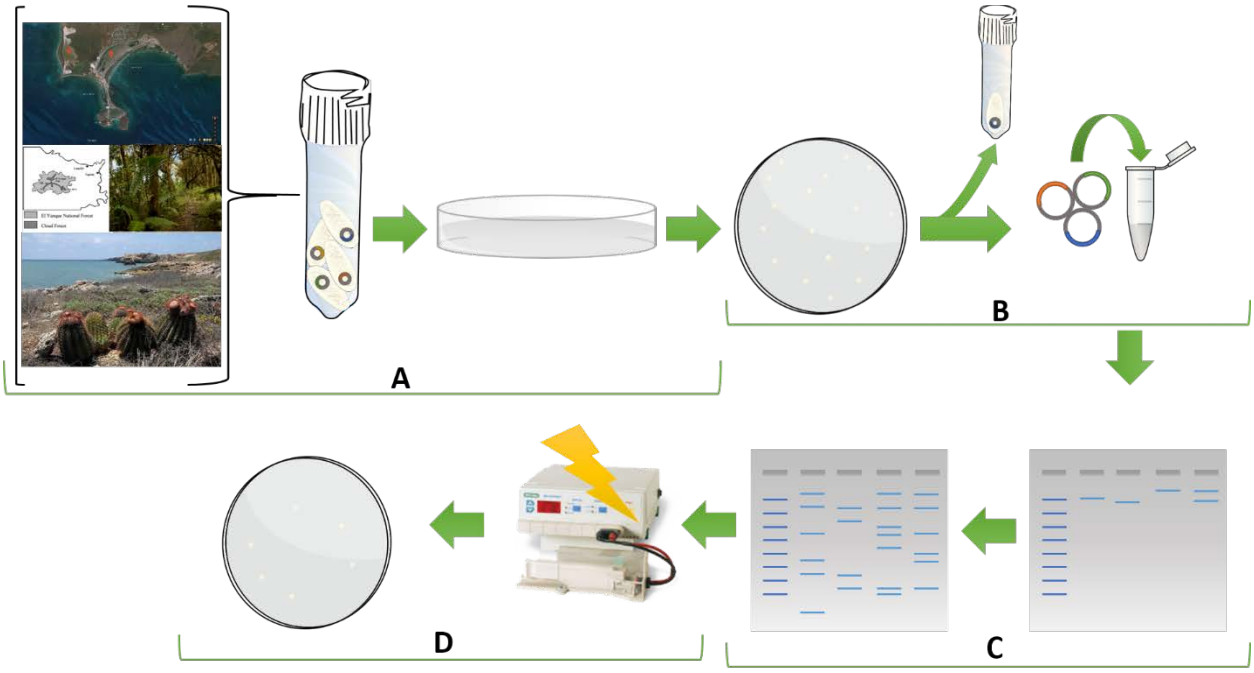

Figure 4. Flowchart based on the screening procedure. A, Different metagenomic libraries (Rain Forest, Dry Forest & Microbial Mat,) were screened over minimal medium (M9) supplemented with N^{ε}-(carboxymethyl)-Lysine (CML). *B, Positive clones were isolated and stored, followed by a miniprep extraction. C, fosmids were then ran in agarose gel follow by digestion using restriction enzymes. D, Confirming the activity based out of the insert, an extracted fosmid was inserted into an isogenic strain of E. coli EPI300 and served on minimal medium (M9) supplemented with CML.*

Positive clone's library "checkup."

Clones capable of growing are processed through a modified Qiagen miniprep protocol stated as follows. An overnight culture(s) was pelleted in a 2mL Eppendorf tube by spinning at 9296 rcf for 3 minutes. Then resuspend each in 250 µl of Buffer P1 (Qiagen) with RnaseA, followed by 250 µl of Buffer P2 (Qiagen). After the addition of the solution, it was mixed by inversion 4-6 times. This step will eventually lyse the cells and shouldn't be allowed to lyse for more than 5 minutes. In order to neutralize the reaction, 350 µl of Buffer N3 (Qiagen) was added and then mixed by inverting 4-6 times. During this step, a white precipitate is formed. Then in a fume hood and with proper safety precautions, 100 µl chloroform was added. The tube was inverted 4-6 times to mix followed by leaving the samples to rest on ice for 5-10 minutes. This

step is so that the chloroform does not get too hot in the centrifuges and leak out of the tubes. The samples were then centrifuged at maximun speed (18220rcf) for 10 minutes. At this step, an aqueous and the organic phase are seen. Then 750µl of supernatant/aqueous layer were transfered into a new 1.7mL Eppendorf tube. In order to precipitate the DNA, 1.0 volume of cold isopropanol was added and vortexed to mix the sample. The sample was left to rest over ice for 10-30 minutes, and then centrifuged at maximum speed for 25 minutes. The supernatant was discarded without disturbing the pellet. The sample was then washed by adding 500ml of 70% EtOH at room temperature, flick inverted 4-6 times to rinse the tube, and centrifuged at maximum speed for 5 minutes. The supernatant was discarded without disturbing the pellet and the EtOH wash was repeated in order to remove all traces of isopropanol. Later on, the pellet was air dried for around 15 minutes. Eventually a change in color (from white to clear) should be seen as the pellet dries. Finally, it was resuspended in 25µL of dH_2O or $T_{10}E_1$ buffer.

The extracted fosmid DNA was then digested using *Not*I (Promega) in a 25µL reaction as stated in the manufacturer protocol for 3 hours. *Briefly: 1µg of DNA was added, followed by 2.5µL of buffer, 5U of enzyme, and water until it reached a volume of 25µL.* The sample was then loaded with 5µL of 6X loading dye and ran in a 1% agarose gel at 70v for 120 minutes. The vector used for the libraries contains restriction sites for *Not*I around the vector insertion site. This allows easy separation of the vector from the insert.

Clones activity confirmation

In order to confirm that the activity comes from the cloned insert, a re-transformation of the extracted clone was performed. To do this, the fosmid was electroporated into an isogenic cell (Epicentre EPI300 Tranformax Electrocompetent cells), utilizing a BIORAD Micro Pulser™ with preset parameters for *E. coli* following the protocol provided by the electrocompetent cells kit

(Epicentre, 2011b). In this case, LB broth was used instead of Super Optimal broth with Catabolic repressor (SOC) for recovery and left two hours recovering. Cells were washed in physiological saline solution (0.85% NaCl) and 100µL of cells was spread with 0.6mm glass beads on the experimental plate (M9 enriched with CML). Incubation followed at 37°C for up to 7 days.

Determining the biocatalytic activity genes

One clone from the rainy bottom collection has been selected due to having lowest molecular size compared to the others for this procedure. The selected clone of interest was characterized by inactivating the gene(s) that provides the activity by performing an *in vitro* Tn5 transposon mutagenesis assay (EZ-Tn5™ KAN2 kit by Epicentre technologies). Mutagenesis was followed as instructed by the kit protocol with a starting material concentration modified. *Briefly: 2µg of DNA were used instead of 0.2µg in order to increase the successful rate of transformants during the electroporation. Followed by addition of 0.1pmol of EZ-Tn5 <KAN-2> Transposon (calculated inferring a size of, 42,000bp on target DNA), 1µl of EZ-Tn5 10X Reaction Buffer, 1µl of EZ-Tn5 Transposase and 5µl of ddH_2O and incubated at 37°C for 2h.* After denaturing the transposase, a 1 µl of solution was mixed with 50µl *E. coli* Epi300 electrocompetent isogenic cells and electroporated into the cells. Rapidly after the procedure 750µl of LB broth culture medium was added to the cells, then transferred into a 4.2mL LB broth test tube and placed at 37°C for 1.5 hours in the orbital shaker. Afterward, cells were plated in Petri dishes containing kanamycin at 50µg/mL and chloramphenicol at 15µg/mL for the selection of cells with resistance to antibiotics, the newly inserted transposon, and vector. Then colonies obtained using this procedure were patched into a new plate called master plate that contained LB agar, kanamycin, and chloramphenicol. The master plate was left incubating overnight at 37°C.

An experimental plate was prepared identically to the screening plate only that this time it was supplemented with kanamycin as well, containing both antibiotics. Replica plating was performed by pressing down the cloth onto the master plate followed by pressing it on the experimental plate and finally onto the control plate which contains only LB Agar with both chloramphenicol and kanamycin. Mutants that lost activity on the experimental plate (unable to grow by using the carbon source supplied in the minimal medium) were then selected and stored for sequencing using the end points of the transposon.

2.3 RESULTS

Screening & confirmation

Metagenomic libraries were screened using M9 minimal medium supplemented with CML. Incubation period lasted 9days; each 24h plate was checked to see if there was growth or not. The figure 5 shows different libraries screened along with the host cell containing the vector (pCC1FOS) without insert. Screening results after 7 days of incubation showed a total of 9 clones from El Yunque I, 4 clones from El Yunque II, no clones were isolated from Dry I due to contamination (all isolated selected contained no fosmid), Dry II no clones were found. The Microbial Mats libraries, Candelaria contained 137 clones, and Fraternidad obtained 28 clones.

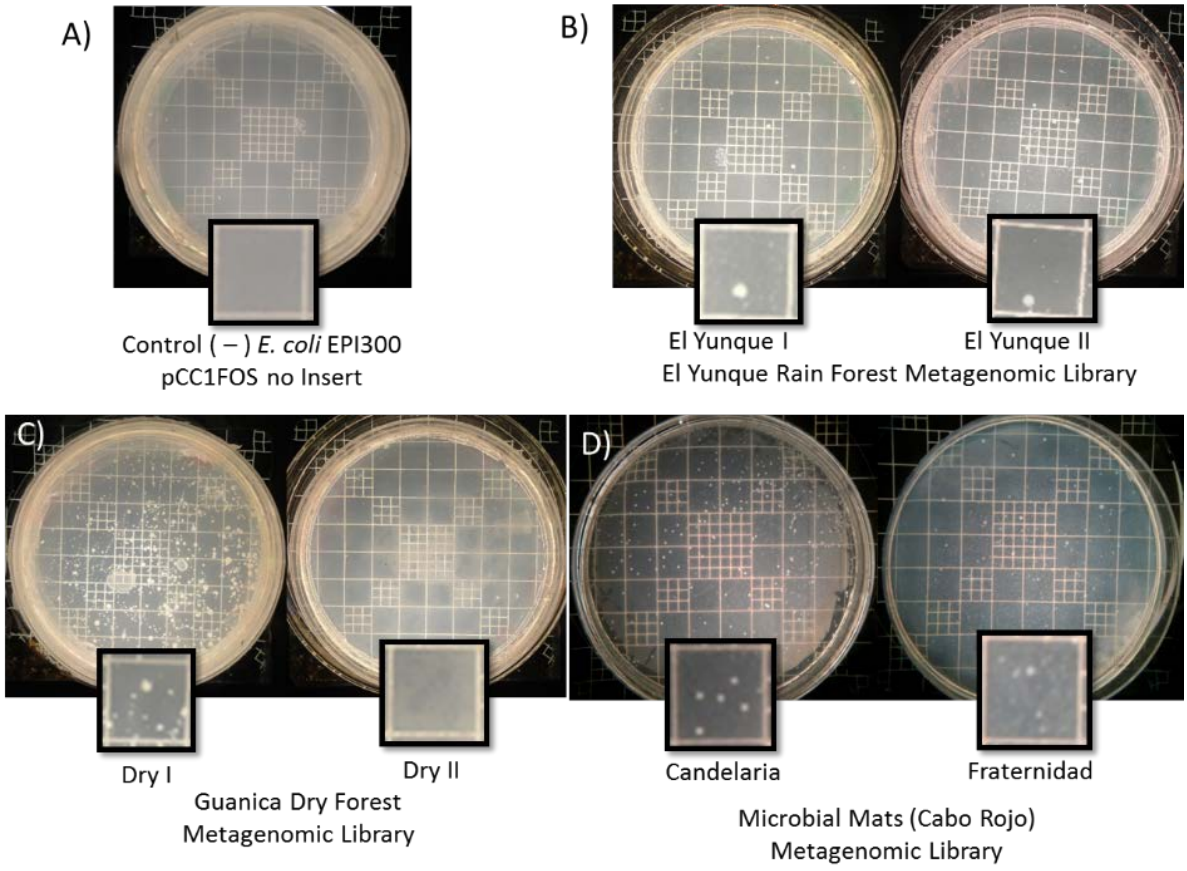

Figure 5. Metagenomic libraries screened for CML assimilation. A) Shows *E. coli* EPI300 unable to grow on this carbon source. B-D) shows clones from each metagenomic libraries after 7 days of incubation.

A total of 5 randomly selected clones from each library, El Yunque sub-tropical rainforest, Guanica's dry forest, and 4 for microbial mats, were used for miniprep extraction followed by digestion with *Not*I. Figure 6 shows clones digestion from El Yunque sub-tropical rainforest, each clone possess a different restriction pattern, indicating that each one is a unique clone. Figure 7 shows digestion obtained from the microbial mats library, All clones obtained from the Fraternidad sampling site possess only one pattern, while the clone from Candelaria sampling site possess an unique distinct pattern.

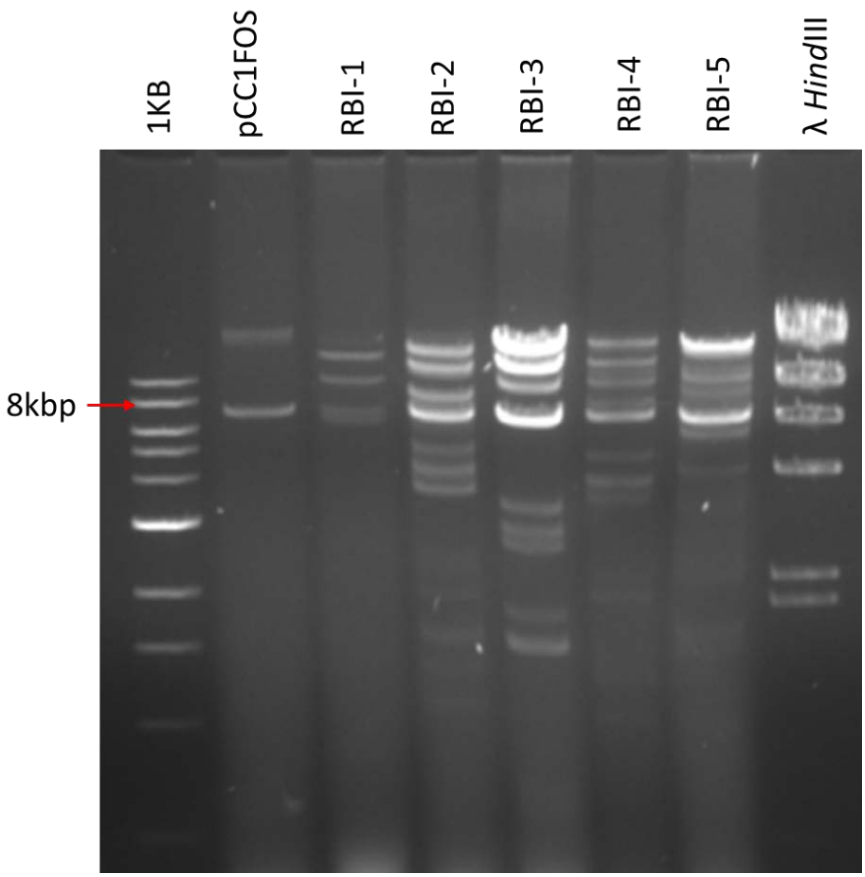

*Figure 6. Fosmid digestion of the El Yunque Sub-Tropical Rain Forest using Not*I *restriction enzyme.* The 8kb arrow indicates the presence of the vector pCC1FOS. RBI (1-5) Represent the selected clones from the metagenomic library.

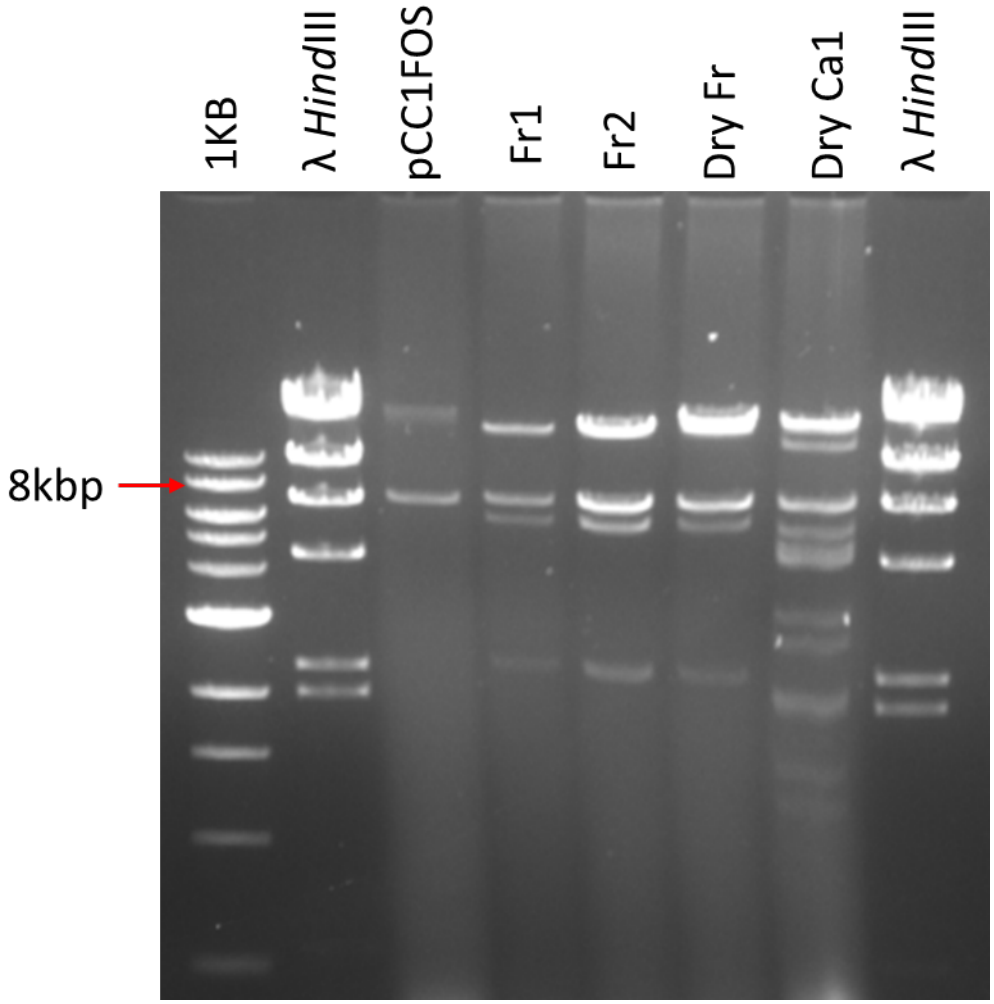

*Figure 7. Fosmid digestion of Candelaria and Fraternidad clones from microbial mats metagenomic libraries using Not*I *restriction enzyme.* The 8kb arrow indicates the presence of the vector pCC1FOS. Fr1 & Fr2 are from the Fraternidad wet season metagenomic library, Dry Fr is from the Fraternidad dry season metagenomic library, each contains the same restriction pattern. Dry Ca1 was obtained from the Candelaria dry season metagenomic library.

After the digestion, clone RBI-3 from the El Yunque library was used for re-transformation into an isogenic cell of *E. coli* EPI300. The same was done for clone Fr1, from the microbial mats. According to Figure 8 and 9, both clones were capable of inducing activity again indicating that the cloned inserts were responsible for the activity. The microbial mat re-transformation was patched instead of spread since they were grown first on LB agar plate with 12.5µg/mL of chloramphenicol.

