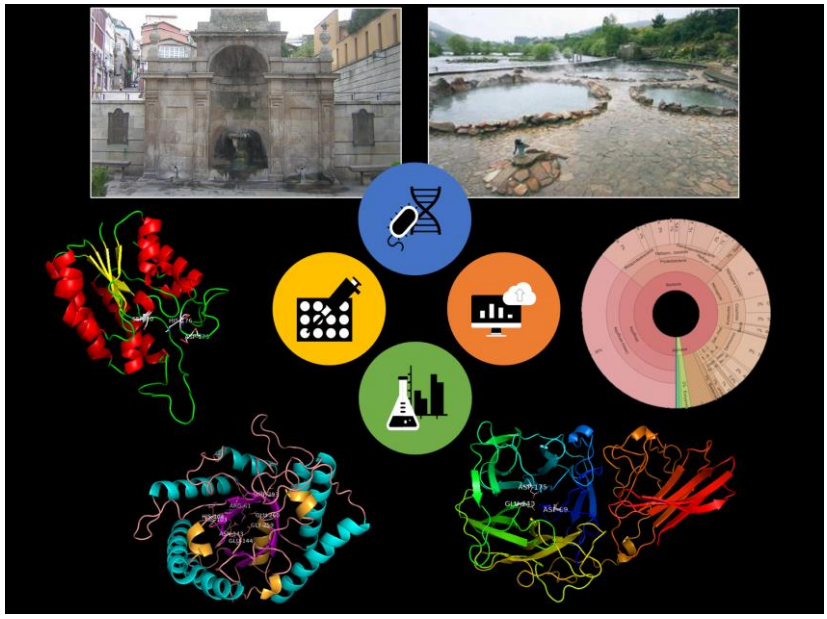
Search for novel thermozymes in metagenomic libraries from hot springs

Autor: Juan José Escuder Rodríguez

Tesis doctoral UDC / 2021



Directores: Mª Isabel González Siso y Manuel Becerra Fernández

Tutor: Manuel Becerra Fernández

Programa oficial de doctorado en Biología Celular y Molecular

UNIVERSIDADE DA CORUÑA

Cover image: Graphical abstract of the whole metagenomic process. Metagenomic DNA extraction (blue circle) from two hot spring sources depicted in the photrographs. Functional screening of metagenomic libraries (yellow circle), with a lipolytic thermozyme that was modelled in silico (leftmost protein model). Bioinformatics analysis of the metagenomes (orange circle), allowing the description of the microbial community (pie chart of taxonomical assignments of reads) and identification of thermozymes, including a endoglucanase (protein model in the downleft position) and a xylosidase (protein model in the downright position). Expression, purification and biochemical characterization of thermozymes (green circle) of thermozymes found, including the endoglucanase and the lipolytic thermozyme mentioned.

El presente trabajo: **Search for novel thermozymes in metagenomic libraries from hot springs**, presentado por D. Juan José Escuder Rodríguez para aspirar al grado de Doctor en Biología Celular y Molecular, con mención de Doctorado Internacional, ha sido realizado bajo nuestra dirección en el departamento de Biología (laboratorio de Bioquímica y Biología Molecular) de la Universidade da Coruña.

Revisado el texto, estamos conformes con su presentación para ser juzgado.

A Coruña, ____ de Septiembre de 2021

V°B°

Dra. María Isabel González Siso

Catedrática de Bioquímica y Biología Molecular

V°B°

Dr. Manuel Becerra Fernández

Profesor de Bioquímica y Biología Molecular

El presupuesto general del laboratorio ha sido financiado por el "*Programa de Axudas para a Consolidación e a Estruturación de Unidades de Investigación Competitivas do Sistema Galego de I+D+I*" Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia) para el grupo EXPRELA co-financiada con fondos FEDER. Proyectos de 2012-2015, 2017-2019 y 2020-2023. También se ha recibido financiación por el programa "Ayuda para la creación, reconocimiento y estructuración de agrupaciones estratégicas del sistema universitario de Galicia (agrupación CICA-INIBIC)" financiado por la Xunta de Galicia. Proyectos de 2015-2017 y 2018-2020. También se ha recibido financiación del 7º programa marco de la Unión Europea, como acción Marie Curie FP7-PEOPLE-2012-IAPP proyecto *HotDrops* No. 324439) financiado por la Unión Europea. Proyecto de 2013-2017.

Juan José Escuder Rodríguez ha realizado su trabajo como personal investigador contratado y su estancia de investigación gracias a dichas ayudas.

A mis padres y hermano

All that is gold does not glitter,
not all those who wander are lost;
the old that is strong does not wither,
deep roots are not reached by the frost.
J. R. R. Tolkien

ACKNOWLEDGMENTS / AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a mis directores de tesis Manuel e Isabel la oportunidad que me han dado para investigar en un campo tan apasionante de la bioquímica como es la metagenómica. Gracias por todos los consejos, indicaciones, ideas y reuniones que han ido dándole forma a este trabajo ya que sin vuestra guía no hubiera sido posible.

Por supuesto también tengo mucho que agradecer a todo el grupo de bioquímica de la facultad de ciencias, a Esperanza, Esther, Ana y Marian, quienes siempre me han ayudado y apoyado durante estos años en el laboratorio. Gracias porque el esfuerzo que demostráis día a día es inspirador.

A Mónica por tener tanta paciencia con todos, por esos pedidos a destiempo y todas las incidencias que nos has tramitado, pero sobre todo por los buenos momentos fuera del laboratorio, donde tu gran iniciativa también resolvió muchas veces nuestra no tan buena organización. A Ángel, nuestro sabio al que acudir cuándo las ideas se acaban, y no soy el primero ni seré el último en decirlo. Por esos más de mil cafés que si echo cuentas habremos compartido (y me quedo corto) con debates de toda índole o simplemente comentando la actualidad, cuándo no estamos hablando de nuestra afición común a la que dedicarle el tiempo y dinero que (no) podemos. A Agustín, por toda la ayuda que me diste cuando aún daba mis primeros pasos como alumno de máster en el laboratorio y corrigiendo esos errores de novato. Y por todas las quedadas en las que siempre por algún motivo estábamos en el grupo de los puntuales. A Olalla, nuestra precursora en esto de la metagenómica, nos allanaste el camino en frentes donde hubiera sido muy difícil trabajar de otra manera. Y por supuesto todo lo vivido fuera del laboratorio, las quedadas, los planes improvisados y los no improvisados. También tengo que mencionar tu talento para todo tipo de manualidades y tareas artesanales (a día de hoy aún se conserva ese taller de velas y jabones). A Mari, al principio en alguna ocasión tuviste que tener paciencia conmigo,

pero puede que de ahí venga lo bien que nos llegamos a entender después. Gracias por todos los buenos momentos tanto dentro como fuera del laboratorio y por confiar en mí para cuidar al bueno de Pancho. A Mariu, principal artífice de que mi estancia fuera en San Diego, y compañera de aventuras y desventuras en el mundo de los metagenomas (y de cierto viaje a Cambridge). Siempre le pones buena cara a todo y por eso es tan fácil hablar contigo, incluso cuando tienes que multiplicarte por tres para dar abasto. A María "pequeña", porque tu alegría es contagiosa. Por aguantarme de compañero de congresos en Barcelona y Santander, con esas visitas turísticas que convalidan con rutas de senderismo nivel medio-alto. Porque con tus