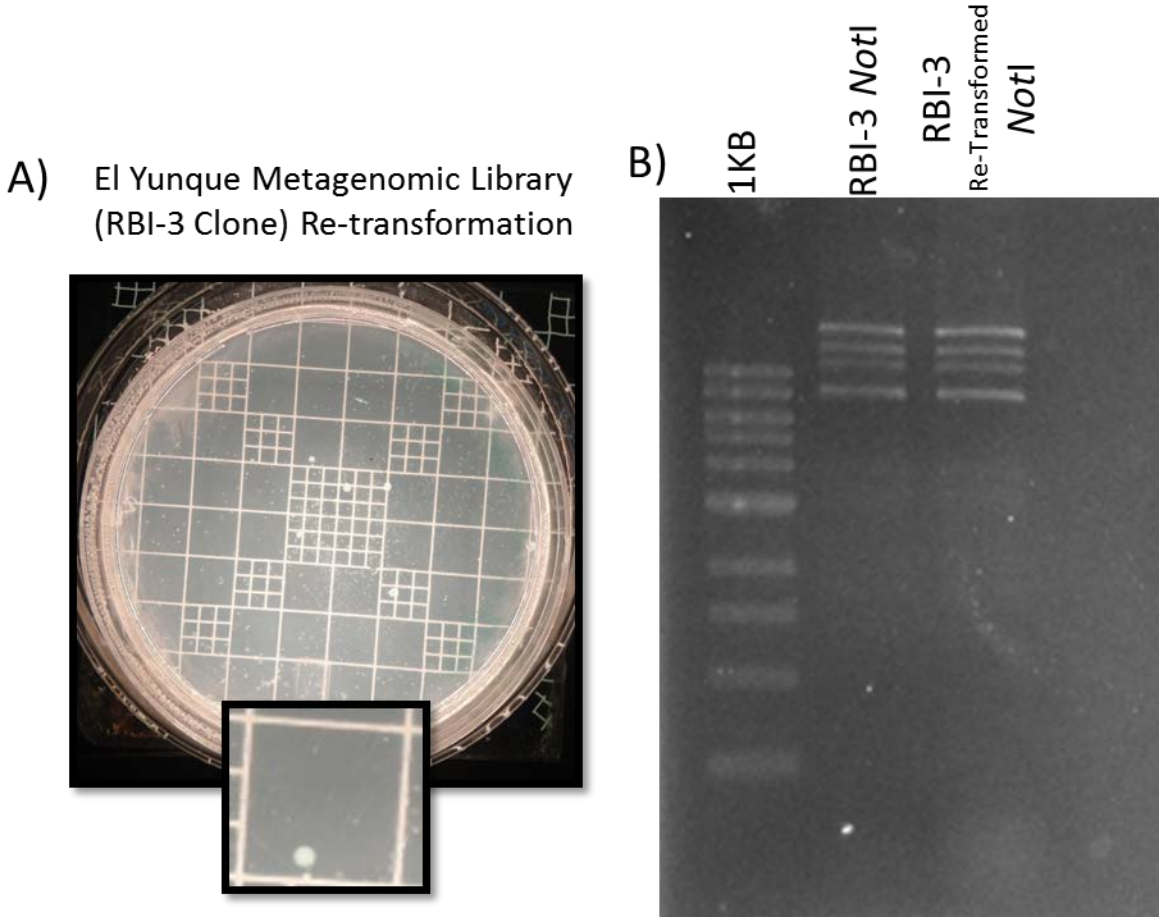

Figure 8. Fosmid re-transformation by electroporation of RB3 clone (El Yunque Metagenomic Library). A) Re-transformed clone grown on M9 supplemented with CML. B) Restriction analysis using *Not*I of the fosmid extracted from the original clone (RBI-3) and the re-transformed, both containing the same insert.

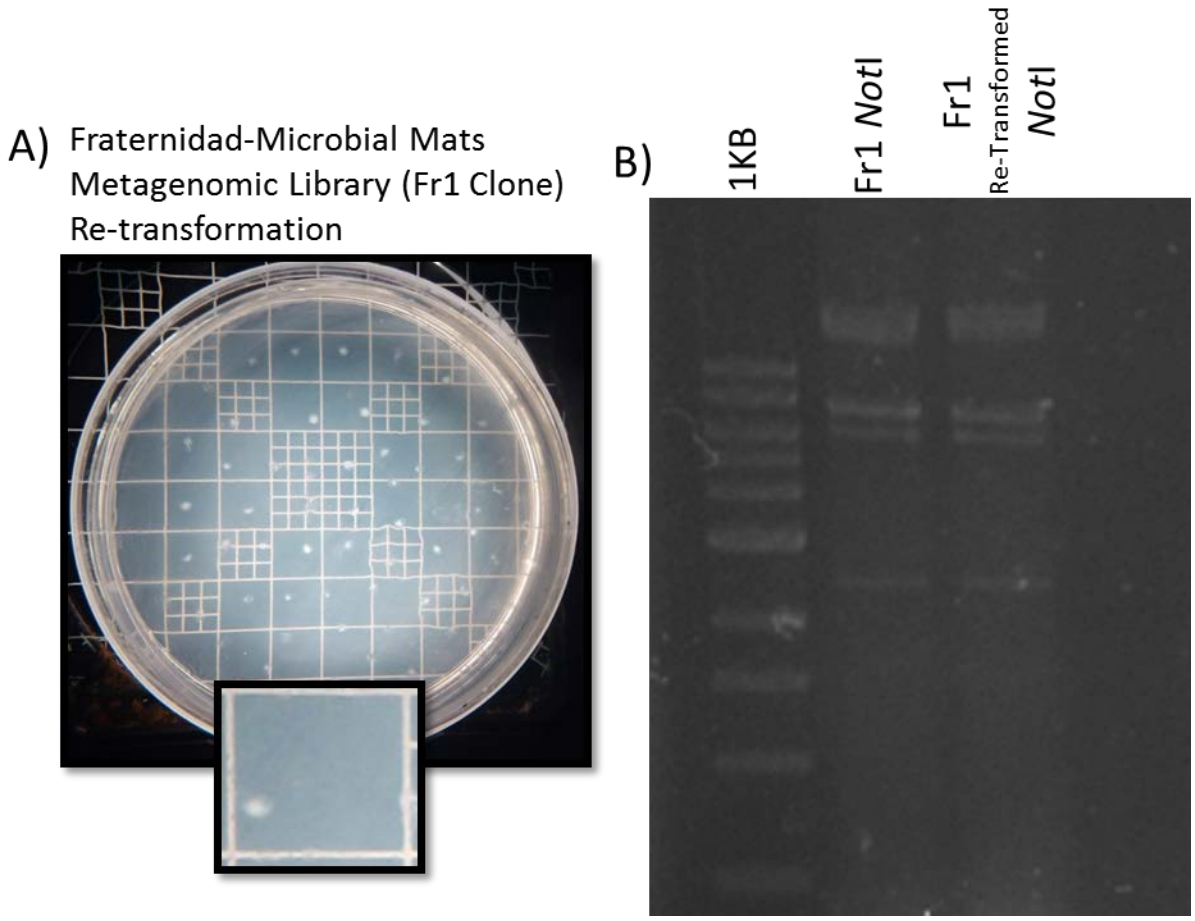

Figure 9. Fosmid re-transformation by electroporation of Fr1 clone (Fraternidad Cabo Rojo Microbial Mats). A) Re-transformed clone grown on M9 supplemented with CML. B) Restriction analysis using *Not*I of the fosmid extracted from the original clone (Fr1) and the re-transformed, both containing the same insert.

Mutagenesis by Tn5 Transposon

In order to understand the mechanics involved, the RB3 clone was selected for mutagenesis studies, due to its ability of growing on the minimal medium supplemented with CML for 48h, taking less time than the rest of the clones. After incubation period, colonies unable to grow on the experimental plate were stored cryogenically and prepared for capillary sequencing on triplicates. A total of 35 out of 36 clones selected from the transformation resulted in activity loss (figure 10).

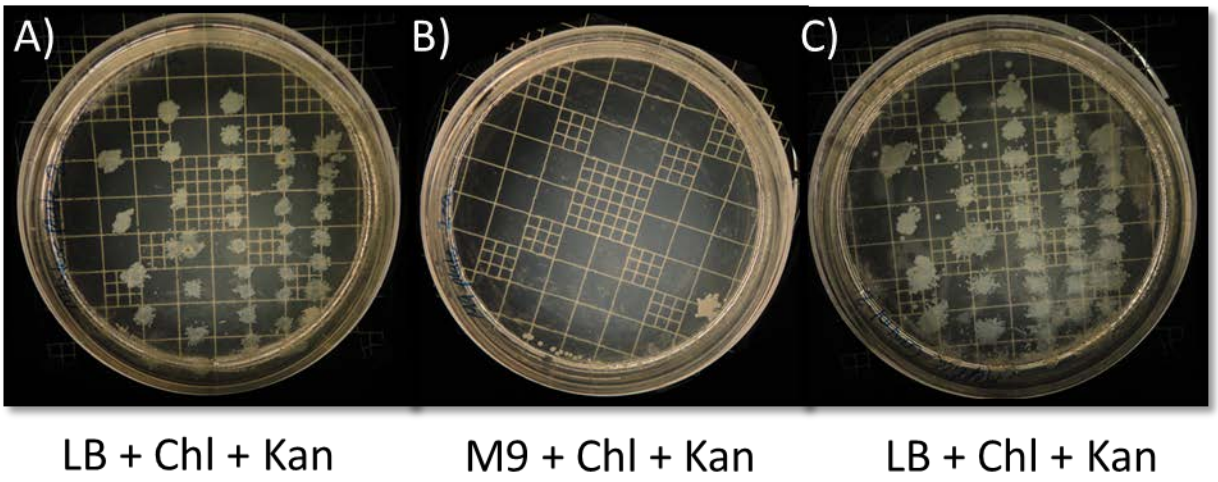

Figure 10. Mutagenesis replica plating of Tn5 induced mutant from the RB3 clone. A) Master plate, contains LB agar with kanamycin and chloramphenicol. B) Experimental plate contains M9 minimal medium supplemented with CML, kanamycin, and chloramphenicol. C) Control plate contains LB agar with kanamycin and chloramphenicol.

2.4 DISCUSSION

As more studies are being held in the development of age-related diseases, Advance Glycation End product N^{ε}-Carboxymethyl-L-Lysine, has been found to play an active role in the progression of the pathologies. This well studiedy molecule has been used for decades as a biomarker for oxidative stress. Even more recent studies associate CML with carcinogenic properties at long-term exposure (Jiao et al., 2014), increasing the list of pathologies it can directly affect. However, there is no known mechanism to reduce the amount of CML that has been accumulated. This is why in this study we screened metagenomics libraries in order to find biocatalytic enzymes capable of modifying this molecule, that could later on be implemented as a potential therapy.

A total of three metagenomic libraries were screened in order to see if they have the information required for using CML as a sole carbon source. For each screening approximately a

total of 365,625 cells of inoculum was estimated with the bacteria generations formula. Even though CML is very similar to lysine and *E. coli* has many mechanisms capable of transforming lysine into other compounds, such as the *cad* operon, which decarboxylases lysine transforming it into cadaverine (Meng & Bennett, 1992), *E. coli* EPI300 isn't able to use CML as sole carbon source. This could indicate the bacteria might lack a system capable of recognizing this molecule. By adding a genetic information of the environment, now the cell is capable of recognizing that molecule and use it to its advantage, while being able to form colonies of similar shape when grown in LBA medium with a change of pigmentation leading to a white colony (data not showed). Yet, when this genetic information gets interrupted by a transposon, the whole activity is lost, confirming that indeed the activity comes from the insert alongside with the proof that retransforming into an isogenic cell grants the phenotype to the cell. With the exception of the mutant that conserved the activity, the mutant colonies tended to be less translucent, thus a small change on pigmentation. The mechanism of action is currently unknown, meaning that the ability to use CML might even come from a heterologous function relationship, commonly found in many activity-based screenings (Couto et al., 2010; Troeschel et al., 2010).

The mechanism involved in the utilization of CML appears to be broad across different environments, according to the initial screening results, demonstrating the ability that functional metagenomics possess. This was demonstrated when we looked at the restriction analyses from both El Yunque and the Microbial mats. In the case of El Yunque (figure 6), each clone studied possessed an unique restriction pattern. This suggests that there is a possibility of more than one mechanism for this activity found in that specific soil sample. By contrast, in the benthic microbial mat (Fraternidad) only one clone was obtained in both rainy and dry seasons (figure 7). While on

the ephemeral microbial mat (Candelaria), the clone differs from the community present found in the benthic mat.

Yet describing how it works is more complicated than a simple screening. Since the inserts were of high molecular weight, it could imply more than one gene involved. Determining the gene(s) responsible for this action is essential in order to develop newer strategies directed at the breakdown of the molecule. The transposition mutagenesis is a great tool for this since mutants are generated by completely random insertion. Usually only one mutant is shown per mutagenesis in a ideal case, and sequencing around the transposon can help us understand what was inactivated. This is an advantage over random mutagenesis protocols on where there is no current way to determine the location of the mutation without knowing the target gene (Kulasekara, 2014).

CHAPTER 3: IMPLEMENTING BIOINFORMATICS AS STRATEGY TO IDENTIFY GENES ASSOCIATED WITH THE BIOCATALYTIC ACTIVITY OF N^E-CARBOXYMETHYL-LYSINE

3.1 INTRODUCTION

Prior next generation sequencing (NGS), genomics field has been constrained by smaller and repetitive content such as viruses and bacteria's. Other than that it was greatly used for genes related to genetic diseases using animal models and the human genome map (Koboldt et al., 2013). Now with the development of these new platforms, we can study variation within each genome, whole system interactions, communities, and single cell genomics. On where each platform has develop its strengths and weaknesses, related to the required application (Liu et al., 2012). However, one of the greatest promises NGS has to offers, is being able to sequence how whole microbial communities are affected by different environmental factors (Yergeau et al., 2012). lso allows us to study their interactions within these communities by analyzing the expression at transcriptomic levels, also known as metatranscriptomics, that could eventually be mapped into a full biochemical interactions model, in order to study the role of the microorganisms living in that community of choice (Gorni, 2014; Wang et al., 2009).

In contrast, NGS platform had been relying on sequencing small fragments many times, and thus it makes it difficult to assemble *de novo* sequences since these are managed without any template or guided reference. Giving bioinformatics the challenge of generating better algorithms capable of aligning these sequence reads, yet at the same time become resource friendly, bioinformatics reduces the amount of time required for assembly, allowing scientist to interpret the data and explore it in less time. Currently, the most common method used for alignment is known as *de Bruijn Graph.* This algorithm breaks down the sequences into short pieces known as *k-mers*. Sequences do not lose their content when fragmented into *k-mers* since they are all connected and preserved within the *de Bruijn Graph.* In this way it generates alignments that allow us to assemble more precisely repeated regions in contrast with linear alignments such as those

used for Sanger based sequences. The algorithm known as SPAdes Assembler has been optimized to deal with Illumina data entry, specifically for small genomes and large cloned fragments (Bankevich et al., 2012; Nurk et al., 2013; Prjibelski et al., 2014). However, sequences that contain repetitive terms eventually are mistakenly aligned toghether due to their common similarities. This is why different strategies had been developed to address this issue, such as combining NGS platforms or supplementing with capillary sequencing.

In the case of metagenomics studies, a direct sequencing has been used mostly to study microbial communities and microbiomes, since it relies mostly on identifying the 16S rDNA and sorting it out using pipelines like MG-RAST (http://metagenomics.anl.gov/) a metagenomic analysis server that automated the whole process of data analysis (Claud et al., 2013). But when it comes to discovering novel genes, it is a whole different strategy since there are no current pipelines capable of sorting the amount of information generated, making it one of the biggest challenges in metagenomics right now. Current methods for metagenomic analysis include analyzing clones by PCR related functions and overexpressing positive hits to analyze products such a polyketide synthase (PKS) clusters (Owen et al., 2013) in order to find biological active therapeutic molecules. Other methods include heterologous expression systems, describe by phenotypical activity followed by sequencing and assembly using *de novo* mechanisms on where the fosmid/cosmid or small inwards sequences obtained by capillary sequencing are used to pull out the most probable sequence along with mutagenesis studies to identify possible new functions to hypothetical proteins (Godzik, 2011; Lam et al., 2014).

This way we can compile the information generated from the Tn5 mutants in our study and supplement it with the sequence of the whole fosmid. In order to understand the underlying

mechanisms of the activity provided by the metagenomic insert found in our clones. This may allow us to predict if this activity comes from a novel function or not.

3.2 MATERIALS AND METHODS

Analyzing mutants

A total of 9 mutants generated from clone RB3 were selected for sequencing. Samples were prepared and sent in triplicates for sequencing at Macrogen USA, Rockville, MD (https://www.macrogenusa.com/) over their capillary sequencer using the following primers (see figure 11) and according to the company requirements.

KAN-2 FP-1 Forward Primer

5′ - ACCTACAACAAAGCTCTCATCAACC - 3′

KAN-2 RP-1 Reverse Primer

5′ - GCAATGTAACATCAGAGATTTTGAG - 3′

Figure 11. Sequence primers provided by the EZ-Tn5™ <KAN-2 protocol. These primers belong to Epicentre technologies (Epicentre, 2011a).

Sequences were then trimmed automatically using quality scores with the built-in trimming function found in the CLC Main Workbench 7.5 bioinformatics suite (developed by Qiagen, http://www.clcbio.com/products/clc-main-workbench/). This removes any low-quality base call and any primer region. Each triplicate was then aligned to see if there were any errors during sequencing and generate a consensus sequences from each triplicate. Then the annotation of the transposition site was held by searching for a mosaic end sequence generated from the insertion as shown in figure 12.

Figure 12. EZ-Tn5 transposon insertion site junction. Mosaic ends are used for locating the insertion (Epicentre, 2011a).

Upon identification of this sequence 19bp of the reverse strand next to the mosaic were then deleted manually due to the duplication of the bases during the insertion of the transposon. Transposition sites were then deleted, and both sequences were then properly merged into one consensus target.

For identifying the promoter regions, the Softberry bacterial promoter software (BPROM) (http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) was used, and each promoter region was annotated manually according to the coordinates given by the online software. Then the Open Reading Frames were predicted using the three main start codons available in *E. coli*, ATG, GTG and TTG (Blattner, 1997) by using an adaptation of the glimmer3 software found on CLC Main Workbench. Each ORF is blasted using NCBI blastx algorithm to determine possible encoding proteins.

Shotgun Next Generation Sequencing

Samples were prepared by implementing the miniprep extraction protocol explained in section 2.2. Purity of the 260/280 ration was obtained using the Eppendorf bio-photometer. Once the pattern matched the requirements (1µg of DNA with 1.8-2.0 260/280 ratios) of the sequencing company (MR DNA Lab, Shallowater, TX). The sample was sent out for sequencing using the

communal sequencing services. The services consist of library preparation and sequencing using Illumina Mi-Seq sequencer on the pair-end 200nt inserts technology.

Once the sequence files were downloaded, assembly began by using SPAdes 3.0 *de novo* assembler algorithm (http://bioinf.spbau.ru/spades) which is based on *de Bruijn Graphs* and optimized for Illumina reads coded by the Russian Academy of Sciences. The algorithm was run on a dedicated Ubuntu Linux 14.04 server with two Intel Xeon L5420 processors (a total of 8-cores at 2.50GHz each) and 16GB of RAM. The input command used to run the assembly follows as directed: "*spades.py –k 21, 33, 55, 77 –careful -1 RBIII3_1.fastq -2 RBIII3_2.fastq –o fosmids/*" where fastq are the pair-end sequence files obtained from the company and the different values of *k-mers* allow the program to generate different contigs by overlapping different sizes. Once finished the output file generated is called "scaffolds.fasta", which is then run through a local BLAST within CLC Main Workbench using the vector sequence (pCC1FOS), identifying the real sample and not those generated by small contaminants of genomic DNA from the host or other sequences that could have been generated.

Manual curating and editing was done using CLC Main Workbench where vector integrity was checked, and files were then exported and a final assembly was done using DNASTAR Lasergene SeqMan NGen template assembly, using the previously generated sequence as the template and *E. coli* K-12 genome sequence for contaminant removal. This commercial application is similar to the Bowtie algorithm but manages resources more efficiently.

Annotation

Primary annotation was held using Rapid Annotation using Subsystem Technology (RAST) by uploading the fully assembled fosmid and annotating with the classical RAST module along with the FIGfam release version 70 and selecting the build metabolic model option present.

Once the pipeline returns with the results, usually 2-3 hours, the annotation provided as a General Feature Format (GFF) file was then complemented with the previously generated mutant's sequences by performing local BLAST tool on the CLC Main Workbench software against the fully assembled fosmid. The surrounding sequences across the mutant's annotation were verified by blasting each of the CDS generated with RAST GFF file using the National Center for Biotechnology Information (NCBI) blastx tool, to assign proper annotation. ORF predictions were made in the CLC Main Workbench software that could start with either AUG, CUG, and UUG as start codons with a minimal of 200 codons present in order to find any possible target left behind by the RAST pipeline annotation server.

3.3 RESULTS

Genomic Map

The fosmid assemblage was performed with 100X coverage using the SPAdes 3.0, as obtaining a 41,089bp fosmid including both the vector (8141bp) and insert (32948bp). A total of 57 ORFs were identified on the insert. Out of those 36 were classified as hypothetical proteins. Four of the non-hypothetical were manually annotated, and the rest were classified according to the RAST pipeline. Possible affected areas by Tn5 insertions were marked as well, shown in figure 13. The insert sequence without the vector has been deposited at NCBI GenBank nucleotide database (reference number KP938202) along with the annotated regions for future references. The closest related genome to the sequence corresponds to *Methylobacterium nodulans* (accession number: CP001349.1, query coverage: 6%, e-value: 8E-83, Identity: 63%) according to the NCBI non-redundant database.

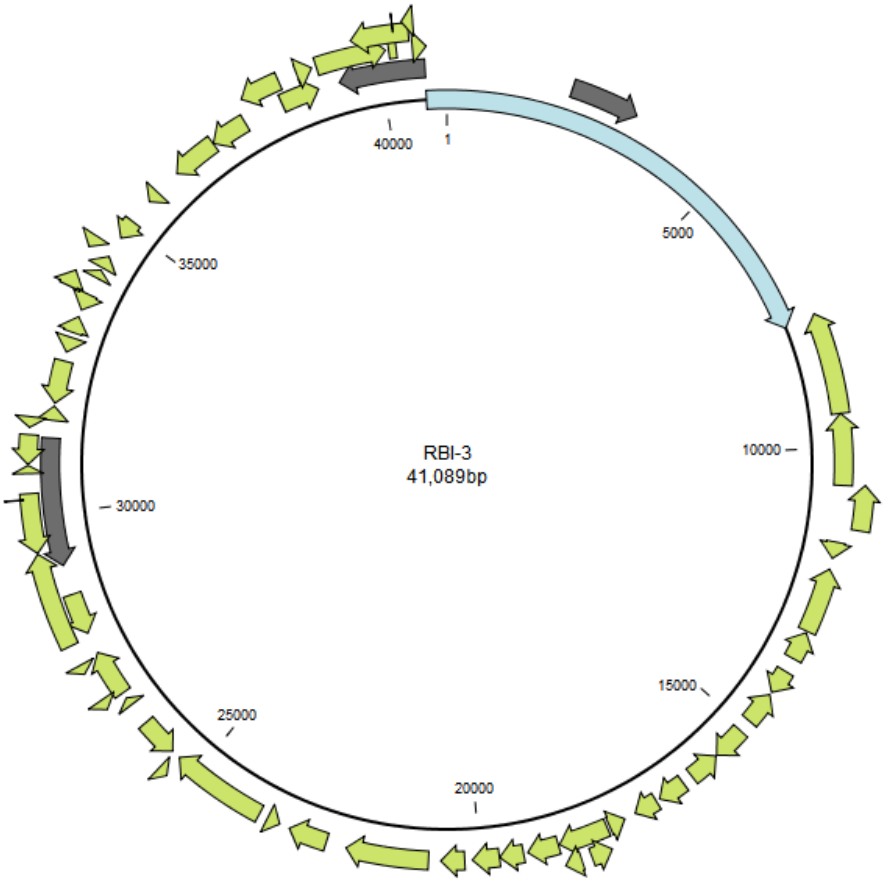

Figure 13: Complete genomic map of the fosmid assembled. Green arrows indicate coding ORF's, black arrows indicate regions affected by transposition mutagenesis due to the DNA displacement. Blue arrow indicates the vector, pCC1FOS. The two black bars present indicate the exact point of insertion of the transposon. ORF found on of the vector where left blank intentionally.

Mutant Analysis

Out of 9 mutants sent for sequencing using capillary sequencing, 3 of them were the only ones whose sequences were reliable and useful due to the noiseless background and high-quality score. Capillary sequencing analysis was able to pinpoint the location of the transposon and provided enough information to predict five possible ORFs as shown in figure 14. Mutant coded as RB3M6 landed on the vector.

According to the figure 14, the transposon in mutant A was inserted around replication genes associated with the pCC1FOS vector. We thus assumed that the mutation prevents CML metabolism by interfering with the fosmid ability to replicate, and eliminated this mutant from further analysis. Mutant B contained two possible targets, a predicted ABC transporter and a predicted CRISPR associated protein. Mutant C predicted a sulfite oxidase interruption.

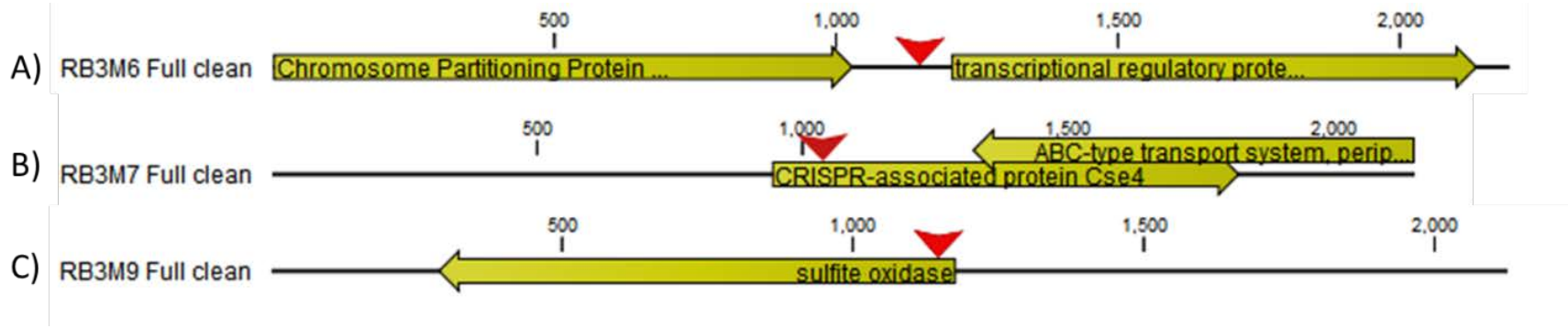

Figure 14: Mutagenesis analysis of RB3 Tn5 insertions. Red arrow indicates transposon insertion site. A) RB3M6 mutant and its possible targets. B) RB3M7 mutant and its possible targets. C) RB3M9 mutant and its possible targets. This data was generated from the Sanger sequences.

Combination of both capillary and Illumina data allows us to predict the complete surroundings of the Tn5 insertion. Zooming in on where the mutant RB3M7 landed shows that it is mostly covered with hypothetical proteins as shown in figure 15 with a plausible ABC transporter and a CRISPR associated protein.

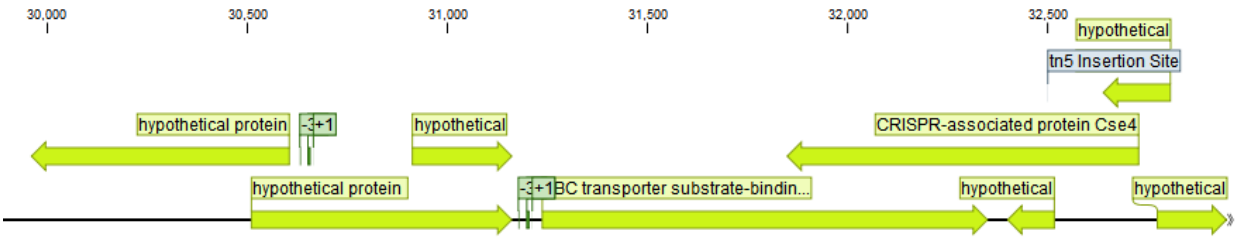

Figure 15: Zoomed region of the RB3M7 mutant. Most of the ORF's found in this region encode for hypothetical proteins of unknown functions. Green flags indicate possible promoter regions. Blue flag indicates Tn5 Insertion. While the green arrows mark the predicted genes.

Zooming in into the RB3M9 region (figure 16) shows a membrane adenylate cyclase regulatory protein and a possible sulfite oxidase along with a cluster of hypothetical proteins.

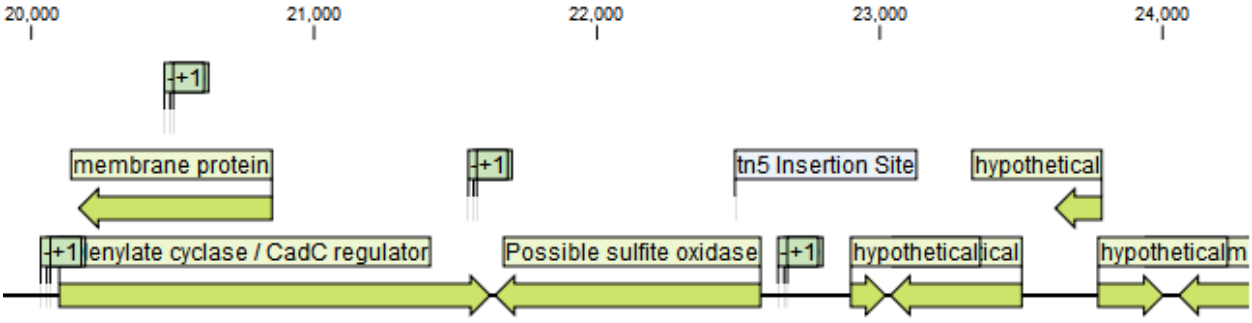

Figure 16: Zoomed region of the RB3M9 mutant. Green flags indicate possible promoter regions. Blue flag indicates Tn5 insertion. While the green arrows mark the predicted genes.

Analysing targets

A list of the comprehensive manual annotation performed by blastx of RB3M7 targets is shown in table 1 according to the most relevant position of Tn5 insertion site. The currently most significant predicted gene from this mutant, would be the ABC transporter since it contains a domain associated with the uptake of amino acids. The other target presents unreliable information due to the high e-values and low query coverage. The ABC transporter originated from a organism that belongs to the rhizobia.

Table 1. Top hits for RB3M7 Mutant Plausible Targets and Annotations

Name	Position on the Insert	CDS Size (bp)	Organism Related	Function	FIGfam ID	NCBI Accession Number	E-Value	Identity %	Query Coverage (%)
CRISPR-associated protein Cse4	32729..31848	882	*Streptacidiphilus neutrinimicus*	Unknown	N/A	WP_042368918.1	0.73	36	25
hypothetical protein	32729..31848	882	*gamma proteobacterium* HTCC2207	Unknown	N/A	WP_007231351.1	4.4	41	18
RNA-helicase	32729..31848	882	*Plasmodium cynomolgi* strain B	Transcription, Biogenesis, Recombination	N/A	XP_004220684.1	9.3	51	13
ABC transporter substrate-binding protein	31236..32351	1116	*Bradyrhizobium sp.* Cp5.3	Type I periplasmic ligand-binding domain of uncharacterized ABC-type transport systems that are predicted to be involved in the uptake of amino acids,peptides, or inorganic ions; cd06325	6666666.96294.peg.56	WP_027552054.1	4.00E-23	36	60
hypothetical protein	31236..32351	1116	*Bradyrhizobium japonicum*	Unknown	N/A	/P_039155830	8.00E-27	37	60
hypothetical protein	31236..32351	1116	*Ralstonia sp* . JGI 0001001-C06	Unknown	N/A	WP_037022658.1	2.00E-32	54	34

A list of the comprehensive manual annotations of RB3M9 targets is shown in table 2 according to the most relevant position of Tn5 insertion site. A predicted sulfite oxidase and a membrane adenlyate cyclase regulator gene were both identified from organisms related to the rhizobia.

Table 2. Top hits for RB3M9 Mutant Plausible Targets and Annotations

Name	Position on the Insert	CDS Size (bp)	Organism Related	Function	FIGfam ID	NCBI Accession Number	E-Value	Identity %	Query Coverage (%)
Sulfite Oxidase	22581..21637	945	*Rhizobium alamii*	SO catalyzes the terminal reaction in the oxidative degradation of the sulfur-containing amino acids cysteine and methionine. Cl00199	6666666.96294.peg.37	WP_037108068.1	9.00E-59	43	87
hypothetical protein N007	22581..21638	945	*Alicyclobacillus acidoterrestris* ATCC 49025	Unkown	N/A	EPZ47813.1	1.00E-68	46	88
sulfite oxidase	22581..21639	945	*Bradyrhizobium sp.* LTSPM299]	SO catalyzes the terminal reaction in the oxidative degradation of the sulfur-containing amino acids cysteine and methionine. Cl00199	6666666.96294.peg.37	KJC56794.1	6.00E-63	45	87
membrane protein	20096..21625	1530	*Bradyrhizobium sp.* URHA0002	Two component system transcriptional regulator, cl21459, cd00383, TIGR02917	N/A	WP_027540690.1	3.00E-87	39	99
adenylate cyclase 3	20096..21626	1530	*Bradyrhizobium japonicum*	Signal Transduction	6666666.96294.peg.36	WP_039151921.1	3.00E-79	36	99
CadC regulator	20096..21627	1530	*Rhizobium sp.* CF258	Regulates *cad* Operon	N/A	WP_037159537.1	2.00E-82	37	99

Protein phylogenetic analysis of the targets along with experimentally based data of similar structure is shown below (figures 17-20). For mutant RB3M9, the analyses of nearby CDS in relation to the Tn5 insertion are shown in figures 17 and 18 for both possible targets. For the RB3M7 mutant, the analysis of nearby CDS in relation to the Tn5 insertion is shown in figure 19 and figure 20 for both possible targets. Target 1 as seen in figure 17 showed that it is closely related to the sulfite oxidase functional structure. Target 2 as seen in figure 18 indicates that it is more closely related to functional structure of the adenylate cyclase class III family. Target 3 as seen on figure 19 indicates close structural relationship with the ABC transporters family. While target 4, shown no plaussible structural clasification in relationship with the blastx hits.

No enzyme of the known lysine degradation pathways was found using KEGGs metabolic models, but there was a positive hit for one related to lysine biosynthesis. The enzyme, aspartate-semialdehyde dehydrogenase, was found 19,000bp away from the sulfite oxidase (SO) gene. Both FIGfam and NCBI classified it as aspartate semialdehyde dehydrogenase (Accession number: ESJ97595.1, ID: 6666666.96294.peg.10).

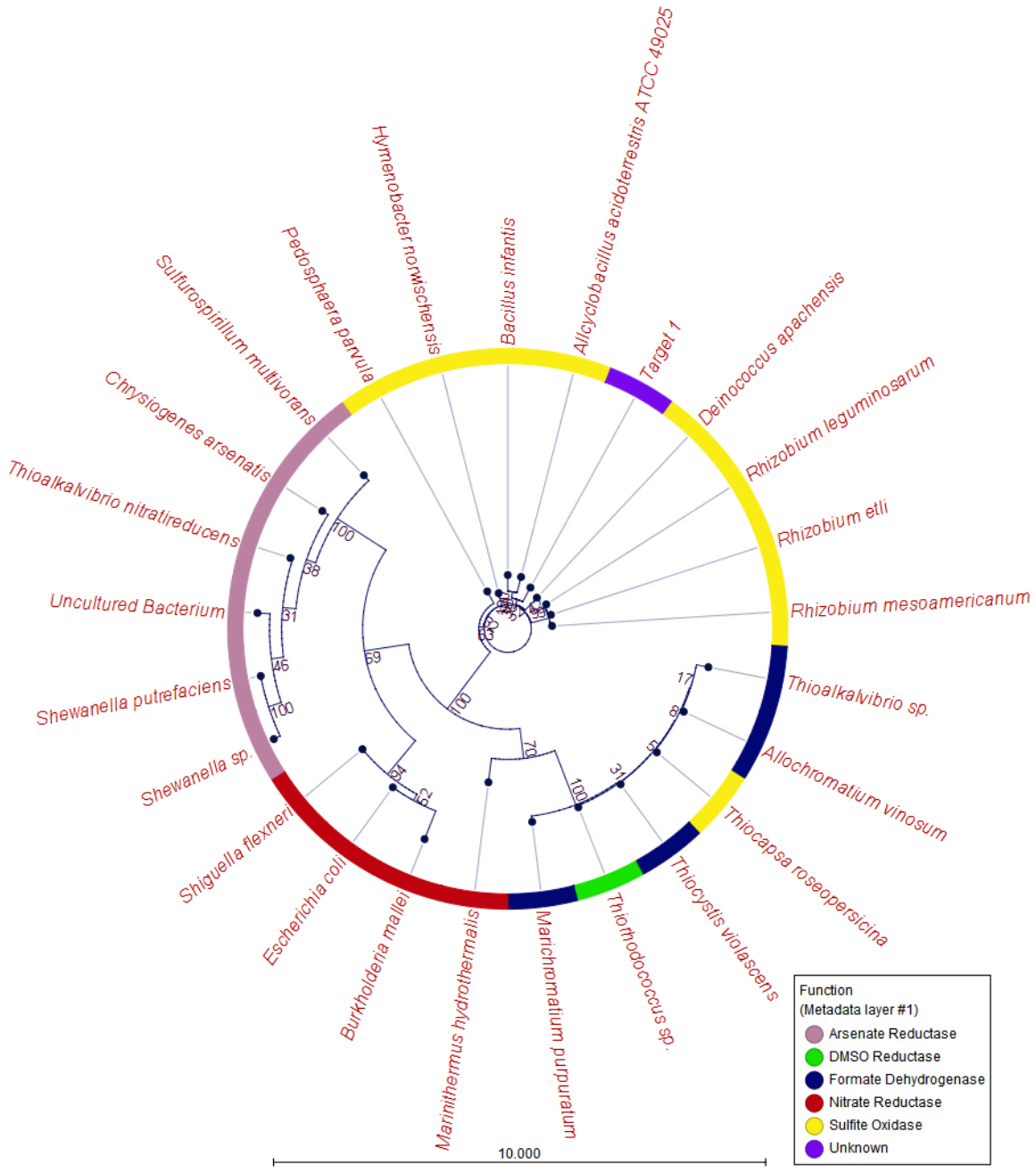

Figure 17: Maximum likelihood circular phylogram of RB3M9 Sulfite Oxide CDS assignment based on the molybdenum cofactor binding site conserved domain.

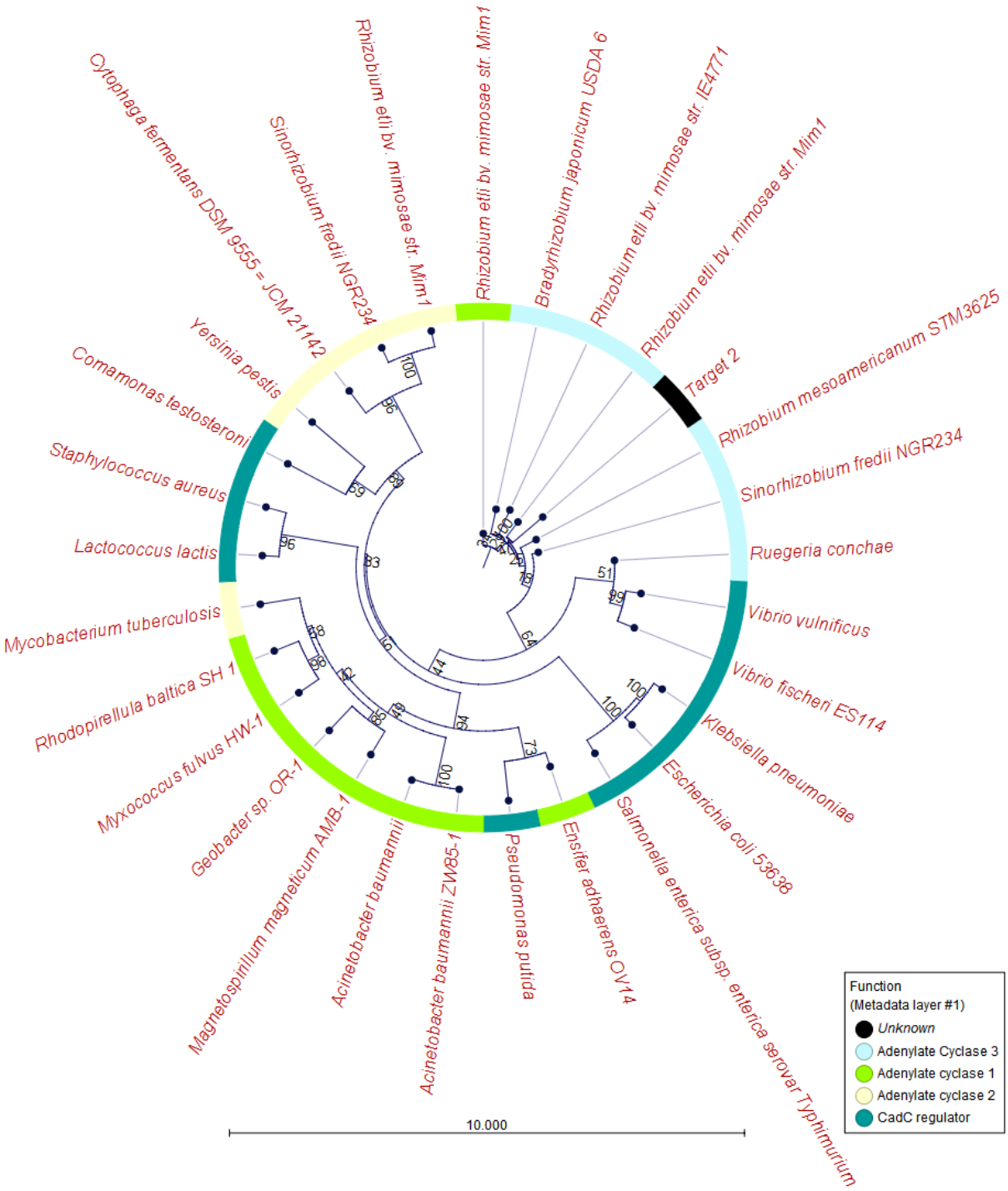

Figure 18: Maximum likelihood circular phylogram of RB3M9 membrane regulatory protein CDS assignment based on the effector prokaryotic domain of response.

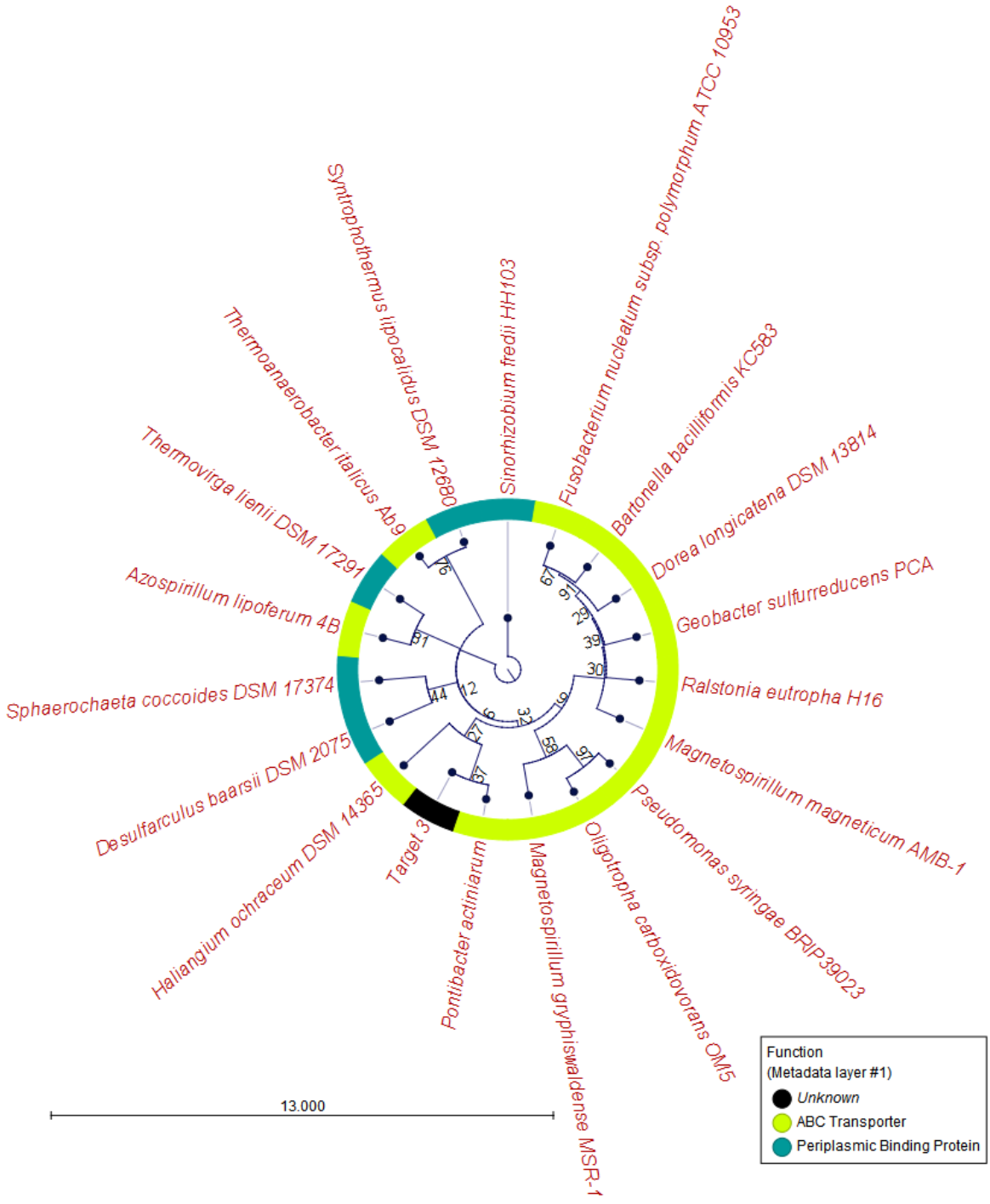

Figure 19: Maximum likelihood circular phylogram of RB3M7 ABC-Transporter CDS assignment type I periplasmic ligand-binding domain.

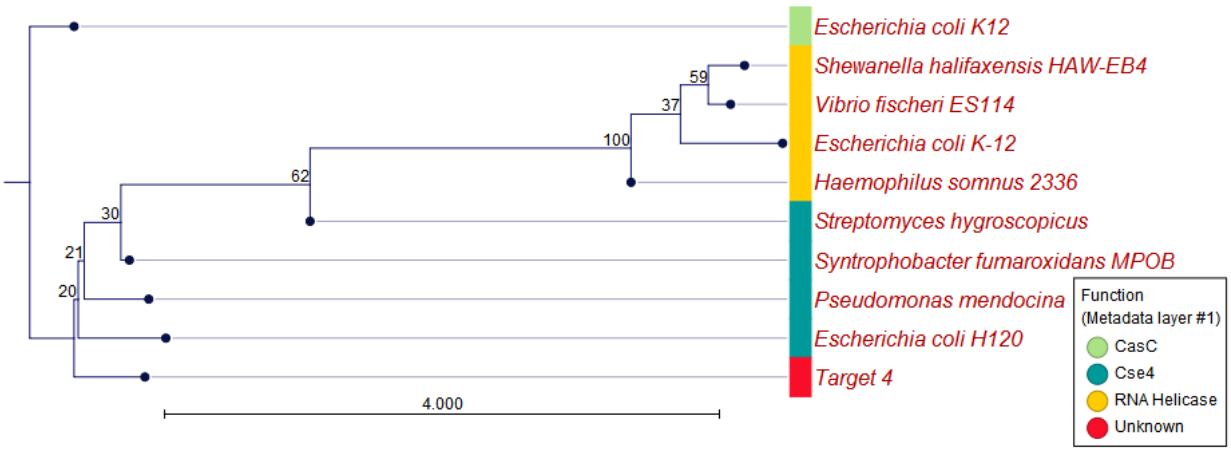

Figure 20: Maximum likelihood phylogram of RB3M7 CRISPR-associated protein Cse4 CDS assignment based on BLASTX hits, described on table 1. No proper assignment found for this specific target (Target 4) according to BLASTX hits obtained in table 1.

3.4 DISCUSSION

Bioinformatics allows us to answer current biological question, yet it is currently limited by computing power and database submission content. Many tasks might require the use of a dedicated server or a super computer to provide the most accurate model and predictions. After assembling a whole fosmid by usage of a dedicated server,. an initial *in silico* analysis was performed to predict possible functions associated with the annotations found by the RAST pipeline. When we took an insight at the fully assembled fosmid, we found that most of its component are at present hypothetical proteins. Hence, little information is available right now to fully understand the mechanics brought by these inserts once expressed in *E. coli* EPI300. Restricting the analysis mainly to the information found surrounding the mutagenesis site and using protein phylogenetic analysis, we can predict better protein function by establishing sequence similarity with known proteins and their structural and functional conserved domains (Pellegrini et al., 1999).

Examining mutagenesis site for RB3M7 we found a possible target related to the phenotype observed in the wild-type strain, barely similar to Cse4 family protein of the CRISPR enzyme, since the insertion occurred directly over this possible protein with an unusual start codon, TTG. According to *E. coli* codon usage, TTG is known to be the lowest affinity start codon for fMet amino acid using it only 3% of the time. However, the prediction values for this family demonstrate that there's a very certain chance of misclassification as shown in table 1. Right next to this sequence we found a gene encoding a periplasmic binding protein. This family of proteins provide a great variety of functions that favor their protein engineering exploitation. Their characteristics include ligand-binding function and their ability to suffer conformational changes and structural adaptations (Dwyer & Hellinga, 2004). They are involved in many signal translational pathways, are capable of responding to extracellular substances, and interacting with biologically active enzymes. Some of them like the ABC transporter family, play a crucial role in importing extracellular components into the intracellular structure (Procko et al., 2009). One possible function of this periplasmic membrane protein is to work as a transporter to let in the CML into the host cell. Inactivation of this gene would prevent the organism to use its oxidase activity which could explain the phenotypic traits found when the Tn5 inserted near the region.

On the other hand, at the mutagenesis site of RB3M9 we found a gene coding for a possible Sulfite Oxidases (SO) right on the Tn5 insertion site. The SO as predicted, are virtually found in all life forms, with greater abundance in prokaryotes, yet just only a few of them have been characterized. They happen to be involved in different metabolic processes that generate energy from sulfur compounds, while others are responsible for host colonization of pathogenic bacteria (Kappler, 2011). Although CML contains no sulfur atom, this enzyme could play a role in the biotransformation of CML.

Next to the SO, we found a predicted gene with a regulatory membrane protein domain; the exact component regulated is currently unknown. As shown in figure 18, evolutionarily it is closer to a class III adenylate cyclase. This is one of the most studied classes in this family since it plays important roles in human health (Linder, 2006), as well in bacterial pathogenesis (Lory et al., 2004). If the regulatory protein has a similar function as a CadC regulator, this would imply that the overexpression of this protein could forcefully activate the *cad* operon found in *E. coli*, forcing the cell to intake lysine. In this case, the structurally similar compound of interest CML could penetrate the cell. But this could trigger fluctuational changes of pH (Rauschmeier et al., 2014) if CML is somehow transformed into cadaverine that would increase pH until lysine or in this case CML has been completely depleted. Either case, the periplasmic membrane protein could be involved in a regulatory process of the biotransformation of CML as well.

The presence of a lysine biosynthesis component could also indicate that the molecule could end up being transformed into lysine once again, meaning that plenty of targets could withhold the key for this phenomenom. That could eventually generate the required information to engineer a strategy that could actually reduce the amount of CML that accumulates in our bodies through time. In order to achieve this, the predicted genes would first need to be confirmed by experimental approaches. Different strategies could be employed, like genetic complementation, real time quantitative PCR, transcriptomic data analysis and purified protein supplementation. For this particular case, since we are using mutants to identify genes, one can establish genetic complementation and verify if the cloned gene is essential or not for the activity. Isolating these genes would eventually allow us to over express and purify proteins that later can be used to explore the mechanism involved in the biotransformation of CML.

No known organism was found over the current database that was closely related to the microorganisms from which the described metagenomic insert originated. However, many of the protein encoding genes described happen to be closely related to the those categorized as rhizobia, a large group of microorganisms capable of fixing nitrogen by developing symbiotic relationships with legumes (Zahran, 2001). Still, not enough evidence has been found to support this claim. In order to confirm this, more clones from El Yunque metagenomic library would need to be selected by using probes generated with the current sequence followed by sequencing and assembly until either a 16S rDNA or a conserved nitrogenase complex is found.

CHAPTER 4: BREAKING IT DOWN: EVALUATING PLAUSIBLE GENES FOR N^E-CARBOXYMETHYL-LYSINE ASSIMILATION BY *IN-TRANS* COMPLEMENTATION

4.1 INTRODUCTION

Bacterial genetics is structured to maximize efficiency, at the genomic scale; their genomes tend to be organized, and genes with related function can be found close to each other. Even when the chromosomal structure could vary within strains of the same species or rapid changes occurring in their genes, functional regions tend to resist change (Rocha, 2008). Even with sequencing whole genomes and all the current understanding of genomic structure, the biological function of half of the *E. coli* proteins is unknown (Vlasblom et al., 2014). There are many ways to study and assign functions to a gene. Bioinformatics uses structure-based pattern analysis to determine possible genes, motifs or domains according to previously studied experimental data uploaded to the databases. Experimentally, we can source genes by cloning and expressing them to determine physical properties of the message that they encode. In addition, we can perform mutations and study how the mutations affect phenotype. With this information, databases can later be generated to support the bioinformatics platform. However, it is important to recall that assignments of genes trough this platform are only predictions generated by algorithms that are written to associate the functions with structural similarities (Lord et al., 2003; Stockwell, 2000), and experimental data would eventually be required to confirm the assignments.

Among mutation-based analyses, there is an important tool that has been used to characterize genes known as genetic complementation test (figure 21). Many variants of the complementation test have been developed to answer specific questions, yet the principle remains the same. How a mutant behaves when a wild-type copy of the gene is incorporated and expressed in a vector, this helps identify functions and dominance (Oh et al., 2003; Shin et al., 1983; Simon et al., 2009). The *in-trans* complementation mechanism allows one to isolate a specific gene and determine its role by complementing a wild-type copy in *trans* position (figure 21). In this way,

viruses have been studied to determine the function of their genes (Khromykh et al., 1998; Myhre et al., 2010). Trans complementation tests have also been used to determine the virulence genes required for infection of *Agrobacterium tumefaciens* (Berger & Christie, 1994).

However, there are several disadvantages regarding a *trans* complementation assay, mostly influenced by the complexity of the study system. In case that more than one copy of a gene is present, the assay cannot be performed unless all copies are completely silenced. Also if the expected complementation belongs to a gene that encodes long lived molecules such as proteins, no remnants of the proteins can be left active in the cell, since they can generate false positive results.

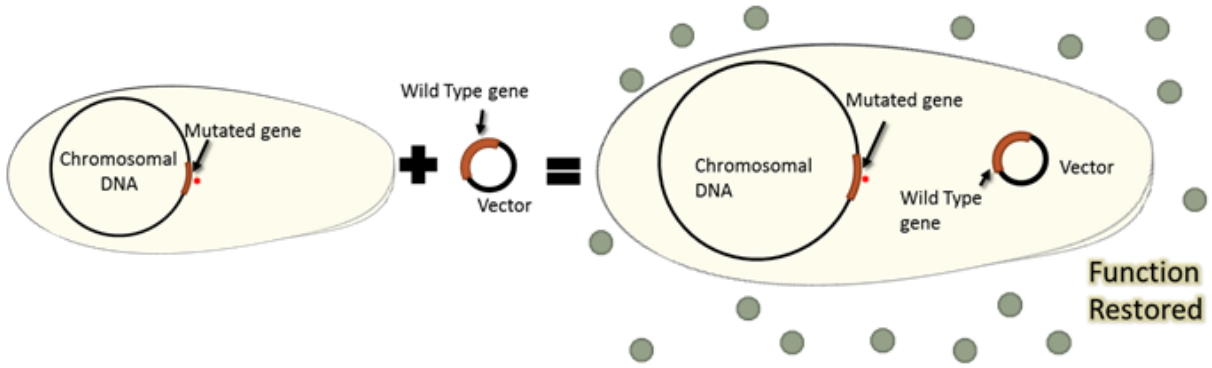

Figure 21: Basic mechanism of Trans-complementation to test if the mutant generated can be restored by complementing with a WT copy.

In this study, we planned to use genetic complementation based on plasmid to fosmid trans-complementation. Specifically, the WT gene would be cloned and placed together in *trans* position to see if the phenotypic activity coded by the WT is recovered in the mutant generated by

transposition mutagenesis. We are also supplementing the mutant with the extracellular matrix exposed to the WT to get an insight if this genetic property can be transmitted extracellularly.

4.2 METHODOLOGY

Secreted products

There is a possibility that the required working properties can be secreted into the medium. In order to rule out this scenario, a supplement made out of the surface of the plated clones capable of growing on the minimal medium by washing with a solution phosphate buffer at 7.2pH, was generated by filtering 3mL of the wash over 0.22µm nitrocellulose filters. Fifty µL of previously washed mutants in physiological saline (0.85% NaCl), were mixed with 50µL of the filtered solution and plated over three different minimal media supplemented with CML, without carbon source and with glucose, respectively. The same procedure was done for the mutant without the supplement and the WT without the supplement. Samples were incubated at 37°C for up to 48h. After incubation period, a table was generated for those capable of growing or not over each condition presented. This experiment was held with a designated personnel who did not know the content of the plates in order to rule out any biases and perform a double blindfold experimentation, in which records are kept by two different individuals and then compared according to the key built by the one who prepared the inoculum.

PCR-based cloning complementation

For the RB3M9 mutant targets, rimers were designed using primer3 (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) with the ORF obtained from the mutant sequences using Sanger capillary and the annotated genes obtained from Illumina sequence. Three sets of primers were designed (shown in Table 3) for the pENTR™/SD/D-TOPO® cloning system

from Life Technologies, one containing a promoter region (expected product size of 1100bp) and another without it (expected product size 918bp) derived from the Sanger based sequence. The last set was designed using the Illumina sequence containing a predicted sulfite oxidase annotated by RAST (expected product size 945bp).

Table 3. Primers generated for the cloning of RB3M9 mutant targets.

Primer	Sequence	Tm (°C)	Direction
RB3M9-P	5'-CACCGCCCAAGCCAGACCGAA-3'	64.8	Forward
RB3M9-O	5'-CACCCGTGTACTTCACACTCTGGA-3'	60.3	Forward
RB3M9-R	5'-CTGGAGACAACGACCGGAAC-3'	57.9	Reverse
SO-RB3M9 Fwd	5'-CACCGTGTTTGCGACCCTCGAGT-3'	63.6	Forward
SO-RB3M9 Rev	5'-TCAGATTCCACCACCTTGCAT-3'	56.6	Reverse

Due to the high GC content, a gradient PCR was performed in order to obtain proper annealing temperature. For this, annealing temperature was run from 70-58°C for the capillary sequences based design primers, using BIORAD T100™ Thermal Cycler.

For the amplification of each of the PCR products the parameters used are listed on table 4. All of the PCR products were amplified using Promega GoTaq Green Master Mix at a reaction volume of 25µL. *Briefly: 12.5µL of GoTaq Green X2 Master Mix, 100ng of RBI-3 fosmid DNA, a final concentration of 1µM of each primer and nuclease free water to complete volume of 25µL.*

Table 4. PCR parameters for each of the reactions needed to clone with pENTR/SD/D-TOPO kit.

Primers	Initial denaturalization	Denaturalization	Annealing	Extension	Cycles	Final Extension	Hold (∞)
RB3M9-P RB3M9-R	95°C for 5min	95°C for 30sec	58°C for 1min	72° for 1.5min	25	72°C for 7min	4°C
RB3M9-O RB3M9-R	95°C for 5min	95°C for 30sec	56°C for 1min	72° for 1min	25	72°C for 7min	4°C
SO-RB3M9 Fwd SO-RB3M9 Rev	95°C for 5min	95°C for 30sec	54°C for 1min	72° for 1min	30	72°C for 7min	4°C

PCR products were cleaned using MO BIO Ultra Clean® PCR clean up kit. Cloning was performed following the Life Technologies protocol as stated (Invitrogen, 2012). *Briefly: On a bucket filled with ice 2µL of purified PCR product, 1µL of the included salt solution, 2µL of the supplied water and 1µL of vector were mixed in a 0.5mL tube.,Incubation was done at room temperature for 30mins, followed by chemical transformation using the provided E. coli TOP10 by mixing 2µL of the ligation resting on ice for 30mins. Heat shock was done by incubating at 42°C for 30s, and rapid transfer to ice, 950mL of SOC medium was added and incubated at 37°C in an orbital shaker for 2 hours. Finally, 100µL of culture were plated on LBA with 50µg/mL Kanamycin and incubated for 24h at 37°C.*

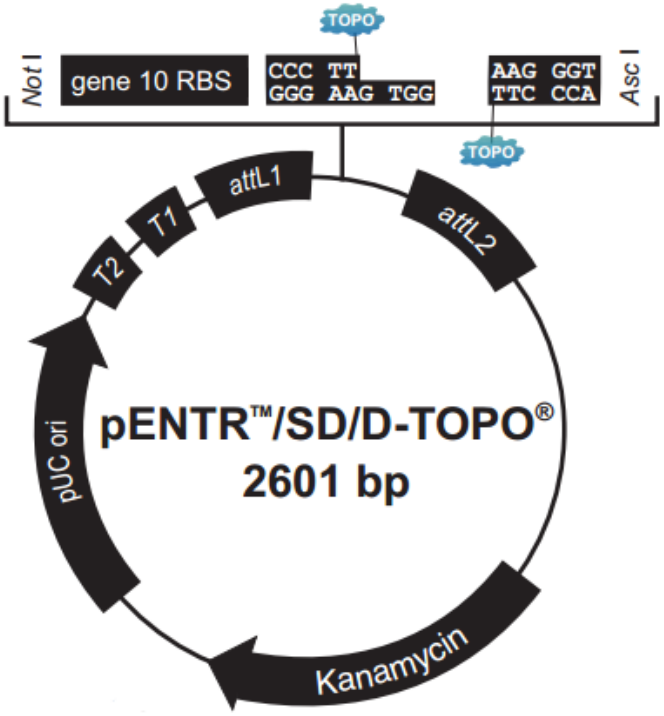

Figure 22: Map and Feature of the pENTR/SD/D-TOPO Vector. Obtained from the pENTR™/SD/TOPO® kit protocol. This vector allows only directional cloning of PCR products according to the primers design.

After recovery and incubation, colonies were selected and grown in 5mL of LB Broth with Kanamycin at 50µg/mL overnight at 37°C in an orbital shaker followed by a miniprep extraction. Since *E. coli* TOP10 strain was used to perform the cloning, and the study has been done in the *E. coli* EPI300 strain, chemical transformation was followed according to Epicentre protocol (Epicentre, 2011b) with the miniprep extraction obtained in the previous step. After transformation cells were spread on LB with Kanamycin 50µg/mL plates and incubated 37°C for 24h. Pure colonies were then transferred to 5mL LB broth with Kanamycin 50µg/mL at 37°C for 2 hours in the orbital shaker, followed by another miniprep extraction to verify presence of the plasmid. The extracted plasmid was then cut with *Eco*RV and *Not*I in order to release the insert and confirm

cloning procedure, using Promega enzymes following the protocol instructed by the manufacturer (Promega, 2011). *Briefly: 5U of each enzyme and 1µg of DNA, was mixed in a 0.5mL tube with 2.5µL of the supplied buffer and distilled water to a final volume of 25µL.* Digestion was left incubating for 3h at 37°C and ran in a 1% agarose gel at 70V for 50mins.

After verifying the plasmid content, a ratio of 1:2 plasmid to fosmid was prepared for electroporation in a 5µL solution. The vector solution was mixed with 50µL of *E. coli* EPI300 electrocompetent cells. Electroporation cuvettes were left cooling for at least 15 minutes on ice before transferring the mixture and electroporated using the *E. coli* 1 profile from the BIORAD MicroPulser™. Cells were then left incubating in LB broth at 37°C for two hours. The cells were then spread onto LBA plates containing 50µg/mL Kanamycin and 12.5µg/mL chloramphenicol, and incubated at 37°C for 24h. Individual colonies were selected and grown in 5mL of LB broth with 50µg/mL Kanamycin and 12.5µg/mL chloramphenicol, followed by a miniprep extraction. Extraction mixture was run in a 1% agarose gel to confirm the presence of both fosmid and plasmid within the cells. Once confirmed, the cells were plated or patched on an M9 minimal medium supplemented with CML for 48h to see if activity was regained.

Cloning fragment complementation

One of the mutants was identified near a complex that contained two possible genes, which encode for a predicted sulfite oxidase and adenylate cyclase, respectively. A fragment including this region has been identified. Restriction enzymes *Not*I and *Eco*RV were used to digest the fosmid and clone the fragment into a pUC21-*Not*I (Addgene plasmid # 51659, Voth, Richards, Shaw, & Stillman, 2001) plasmid with an expected insert size of 2,794bp. In order to reduce the amount of non-desired inserts cloned, the double digested fosmid was precipitated by sodium acetate-ethanol precipitation followed by a third digestion using *Xba*I restriction enzyme. *Briefly:*

A total of 5 units each of NotI and EcoRV are used along with 15µg of RBI-3 fosmid DNA on a 25µL reaction using 2.5µL of Promega buffer D and incubated at 37°C for 12h.

The complete reaction was then loaded to an agarose 1% gel and ran for 4 hours at 50V. Gel was slightly stained, and a sizing was selected around the 3Kbp range. A gel extraction was performed using QIAGEN QIAquick® Gel Extraction Kit in order to purify the size-selected fragment. *Briefly: 400mg of the sizing product was weighted and placed in a 2mL tube. 3 volumes of buffer QC was added (1.2mL) and incubated at 50°C for 10min. One volume of isopropanol (400µL) was then added and mixed by vortexing. Sample was then centrifuged in the provided spin column followed by the addition of 750µL of Buffer PE and centrifuged. Lastly, 30µL of buffer EB (elution buffer) was added and left at rest for 5 minutes. The spin column was transferred to a 1.5mL collection tube and then centrifuged.*

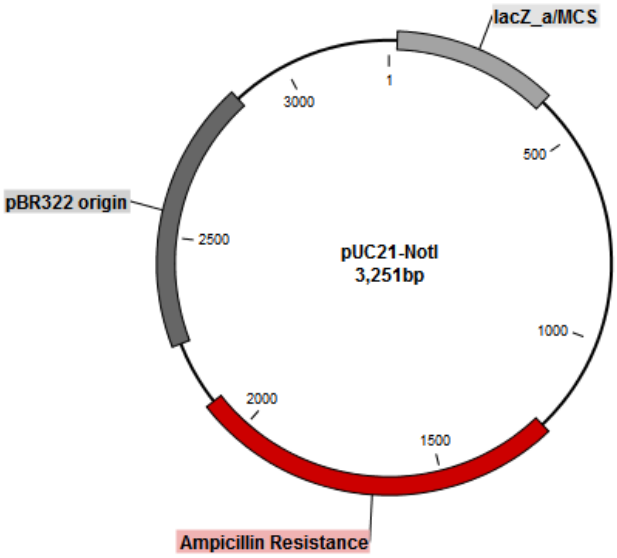

Figure 23: pUC21-NotI map, used for restriction enzyme cloning. Map was generated using CLC Main Workbench with the sequence provided by Addgene.

Ligation was held using T4 DNA ligase following manufacturer protocol (New England Biolabs, 2014) with an 3:1 insert to vector ratio incubating at 16°C overnight in order to properly ligate both sticky and blunt ends. *Briefly: 210ng of the insert was added to a 0.5mL reaction tube along 70ng of plasmid, 2µL of reaction buffer and 1µL of T4DNA ligase. Reaction was completed to 20µL with ddH_2O and incubated overnight at 16°C.*

After ligation, chemical transformation was held using *E. coli* EPI300 following manufacturer protocols (Epicentre, 2011b). *Briefly: 10µL of the ligation reaction was mixed with one vial of E. coli EPI300 chemical competent cell and left at rest for 30min on ice. Heat shock was induced by placing the sample at 42°C for 30sec and immediately placed on ice. Addition of 950µL of LB was done and the sample left incubating at 37°C in an orbital shaker.* After the recovery phase, cells were spread in LBA with 50µg/mL of ampicillin and incubated at 37°C for 24h. Clones were cryogenically preserved, followed by a miniprep extraction of the same and verified on agarose gel electrophoresis.

After verifying the plasmid content, a ratio of 1:2 plasmid to fosmid was prepared for electroporation in a 5µL solution. The vector solution was mixed with 50µL of *E. coli* EPI300 electrocompetent cells. Electroporation cuvettes were left cooling for at least 15 minutes on ice before transferring the mixture and electroporated using the *E. coli* 1 profile from the BIORAD MicroPulser™. Cells were then left incubating in LB broth at 37°C for two hours. The cells were then spreaded onto LBA plates containing 50µg/mL ampicillin, 50µg/mL kanamycin and 12.5µg/mL of chloramphenicol, and incubated at 37°C for 24h. Individual colonies were selected and grown in 5mL of LB broth with 50µg/mL ampicillin, 50µg/mL kanamycin and 12.5µg/mL chloramphenicol followed by a miniprep extraction. Extraction was ran in a 1% agarose gel to confirm the presence of both fosmid and plasmid within the cells. Once confirmed, the cells were

plated or patched on an M9 minimal medium supplemented with CML for 48h to see if activity was regained.

4.3 RESULTS

Cellular Extract Complementation

After 48h of incubation, each plate was evaluated individually by two different persons who were unaware of the content of each plate. Once observations were verified with the annotation key, a table was generated with the content of each plate as shown in Table 5..

Table 5. Extracellular filtrate solution over Phosphate Buffer Saline pH 7.2 grown in M9 minimal medium.

	M9 + Glucose		M9 + CML		M9 No Carbon	
Sample	Expected	Observed	Expected	Observed	Expected	Observed
E. coli EPI300 with pCC1FOS without insert	+	+	-	-	-	-
RBI-3 WT	+	+	+	+	-	-
RB3M9	+	+	-	-	-	-
RB3M9 + Extract	+	+	+	+*	-	-
RB3M9 + Heated Extract	+	+	-	+‡	-	-

+ = Growth observed
- = No growth observed
*Microcolonies
‡Bacterial Lawn

PCR complementation ORF1 Target

First complementation product was isolated using primers designed with the Sanger capillary sequence at 3X coverage. The ORF1 primer targets a predicted molybdenum sulfite oxidase while the P-ORF1 targets the molybdenum sulfite oxidase including a nearby promoter region. Figure 24 shows an optimal temperature for the ORF1 product of 56.3°C and most optimal temperature for the P-ORF1 of 58.6°C.

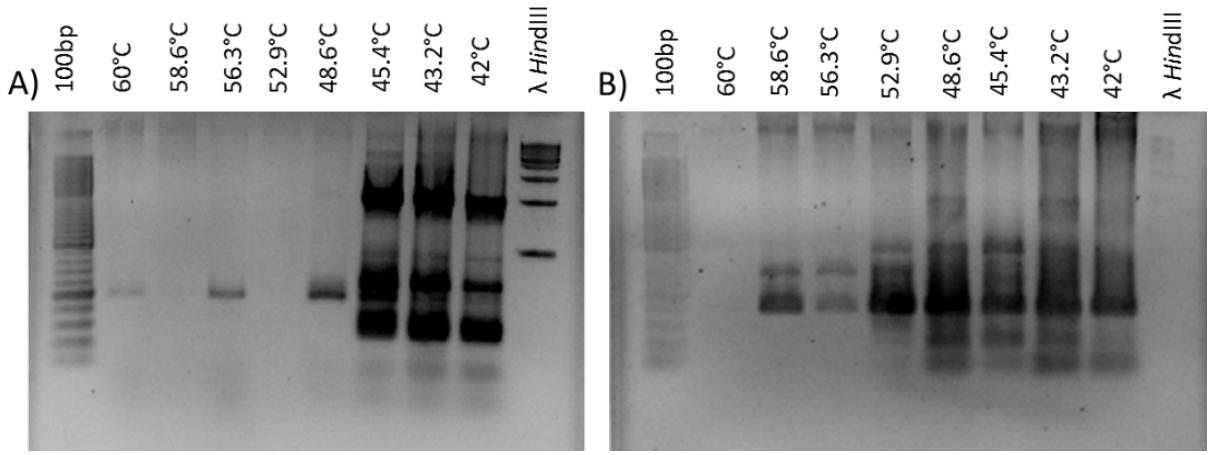

Figure 24. Gradient PCR for RB3M9 ORF target 1 and RB3M9 P-ORF target1. A) PCR gradient annealing temperature for ORF1 Target. B) PCR Gradient temperature for P-ORF1 Target.

The amplicon was then used for directional cloning and transformed into *E. coli* TOP10 cells as shown in figure 25, after 24h of incubation at 37°C in LBA with 50µg/mL kanamycin. After transformation, a miniprep extraction was prepared to verify the cloning by restriction enzyme analysis as shown in figure 26. In this case, no successful cloning was demonstrated by restriction analysis.

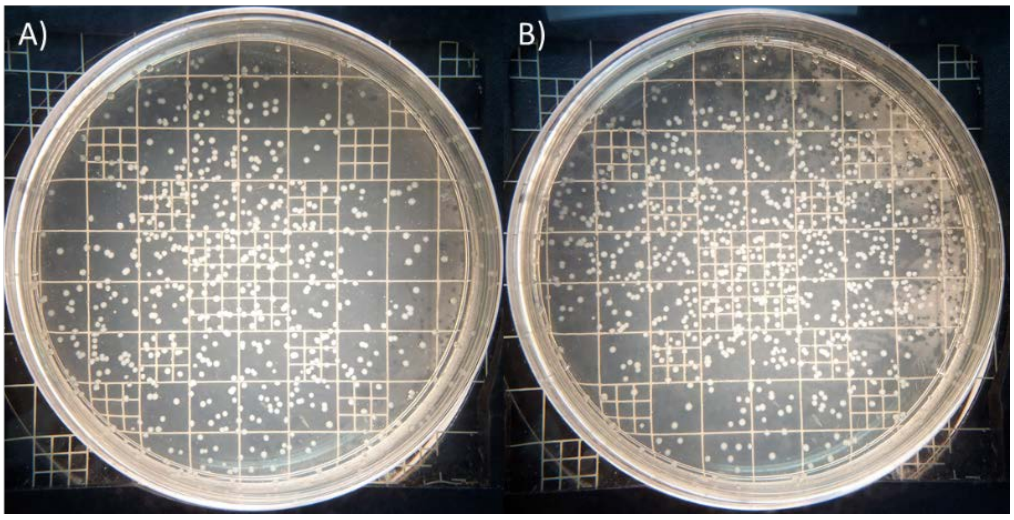

Figure 25. Transformation of pENTR cloned PCR products on pENTR™/SD/D-TOPO® on E. coli TOP10. A) pENTR-ORF1 clones, B) pENTR-P-ORF1 clones.

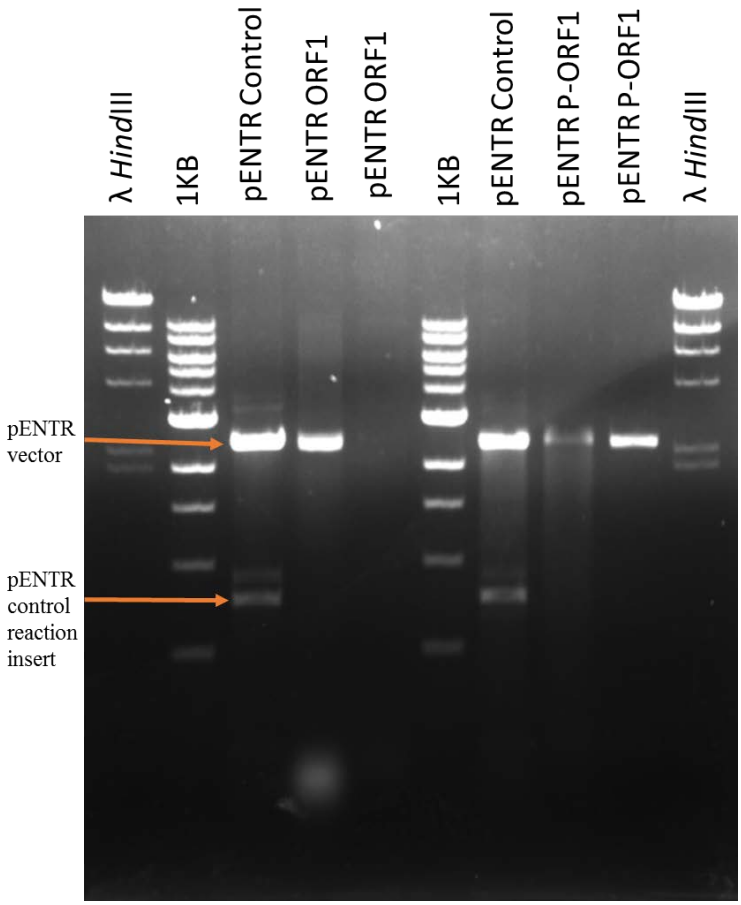

Figure 26. Restriction enzyme analysis of cloned fragments of pENTR-ORF1 and *pENTR-P-ORF1 by double digest using Not*I *and Eco*RV. No visible insert was found on over the double digestion, indicating that the plasmid didn't have the PCR product.

PCR complementation SO target

The SO primer set was designed using the Illumina sequence with a coverage of 100X and identified by the RAST annotation pipeline in complement of the Sanger capillary sequence obtained from mutant RB3M9 for the Sulfite Oxidase gene interrupted by the Tn5 insertion. PCR amplification of the expected 945bp product is shown in figure 27.

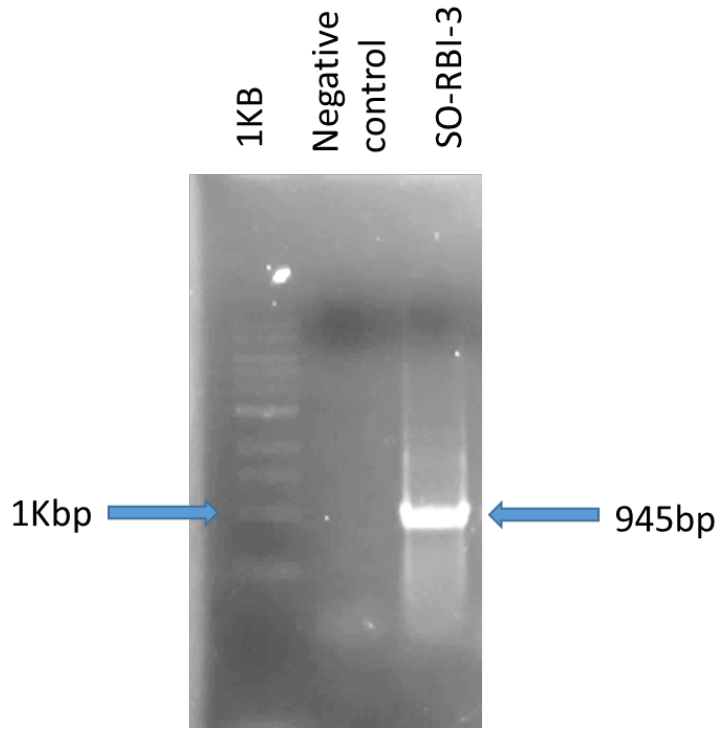

Figure 27. PCR amplification of the predicted Sulfite Oxidase gene interrupted by the Tn5 insertion found on the RB3M9 mutant. The amplification was obtained from the wild type counterpart found in the RBI-3 clone.

Re-transformation of the plasmid in *E. coli* EPI300 strain was successful as seen on figure 28. These clones were later used for the complementation analysis. In order to verify a successful cloning, the extracted plasmid was digested using *Not*I and *Eco*RV as seen on figure 29a. The insert found in the control reaction contains a 750bp product generated using the kit control protocol. The digested plasmid cloned with the product of interest shows a band of approximately 1115bp (945bp from the PCR amplicon and 170bp from the vectors restriction sites ends). A PCR amplification was performed using the SO-pENTR in order to confirm that the cloned insert was present in the plasmid before the complementation test was carried out, as shown in figure 29b.

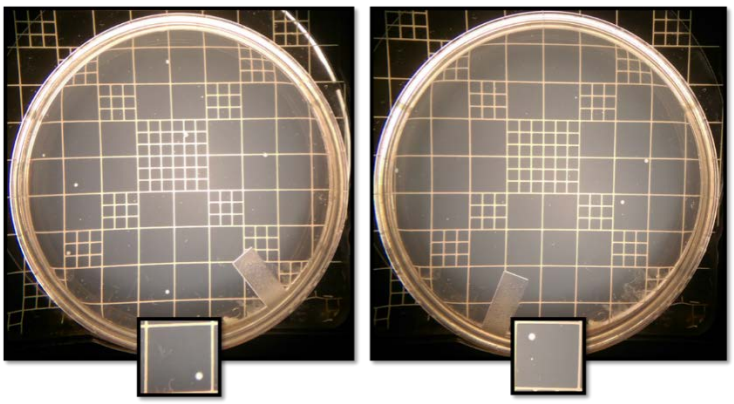

Figure 28. Transformation of SO PCR product cloned using the pENTR™/SD/D-TOPO® vector on E. coli EPI300 on LBA with 50µg/mL kanamycin.

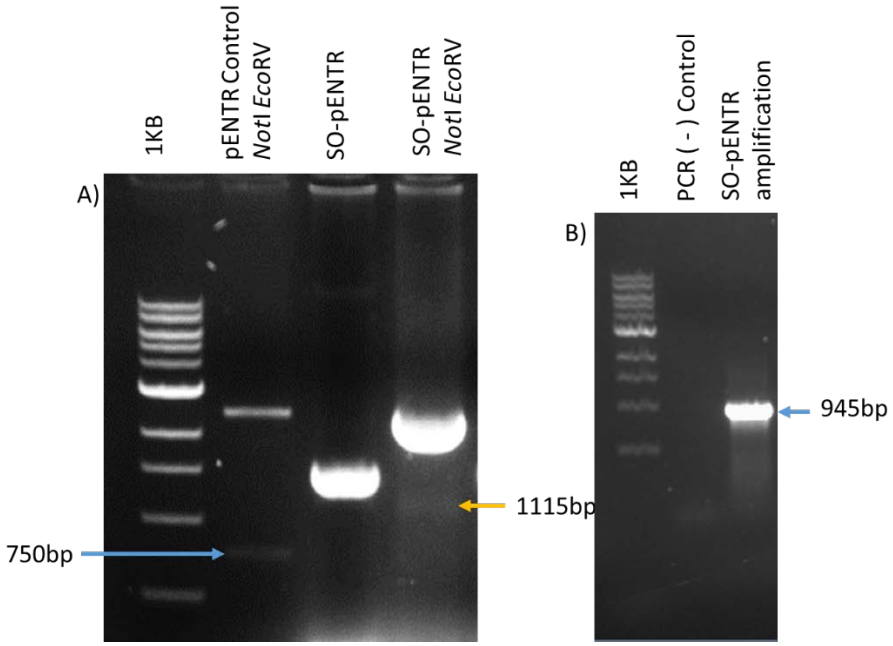

Figure 29. Restriction enzyme analysis of cloned fragment SO-pENTR by double digestion with NotI and EcoRV; along with the PCR amplification of the cloned fragment. A) Restriction analysis of the SO-pENTR plasmid. Arrows indicate cloned insert. B) PCR reaction using SO-RB3M9 primers amplification from the SO-pENTR clone.

Restriction fragment cloning

By digesting the fosmid with *Eco*RV and *Not*I, several fragments would be generated and eventually a fragmented library of the RBI-3 clone could be created. According to CLC Main

Workbench, two products of similar size were to be generated this way. Thus, the third enzyme *Xba*I was implemented in order to prevent the wrong product from cloning into the vector. Gel extraction product from the sizing obtained a total of 60ng/µL and shown in figure 30.

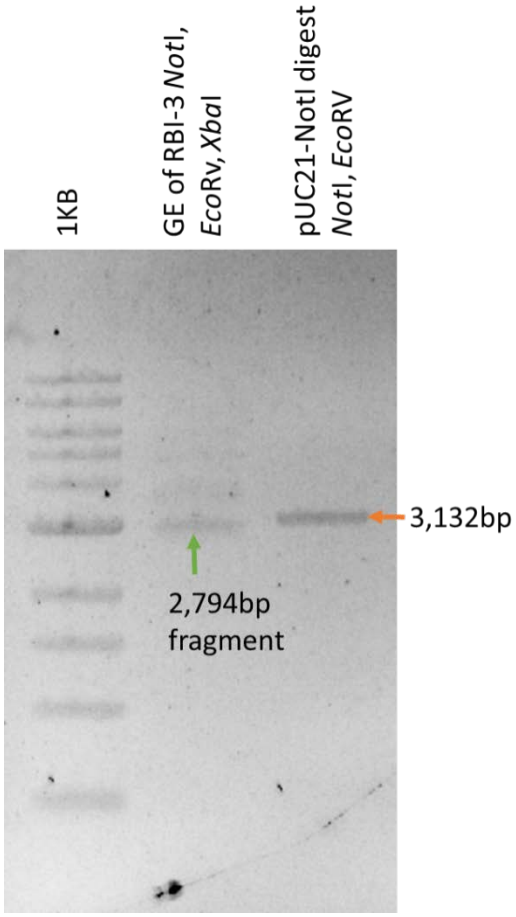

Figure 30. RBI-3 clone triple digestion sizing purification by Gel Extraction. GE indicated the gel extraction product obtained while the next lane shows the linearized plasmid after the double digestion.

Ligated clones from the gel extraction product were transformed and grown on LBA with 50µg/mL ampicillin as shown in figure 31. A total of 32 colonies were obtained using the T4DNA ligase, while no colonies were obtained when the ligase was absent.

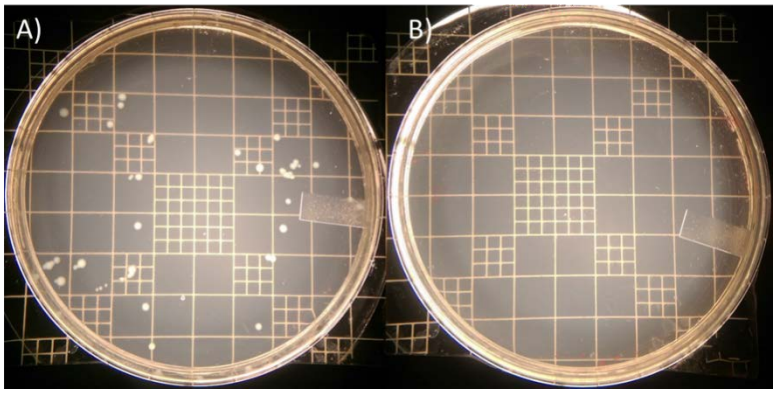

Figure 31. Transformation of the ligated product using the fragments purified by gel extraction. A) Transformants were grown on LBA with 50µg/mL ampicillin after overnight ligation. B) Control plate of the transformation reaction without the addition of T4DNA ligase enzyme.

Clones were then analyzed by gel electrophoresis as shown in figure 32. Out of the selected clones RB3TD1, RBTD2, RB3TD4 and RBTD5 appear to contain plasmid clone with the expected size from the gel extraction.

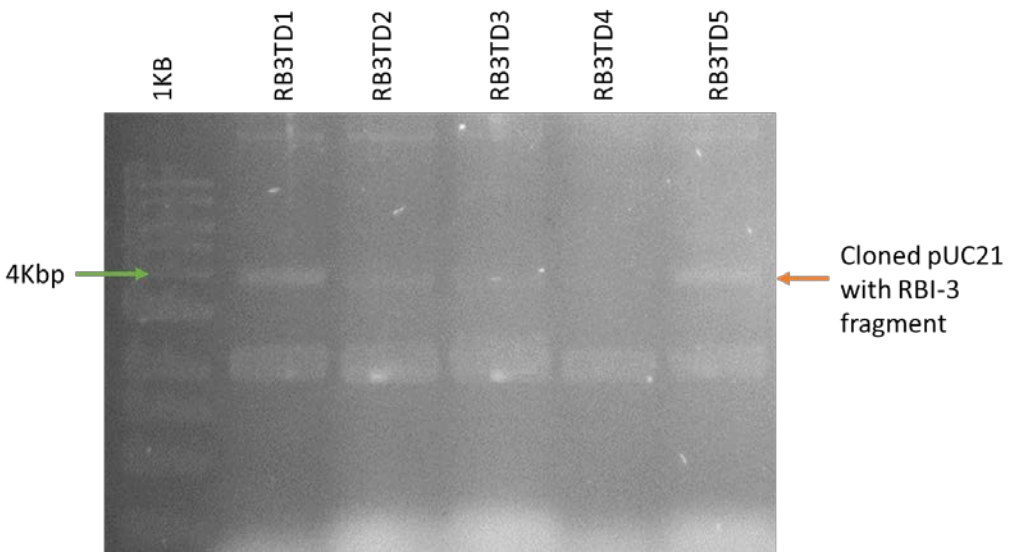

Figure 32. Miniprep extraction of the retransformants obtained from restriction enzyme ligation. The cloned plasmid is indicated by the orange arrow with an approximated size of 4Kbp when supercoiled.

The RB3TD1 plasmid was used for transformation along with the RB3M9 mutant; successful transformant started to form colonies after 30h of incubation as shown in figure 33.

Incubation was done over LBA with 50µg/mL ampicillin, 50ug/mL kanamycin and 12.5µg/mL chloramphenicol.

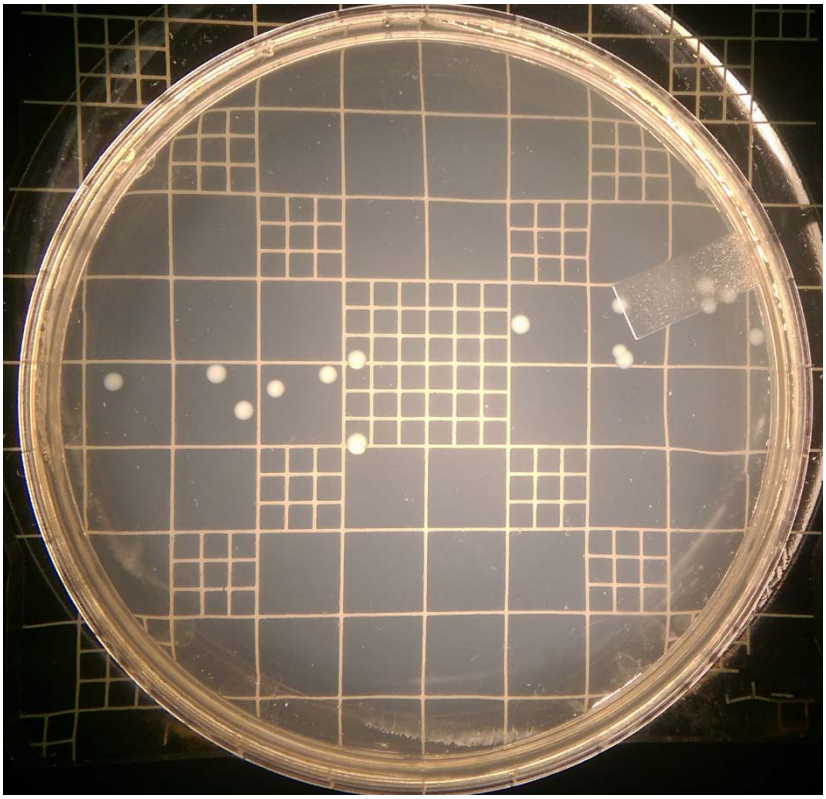

Figure 33. Transformation of RB3TD1 plasmid with RB3M9 mutant fosmid. Grown in LBA with 50µg/mL ampicillin, 50ug/mL kanamycin and 12.5µg chloramphenicol, after 24h of incubation at 37°C.

These clones were then cryogenically preserved. Four of these clones were selected randomly to verify the presence of the plasmid and fosmid by gel electrophoresis after a miniprep extraction (figure 34). Currently, three of the four candidates contain both fosmid and plasmid.

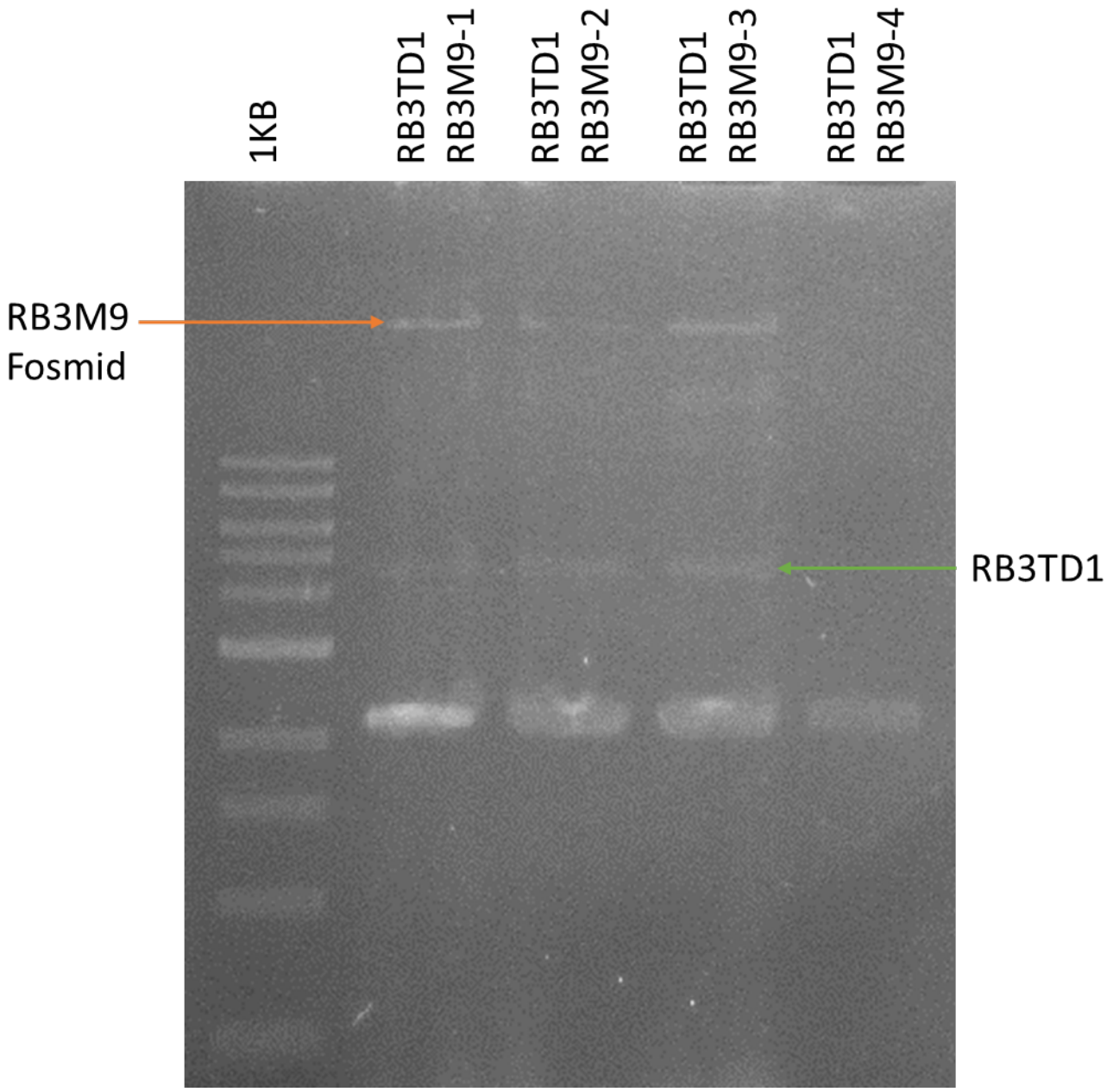

Figure 34. Miniprep extraction of transformants containing both fosmid and plasmid.

Complementation over minimal medium supplemented with CML

Clones from both PCR based cloning and restriction enzyme cloning that demonstrated to contain both fosmid and plasmid were patched over M9 minimal medium plate supplemented with

CML and left incubating for 48h as shown in figure 35. A total of four candidates managed to grow on the minimal medium, all derived from the PCR cloning methodology.

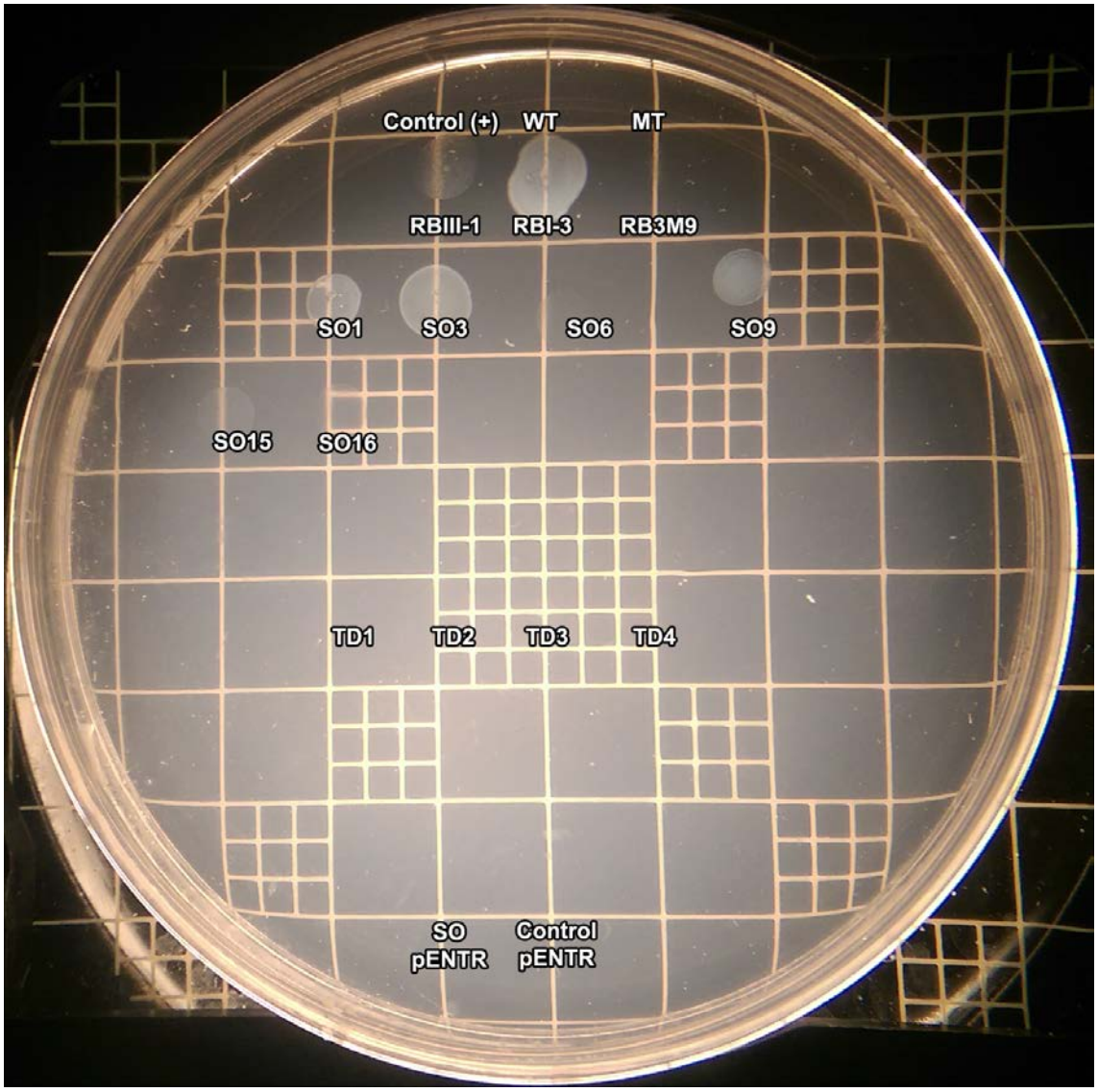

Figure 35. Trans-Complementation assay after 48h of incubation on minimal medium supplemented with CML. The SO (1-16) indicate clones containing both RB3M9 mutant and pENTR plasmid with PCR cloned product for the Sulfite Oxidase gene. TD (1-4) indicate clones containing both RB3M9 mutant and pUC21-*Not*I plasmid cloned with the restriction fragment. The RBIII-1 control belongs to a positive clone sub-pooled from previous screenings.

Clones SO1, SO3, SO9 and SO15 where induced using Epicentre CopyControl™ solution followed by a miniprep extraction as shown in figure 36. This indicated that indeed

these positive complements contained both mutant and the selected clone fragment. The clones generated by restriction enzyme cloning presented no activity after 48h.

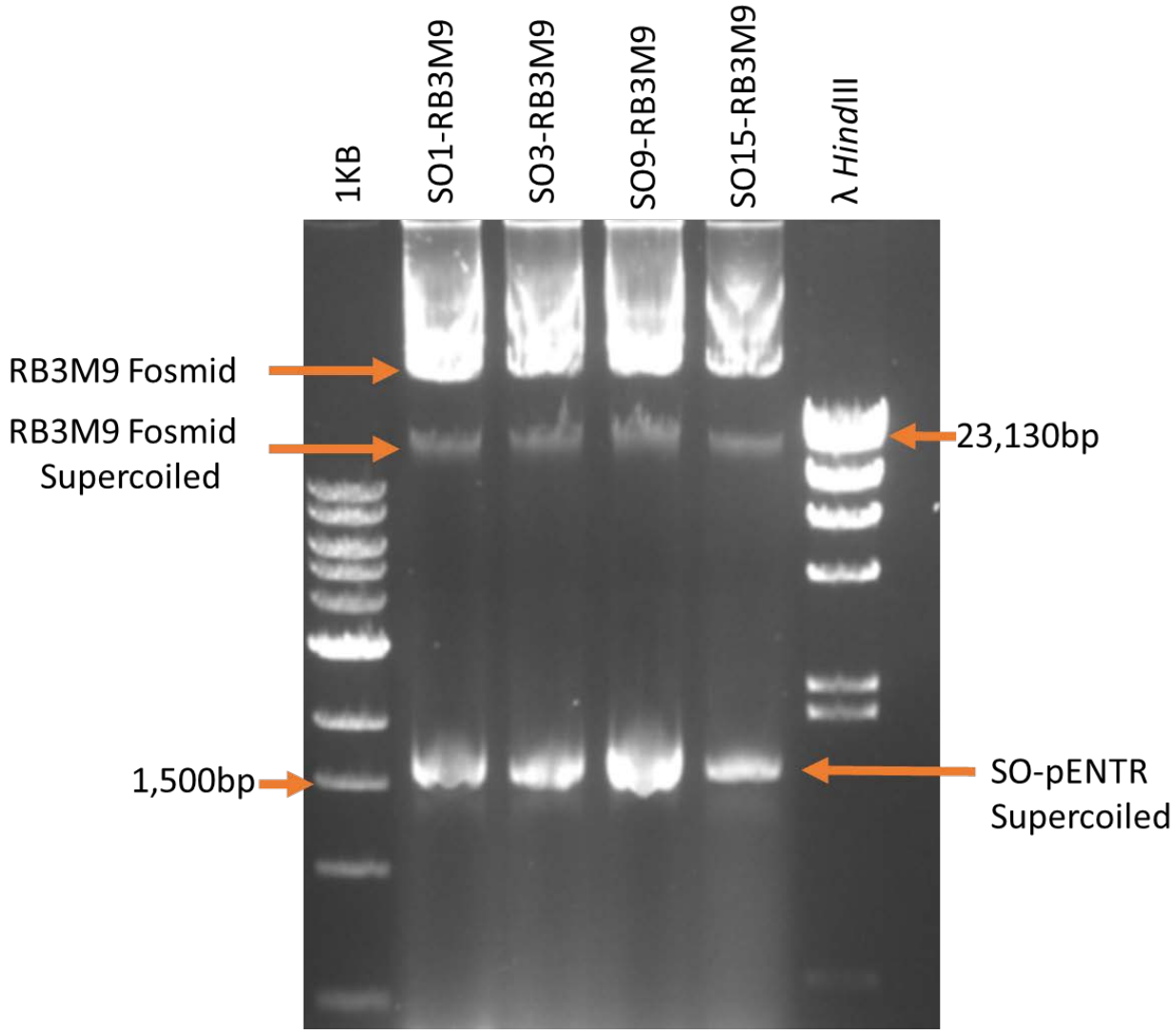

Figure 36. Miniprep extraction of induced complementation clones that resulted in positive growth on minimal medium supplemented with CML. As seen on the image, both plasmid and fosmid are present and indicated by the arrows.

4.4 DISCUSSION

Complementation has been used widely to study the functionality within bacterial genes for decades. Usually, these experiments are designed to work as complement of genetic information missing or damaged in the bacterial chromosome with a functional copy of an expression vector. There are different complementation strategies available such as, DNA repair based (Sedgwick et al., 1991), recombination events (Azizoglu et al., 2014), heterologous expression (Xiang & Moore, 2002) among others. In this case, we performed a complementation study based solely on extrachromosomal DNA by a heterologous expression system. In order to do so, one of the biggest challenges was to develop a system compatible with sustaining both fosmid and plasmid in the same cell. To adress this problem our system required different origins of replication for each of the vectors in order to avoid plasmid and fosmid incompatibility. The fosmid pCC1FOS contains an inducible origin of replication that maintains single copy number (*ori*V) until an induction solution is added to *E. coli* EPI300 strain that causes overexpression of the *trfA* gene. In contrast, the pENTR/SD/TOPO plasmid was engineered using pUC origin, a high copy plasmid (around 500 copies per cell). Having two different origins of replication allows these plasmids to be sustainable within the same host.

Before actually proceeding to a heterologous complementation approach. Understanding if the activity of extracellular proteins could interfere with our study was necessary.. Thus, a filtrate solution made from the conditioned medium obtained from the recently grown RBI-3 clone over the M9 minimal medium supplemented with CML was generated and used to perform a series of conditions listed in Table 5. However these conditions ended adding a new carbon source to the medium and no evidence of functionally was observed since it invalidate the minimal medium

approach by enriching the medium with the extracted protein solution discontinuing any further experimentation involving protein extracts.

Our first primers design set to be used with the pENTR system were generated by a low coverage (3X) Sanger capillary sequences. The pENTR/SD/TOPO contains a Shine-Dalgarno element and ribosome binding sites allowing cloned products to be expressed more efficiently. It also containes a directional cloning design for PCR cloning. The clone could later be transferred into a protein purification plasmid trough the Gateway cloning system. However the results from this first primer design was a complete failure due to the short and low coverage sequences used to generate them, as demonstrated in figure 26. In addition, figure 24 demonstrated that none of the amplicons obtained using these primers were of the desired size.

When the fosmid sequences obtained by Illumina were fully assembled, heyt showed several discrepancies with the Sanger capillary sequence, indicating that this could have contributed to the failure of the first primer set. Now that the whole sequence is known, the newly generated primer set managed to be properly cloned into the pENTR vector as seen in figure 29. Nevertheless this cloned product tends to be overexpressed due to the strong T7 promoter. To compensate the expression levels, the restriction enzyme cloning procedure using pUC21-*Not*I was implemented in order to study the whole gene including its predicted promoter region and prevent the possibilities of cloning multiple or incomplete products using PCR..

After generating our cells containing both plasmids and fosmids, they were patched on a minimal medium using a 48 square divider in order to place as many candidates as possible in the M9 minimal medium supplemented with CML. After 48h, four complemental hosts managed to grow using CML as carbon source. Yet none of the cloned products by restriction enzyme managed to grow on in the complementation assay. This could be due to a low expression phenomenom

generated from the plasmid used, or a weak promoter regulating the product expression of the gene. These results demonstrate that our predicted sulfite oxidase gene plays an important role in the biotransformation of the cells. However, little is known about the sulfite oxidase family in bacteria and all the possible metabolic reactions they might be involved in. Yet some of them appear to be involved in very different metabolic processes such as energy generation from sulfur compounds, host colonization, sulfite detoxification and organosulfonate degradation (Kappler, 2011).

This complementation assay also demonstrated that there was no genetic polarity effect being cause from the Tn5 insertion that could had affected the regulation gene found downstream of the sulfite oxidase gene. Yet this isolated gene alone demonstrated that it is not able to produce the observed phenotype in the WT. This indicates that more than one gene outside of this clone is involved in the mechanism. As the *in silico* analysis suggests, two different mutants were identified with a distance of 10Kbp from one another. Our results open the possibility of finding a single enzyme that could transform CML into a less unhealthy molecule and eventually generate a medical bioremediation therapy approach.

CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATIONS, BIBLIOGRAPHY AND LITERATURE CITED

5.1 SUMMARY

- Through the whole experiment, we managed to identify a novel function by a functional metagenomic approach in which clones managed to grow in a minimal medium containing N^{ε}-Carboxymethyl-Lysine as the sole active carbon source.
- The activity was found across different environments and probably contains different mechanisms of action.
- We managed to generate a full map of all the genes involved in the metagenomic insert of study by using the Illumina platform., The closest relatives obtained from protein phylogenetic analysis indicated that perhaps our donor is closely related to the microorganisms of the Rhizobia branch.
- A total of 57 genes were identified from this fragment using the RAST pipeline, 36 of them are considered hypothetical proteins.
- This information also helped to identify target genes involved with this novel function.
- We demonstrated that heterologous expression complementation is possible using a vector to vector system, establishing a methodology in functional metagenomics complementation in *E. coli* EPI300 host.
- Thanks to our complementation assay, we managed to isolate one of the genes involved in the biocatalytic pathway that allows our *E. coli* EPI300 to utilize N^{ε}-Carboxymethyl-Lysine as a sole carbon source.

5.2 CONCLUSION

N^{ε}-Carboxymethyl-Lysine, a well-known oxidative stress biomarker that in recent years has been shown to play an important role in the development of different pathologies, was used as sole carbon source for functional metagenomics screening. The ability of some clones from

different environments tells us about the diverse metabolisms present in bacteria that are currently uncharacterized. Being able to use this molecule opens up the possibility of generating an enzymatic treatment that might reduce the levels of CML present in affected tissues.

These findings open a new chapter for a medical bioremediation approach, an emerging field that has been actually tested only with cultivable organisms, land demonstrate the usefulness of functional metagenomic studies when it comes to finding novel products and solutions for today's problems. In combination with bioinformatics, we also managed to predict several genes that could be associated with the activity of interest. Among those predicted genes, we can find transporters, oxidases, regulatory complexes and dehydrogenases, and a large group of hypothetical proteins.

At the same time we validated the presence of an essential gene involved in the processes of biotransforming CML into something that *E. coli* can use to grow and generate energy by mechanisms of heterologous complementation between a plasmid and a fosmid. However, the family of the predicted sulfite oxidase is not that well characterized in prokaryotes, but it is speculated to play an important role in many metabolic activities. This newly isolated gene can eventually be used for both, understanding the sulfite oxidase roles in microorganisms, as well as its importance in the biotransformation of CML by biochemical approaches. That eventually could lead to the development of a possible treatment by combining protein engineering and nanotechnology.

5.3 RECOMMENDATIONS

- Performing a complementation assay for targets around the RB3M7 mutant would help us to understand the role of the genes in that region, since the distances between two mutants was around 10,000bp and in either point of mutation the activity is lost. A good candidate

is the predicted ABC-Transporter that could play a role in the intake of CML into *E. coli* EPI300.

- To clone the whole 10Kbp from mutant region RB3M7 to the mutant region of RB3M9 and see if *E. coli* EPI300 can grow in the M9 minimal medium supplemented with CML.
- Generate directed mutations to each of the hypothetical proteins found over the 10Kbp region could provide us with the essential genes required to create a synthetic operon and at the same time it would help us assign proper functions to these proteins.
- Construct clones containing only the mutated region of RB3M7 and only that of RB3M9 to test if they are able to confer the phenotype by themselves.
- Purify the protein product of the isolated gene and perform biochemical analysis that could help us to understand the mechanism of action by using High Performance Liquid Chromatography (HPLC).

5.5 BIBLIOGRAPHY AND CITED LITERATURE

Alcaide, M., Tchigvintsev, A., Martínez-Martínez, M., Popovic, A., Reva, O. N., Lafraya, Á., Bargiela, R., Nechitaylo, T. Y., Matesanz, R., Cambon-Bonavita, M.-A., Jebbar, M., Yakimov, M. M., Savchenko, A., Golyshina, O. V., Yakunin, A. F., Golyshin, P. N., & Ferrer, M. (2015). Identification and characterization of carboxyl esterases of gill chamber-associated microbiota in the deep-sea shrimp Rimicaris exoculata using functional metagenomics. *Applied and Environmental Microbiology*, AEM.03387–14. doi:10.1128/AEM.03387-14

Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A., & Handelsman, J. (2009). Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *The ISME Journal*, *3*(2), 243–251. doi:10.1038/ismej.2008.86

Azizoglu, R. O., Elhanafi, D., & Kathariou, S. (2014). Mutant Construction and Integration Vector-Mediated Gene Complementation in Listeria monocytogenes. *Methods in Molecular Biology (Clifton, N.J.)*, *1157*, 201–211.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*. doi:10.1089/cmb.2012.0021

Bartossek, R., Spang, A., Weidler, G., Lanzen, A., & Schleper, C. (2012). Metagenomic analysis of ammonia-oxidizing archaea affiliated with the soil group. *Frontiers in Microbiology*, *3*. doi:10.3389/fmicb.2012.00208

Berger, B. R., & Christie, P. J. (1994). Genetic complementation analysis of the Agrobacterium tumefaciens virB operon: virB2 through virB11 are essential virulence genes. *Journal of Bacteriology*, *176*(12), 3646–3660.

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, *277*(5331), 1453–1462. doi:10.1126/science.277.5331.1453

Castellani, R. J., Harris, P. L. R., Sayre, L. M., Fujii, J., Taniguchi, N., Vitek, M. P., Founds, H., Atwood, C. S., Perry, G., & Smith, M. A. (2001). Active glycation in neurofibrillary pathology of Alzheimer disease: Nε-(Carboxymethyl) lysine and hexitol-lysine. *Free Radical Biology and Medicine*, *31*(2), 175–180.

Claud, E. C., Keegan, K. P., Brulc, J. M., Lu, L., Bartels, D., Glass, E., Chang, E. B., Meyer, F., & Antonopoulos, D. A. (2013). Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. *Microbiome*, *1*(1), 20. doi:10.1186/2049-2618-1-20

Couto, G. H., Glogauer, A., Faoro, H., Chubatsu, L. S., Souza, E. M., & Pedrosa, F. O. (2010). Isolation of a novel lipase from a metagenomic library derived from mangrove sediment from the south Brazilian coast. *Genetics and Molecular Research : GMR*, *9*(1), 514–523. doi:10.4238/vol9-1gmr738

Cruz, J. M., Ortega, M. A., Cruz, J. C., Ondina, P., & Santiago, R. (2010). Unraveling activities by functional-based approaches using metagenomic libraries from dry and rain forest soils in Puerto Rico. *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology*, *2*(1), 1471–1478.

Dwyer, M. A., & Hellinga, H. W. (2004). Periplasmic binding proteins: A versatile superfamily for protein engineering. *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2004.07.004

Epicentre. (2011a). EZ-Tn5 <KAN-2> Insertion Kit. Retrieved from http://www.epibio.com/pdftechlit/134pl0611.pdf

Epicentre. (2011b). TransforMax ™ EPI300 ™ Chemically Competent E . coli. Retrieved from http://www.epibio.com/docs/default-source/protocols/transformax-epi300-electrocompetent-e-coli-transformax-epi300-chemically-competent-e-coli.pdf?sfvrsn=6

Fernández-Arrojo, L., Guazzaroni, M. E., López-Cortés, N., Beloqui, A., & Ferrer, M. (2010). Metagenomic era for biocatalyst identification. *Current Opinion in Biotechnology*. doi:10.1016/j.copbio.2010.09.006

Figueroa Matías, N. (2012). *Screening Soil Metagenomic Libraries Searching for Degrading Enzymes Useful for the Production of Biofuels and Value Added Chemicals*. University of Puerto Rico at Mayagüez.

Forbes, J. M., Cooper, M. E., Thallas, V., Burns, W. C., Thomas, M. C., Brammar, G. C., Lee, F., Grant, S. L., Burrell, L. A., Jerums, G., & Osicka, T. M. (2002). Reduction of the accumulation of advanced glycation end products by ACE inhibition in experimental diabetic nephropathy. *Diabetes*, *51*(12), 3274–3282.

Furber, J. (2010). The Future of Aging Therapies. (G. Fahy, M. West, L. S. Coles, & S. Harris, Eds.)*The Future of Aging SE - 19*. Springer Netherlands. doi:10.1007/978-90-481-3999-6

Gaens, K. H. J., Goossens, G. H., Niessen, P. M., Van Greevenbroek, M. M., Van Der Kallen, C. J. H., Niessen, H. W., Rensen, S. S., Buurman, W. a, Greve, J. W. M., Blaak, E. E., Van Zandvoort, M. A., Bierhaus, A., Stehouwer, C. D. a, & Schalkwijk, C. G. (2014). Nε-(carboxymethyl)lysine-receptor for advanced glycation end product axis is a key modulator of obesity-induced dysregulation of adipokine expression and insulin resistance. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *34*, 1199–1208. doi:10.1161/ATVBAHA.113.302281

Gaens, K. H. J., Niessen, P. M. G., Rensen, S. S., Buurman, W. A., Greve, J. W. M., Driessen, A., Wolfs, M. G. M., Hofker, M. H., Bloemen, J. G., Dejong, C. H., Stehouwer, C. D. A., & Schalkwijk, C. G. (2012). Endogenous formation of N??-(carboxymethyl)lysine is increased in fatty livers and induces inflammatory markers in an in vitro model of hepatic steatosis. *Journal of Hepatology*, *56*(3), 647–655. doi:10.1016/j.jhep.2011.07.028

Godzik, A. (2011, June). Metagenomics and the protein universe. *Current Opinion in Structural Biology*. Elsevier Ltd. doi:10.1016/j.sbi.2011.03.010

Goh, S.-Y., & Cooper, M. E. (2008). Clinical review: The role of advanced glycation end products in progression and complications of diabetes. *The Journal of Clinical Endocrinology and Metabolism*, *93*, 1143–1152. doi:10.1210/jc.2007-1817

Gorni, C. (2014). Transcriptomics : Open Access Metatranscriptomics in the NGS Era. *Transcriptomics*, *2*(1), 2013–2014. doi:10.4172/2329-8936.1000e107

Guazzaroni, M.-E., Silva-Rocha, R., & Ward, R. J. (2014). Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. *Microbial Biotechnology*. doi:10.1111/1751-7915.12146

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews : MMBR*, *68*(4), 669–685. doi:10.1128/MMBR.68.4.669-685.2004

Henne, A., Schmitz, R. A., Bömeke, M., Gottschalk, G., & Daniel, R. (2000). Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. *Applied and Environmental Microbiology*, *66*(7), 3113–3116. doi:10.1128/AEM.66.7.3113-3116.2000

Hernández-García, D., Wood, C. D., Castro-Obregón, S., & Covarrubias, L. (2010). Reactive oxygen species: A radical role in development? *Free Radical Biology & Medicine*, *49*(2), 130–43. doi:10.1016/j.freeradbiomed.2010.03.020

Hull, G. L. J., Woodside, J. V., Ames, J. M., & Cuskelly, G. J. (2012). N *-(carboxymethyl)lysine content of foods commonly consumed in a Western style diet. *Food Chemistry*, *131*(1), 170–174. doi:10.1016/j.foodchem.2011.08.055

Invitrogen. (2012). pENTR Directional TOPO Cloning, (25). Retrieved from https://tools.lifetechnologies.com/content/sfs/manuals/pentr_dtopo_man.pdf

Jiao, L., Stolzenberg-Solomon, R., Zimmerman, T. P., Duan, Z., Chen, L., Kahle, L., Risch, A., Subar, A. F., Cross, A. J., Hollenbeck, A., Vlassara, H., Striker, G., & Sinha, R. (2014). Dietary consumption of advanced glycation end products and pancreatic cancer in the prospective NIH-AARP Diet and Health Study. *American Journal of Clinical Nutrition*, *101*(1), 126–134. doi:10.3945/ajcn.114.098061

Kamata, K., Ozawa, Y., Kobayashi, T., & Matsumoto, T. (2009). Effect of N-epsilon-(carboxymethyl)lysine on coronary vasoconstriction in isolated perfused hearts from control and streptozotocin-induced diabetic rats. *Journal of Smooth Muscle Research = Nihon Heikatsukin Gakkai Kikanshi*, *45*(2-3), 125–137. doi:10.1540/jsmr.45.125

Kappler, U. (2011). Bacterial sulfite-oxidizing enzymes. *Biochimica et Biophysica Acta - Bioenergetics*. doi:10.1016/j.bbabio.2010.09.004

Karigar, C. S., & Rao, S. S. (2011). Role of microbial enzymes in the bioremediation of pollutants: a review. *Enzyme Research*, *2011*(Figure 1), 805187. doi:10.4061/2011/805187

Khromykh, A. A., Kenney, M. T., & Westaway, E. G. (1998). trans-Complementation of flavivirus RNA polymerase gene NS5 by using Kunjin virus replicon-expressing BHK cells. *Journal of Virology*, *72*(9), 7270–7279.

Kislinger, T., Fu, C., Huber, B., Qu, W., Taguchi, A., Yan, S. Du, Hofmann, M., Yan, S. F., Pischetsrieder, M., Stern, D., & Schmidt, A. M. (1999). N(ε)-(carboxymethyl)lysine adducts of proteins are ligands for receptor for advanced glycation end products that activate cell signaling pathways and modulate gene expression. *Journal of Biological Chemistry*, *274*(44), 31740–31749. doi:10.1074/jbc.274.44.31740

Knecht, K. J., Dunn, J. A., McFarland, K. F., McCance, D. R., Lyons, T. J., Thorpe, S. R., & Baynes, J. W. (1991). Effect of diabetes and aging on carboxymethyllysine levels in human urine. *Diabetes*, *40*(2), 190–196. doi:10.2337/diabetes.40.2.190

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*. Elsevier Inc. doi:10.1016/j.cell.2013.09.006

Kulasekara, H. D. (2014). Transposon mutagenesis. *Methods in Molecular Biology (Clifton, N.J.)*, *1149*, 501–19. doi:10.1007/978-1-4939-0473-0_39

Lam, K. N., Hall, M. W., Engel, K., Vey, G., Cheng, J., Neufeld, J. D., & Charles, T. C. (2014). Evaluation of a pooled strategy for high-throughput sequencing of cosmid clones from metagenomic libraries. *PLoS ONE*, *9*(6), 1–12. doi:10.1371/journal.pone.0098968

Lee, B. H., Hsu, W. H., Hsu, Y. W., & Pan, T. M. (2013). Suppression of dimerumic acid on hepatic fibrosis caused from carboxymethyl-lysine (CML) by attenuating oxidative stress depends on Nrf2 activation in hepatic stellate cells (HSCs). *Food and Chemical Toxicology*, *62*, 413–419. doi:10.1016/j.fct.2013.09.007

Li, L., Han, L., Fu, Q., Li, Y., Liang, Z., Su, J., & Li, B. (2012, January). Formation and Inhibition of Nε-(Carboxymethyl)lysine in Saccharide-Lysine Model Systems during Microwave Heating. *Molecules*. doi:10.3390/molecules171112758

Li, Y. C., Tsai, S. H., Chen, S. M., Chang, Y. M., Huang, T. C., Huang, Y. P., Chang, C. T., & Lee, J. A. (2012). Aristolochic acid-induced accumulation of methylglyoxal and Nε-(carboxymethyl)lysine: An important and novel pathway in the pathogenic mechanism for aristolochic acid nephropathy. *Biochemical and Biophysical Research Communications*, *423*(4), 832–837. doi:10.1016/j.bbrc.2012.06.049

Liebl, W., Angelov, A., Juergensen, J., Chow, J., Loeschcke, A., Drepper, T., Classen, T., Pietruzska, J., Ehrenreich, A., Streit, W. R., & Jaeger, K. E. (2014). Alternative hosts for functional (meta)genome analysis. *Applied Microbiology and Biotechnology*. doi:10.1007/s00253-014-5961-7

Lin, L., Park, S., & Lakatta, E. G. (2009). RAGE signaling in inflammation and arterial aging. *Frontiers in Bioscience : A Journal and Virtual Library*, *14*, 1403–1413. doi:10.2741/3315

Linder, J. U. (2006). Class III adenylyl cyclases: Molecular mechanisms of catalysis and regulation. *Cellular and Molecular Life Sciences*. doi:10.1007/s00018-006-6072-0

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*. doi:10.1155/2012/251364

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. a. (2003). Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics*, *19*(10), 1275–1283. doi:10.1093/bioinformatics/btg153

Lory, S., Wolfgang, M., Lee, V., & Smith, R. (2004). The multi-talented bacterial adenylate cyclases. *International Journal of Medical Microbiology : IJMM*, *293*, 479–482. doi:10.1078/1438-4221-00297

Mathieu, J. M., Schloendorn, J., Rittmann, B. E., & Alvarez, P. J. (2009). Medical bioremediation of age-related diseases. *Microbial Cell Factories*, *8*, 21. doi:10.1186/1475-2859-8-21

Mathieu, J., Schloendorn, J., Rittmann, B. E., & Alvarez, P. J. J. (2008). Microbial degradation of 7-ketocholesterol. *Biodegradation*, *19*(6), 807–813. doi:10.1007/s10532-008-9183-5

Meeprom, A., Sompong, W., Chan, C. B., & Adisakwattana, S. (2013). Isoferulic acid, a new anti-glycation agent, inhibits fructoseand glucose-mediated protein glycation in vitro. *Molecules*, *18*(6), 6439–6454. doi:10.3390/molecules18066439

Meng, S.-Y., & Bennett, G. N. (1992). Regulation of the Escherichia coli cad Operon: Location of a Site Requiered for Acid Induction. *Journal of Bacteriology*, *174*(8), 2670–2678.

Mizutari, K., Ono, T., Ikeda, K., Kayashima, K., & Horiuchi, S. (1997). Photo-enhanced modification of human skin elastin in actinic elastosis by N(epsilon)-(carboxymethyl)lysine, one of the glycoxidation products of the Maillard reaction. *The Journal of Investigative Dermatology*, *108*(5), 797–802. doi:10.1111/1523-1747.ep12292244

Monnier, V. M., Mustata, G. T., Biemel, K. L., Reihl, O., Lederer, M. O., Zhenyu, D., & Sell, D. R. (2005). Cross-linking of the extracellular matrix by the Maillard reaction in aging and diabetes: An update on “a puzzle nearing resolution.” In *Annals of the New York Academy of Sciences* (2005/07/23 ed., Vol. 1043, pp. 533–544). doi:10.1196/annals.1333.061

Murphy, S. L., Xu, J., Kochanek, K. D., & Statistics, V. (2013). National Vital Statistics Reports Deaths : Final Data for 2010. *U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES Centers for Disease Control and Prevention National Center for Health Statistics National Vital Statistics System*, *61*(4).

Myhre, M. R., Olsen, G. H., Gosert, R., Hirsch, H. H., & Rinaldo, C. H. (2010). Clinical polyomavirus BK variants with agnogene deletion are non-functional but rescued by trans-complementation. *Virology*, *398*(1), 12–20. doi:10.1016/j.virol.2009.11.029

New England Biolabs. (2014). Ligation Protocol with T4 DNA Ligase (M0202). New England Biolabs. Retrieved from https://www.neb.com/protocols/1/01/01/dna-ligation-with-t4-dna-ligase-m0202

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., McLean, J., Lasken, R., Clingenpeel, S. R., Woyke, T., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7821 LNBI, pp. 158–170). Springer Berlin Heidelberg. doi:10.1007/978-3-642-37195-0_13

Oh, C. J., Kim, H. B., & An, C. S. (2003). Molecular cloning and complementation analysis of nifV gene from Frankia EuIK1 strain. *Molecules and Cells*, *15*(1), 27–33.

Owen, J. G., Reddy, B. V. B., Ternei, M. a, Charlop-Powers, Z., Calle, P. Y., Kim, J. H., & Brady, S. F. (2013). Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11797–802. doi:10.1073/pnas.1222159110

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(8), 4285–4288. doi:10.1073/pnas.96.8.4285

Pollreisz, A., & Schmidt-Erfurth, U. (2010). Diabetic cataract-pathogenesis, epidemiology and treatment. *Journal of Ophthalmology*, *2010*, 608751. doi:10.1155/2010/608751

Prjibelski, A. D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., Pham, S., Korobeynikov, A., Lapidus, A., & Pevzner, P. A. (2014). ExSPAnder: A universal repeat

resolver for DNA fragment assembly. *Bioinformatics*, *30*(12), 293–301. doi:10.1093/bioinformatics/btu266

Procko, E., O'Mara, M. L., Bennett, W. F. D., Tieleman, D. P., & Gaudet, R. (2009). The mechanism of ABC transporters: general lessons from structural and functional studies of an antigenic peptide transporter. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *23*, 1287–1302. doi:10.1096/fj.08-121855

Promega. (2011). Assembly of Restriction Enzyme Digestion. *Technical Manual*, 1–8.

Rauschmeier, M., Schüppel, V., Tetsch, L., & Jung, K. (2014). New insights into the interplay between the lysine transporter LysP and the pH sensor CadC in escherichia coli. *Journal of Molecular Biology*, *426*, 215–229. doi:10.1016/j.jmb.2013.09.017

Riesenfeld, C. S., Goodman, R. M., & Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology*, *6*(9), 981–989. doi:10.1111/j.1462-2920.2004.00664.x

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics*, *38*, 525–552. doi:10.1146/annurev.genet.38.072902.091216

Rocha, E. P. C. (2008). The organization of the bacterial genome. *Annual Review of Genetics*, *42*, 211–233. doi:10.1146/annurev.genet.42.110807.091653

Romano, A. D., Serviddio, G., De Matthaeis, A., Bellanti, F., & Vendemiale, G. (2010). Oxidative stress and aging. *Journal of Nephrology*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20872368

Roy, C. N., Semba, R. D., Sun, K., Bandinelli, S., Varadhan, R., Patel, K. V, Guralnik, J. M., & Ferrucci, L. (2012). Circulating selenium and carboxymethyl-lysine, an advanced glycation endproduct, are independent predictors of anemia in older community-dwelling adults. *Nutrition*, *28*(7-8), 762–766. doi:10.1016/j.nut.2011.11.005

Schleicher, E. D., Wagner, E., & Nerlich, A. G. (1997). Increased accumulation of the glycoxidation product Nε- (carboxymethyl)lysine in human tissues in diabetes and aging. *Journal of Clinical Investigation*, *99*(3), 457–468. doi:10.1172/JCI119180

Schloendorn, J., Webb, T., Kemmish, K., Hamalainen, M., Jackemeyer, D., Jiang, L., Mathieu, J., Rebo, J., Sankman, J., Sherman, L., Tontson, L., Qureshi, A., Alvarez, P., & Rittmann, B. (2009). Medical bioremediation: a concept moving toward reality. *Rejuvenation Research*, *12*(6), 411–419. doi:10.1089/rej.2009.0917

Schloss, P. D., & Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology*. doi:10.1016/S0958-1669(03)00067-3

Schloss, P. D., & Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology*, *6*(8), 229. doi:10.1186/gb-2005-6-8-229

Sedgwick, S. G., Lodwick, D., Doyle, N., Crowne, H., & Strike, P. (1991). Functional complementation between chromosomal and plasmid mutagenic DNA repair genes in bacteria. *Molecular & General Genetics : MGG*, *229*, 428–436.

Sell, D. R., & Monnier, V. M. (2011). Aging of Long-Lived Proteins: Extracellular Matrix (Collagens, Elastins, Proteoglycans) and Lens Crystallins. In C. W. R. U. Institute of Pathology (Ed.), *Comprehensive Physiology* (pp. 235–305). Ohio: WILEY.

Semba, R. D., Nicklett, E. J., & Ferrucci, L. (2010). Does accumulation of advanced glycation end products contribute to the aging phenotype? *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *65 A*(9), 963–975. doi:10.1093/gerona/glq074

Sharma, P., Jindal, S., Bala, K., Kumari, K., Niharika, N., Kaur, J., Pandey, G., Pandey, R., Russell, R. J., Oakeshott, J. G., & Lal, R. (2014). Functional screening of enzymes and bacteria for the dechlorination of hexachlorocyclohexane by a high-throughput colorimetric assay. *Biodegradation*, *25*, 179–187. doi:10.1007/s10532-013-9650-5

Shin, H. S., McCormick, P., Artzt, K., & Bennett, D. (1983). Cis-trans test shows a functional relationship between non-allelic lethal mutations in the T/t-complex. *Cell*, *33*(3), 925–929. doi:10.1016/0092-8674(83)90035-1

Simon, C., Herath, J., Rockstroh, S., & Daniel, R. (2009). Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Applied and Environmental Microbiology*, *75*(9), 2964–2968. doi:10.1128/AEM.02644-08

Singh, B. K., & Macdonald, C. A. (2010). Drug discovery from uncultivable microorganisms. *Drug Discovery Today*. doi:10.1016/j.drudis.2010.07.002

Singh, R., Barden, A., Mori, T., & Beilin, L. (2001). Advanced glycation end-products: A review. *Diabetologia*. doi:10.1007/s001250051591

Singh, R., Kishore, L., & Kaur, N. (2014, December 25). Diabetic peripheral neuropathy: Current perspective and future directions. *Pharmacological Research*. Elsevier Ltd. doi:10.1016/j.phrs.2013.12.005

Stockwell, B. R. (2000). Chemical genetics: ligand-based discovery of gene function. *Nature Reviews Genetics*, *1*(2), 116–125. Retrieved from http://dx.doi.org/10.1038/35038557

Torres-Zapata, I. (2012). *Generation of large-insert metagenomic libraries from subtropical hypersaline microbial mats and their screening for antibiotic resistance by*. University of Puerto Rico at Mayaguez.

Toth, C., Rong, L., Yang, C., & Martinez, J. (2008). Receptor for advanced glycation end products (RAGEs) and experimental diabetic neuropathy. *Diabetes*, *57*(April 2008). doi:10.2337/db07-0339.Additional

Troeschel, S. C., Drepper, T., Leggewie, C., Streit, W. R., & Jaeger, K.-E. (2010). Novel tools for the functional expression of metagenomic DNA. *Methods in Molecular Biology (Clifton, N.J.)*, *668*, 117–139. doi:10.1007/978-1-60761-823-2_8

U.S. Department of Agriculture. (2013). The Cloud forest of the El Yunque National Forest. Retrieved from www.fs.usda.gov/elyunque

Uchiyama, T., & Miyazaki, K. (2009, December). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current Opinion in Biotechnology*. doi:10.1016/j.copbio.2009.09.010

Van Dam, P. S., Cotter, M. a, Bravenboer, B., & Cameron, N. E. (2013, November 5). Pathogenesis of diabetic neuropathy: Focus on neurovascular mechanisms. *European Journal of Pharmacology*. Elsevier. doi:10.1016/j.ejphar.2013.07.017

Vicencio, J. M., Galluzzi, L., Tajeddine, N., Ortiz, C., Criollo, A., Tasdemir, E., Morselli, E., Ben Younes, A., Maiuri, M. C., Lavandero, S., & Kroemer, G. (2008). Senescence, apoptosis or autophagy? When a damaged cell must decide its path - A mini-review. *Gerontology*, *54*(2), 92–99. doi:10.1159/000129697

Vlasblom, J., Zuberi, K., Rodriguez, H., Arnold, R., Gagarinova, A., Deineko, V., Kumar, A., Leung, E., Rizzolo, K., Samanfar, B., Chang, L., Phanse, S., Golshani, A., Greenblatt, J. F., Houry, W. A., Emili, A., Morris, Q., Bader, G., & Babu, M. (2014). Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli. *Bioinformatics*, *31*(3), 306–310. doi:10.1093/bioinformatics/btu671

Voth, W. P., Richards, J. D., Shaw, J. M., & Stillman, D. J. (2001). Yeast vectors for integration at the HO locus. *Nucleic Acids Research*, *29*(12), E59–E59. doi:10.1093/nar/29.12.e59

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. doi:10.1038/nrg2484

Wenzel, S. C., & Müller, R. (2005). Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Current Opinion in Biotechnology*. doi:10.1016/j.copbio.2005.10.001

Wexler, M., Bond, P. L., Richardson, D. J., & Johnston, A. W. B. (2005). A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environmental Microbiology*, *7*, 1917–1926. doi:10.1111/j.1462-2920.2005.00854.x

Xiang, L., & Moore, B. S. (2002). Inactivation, complementation, and heterologous expression of encP, a novel bacterial phenylalanine ammonia-lyase gene. *Journal of Biological Chemistry*, *277*, 32505–32509. doi:10.1074/jbc.M204171200

Yan, H., Li, X., Xie, J., & Li, M. (2007). Effects of advanced glycation end products on renal fibrosis and oxidative stress in cultured NRK-49F cells. *Chinese Medical Journal*, *120*, 787–793.

Yergeau, E., Lawrence, J. R., Sanschagrin, S., Waiser, M. J., Korber, D. R., & Greer, C. W. (2012). Next-generation sequencing of microbial communities in the athabasca river and its tributaries in relation to oil sands mining activities. *Applied and Environmental Microbiology*, *78*(21), 7626–7637. doi:10.1128/AEM.02036-12

Zahran, H. H. (2001). Rhizobia from wild legumes: Diversity, taxonomy, ecology, nitrogen fixation and biotechnology. *Journal of Biotechnology*, *91*(2-3), 143–153. doi:10.1016/S0168-1656(01)00342-X

Zhao, H., Dreses-Werringloer, U., Davies, P., & Marambaud, P. (2008). Amyloid-beta peptide degradation in cell cultures by Mycoplasma contaminants. *BMC Research Notes*, *1*, 38. doi:10.1186/1756-0500-1-38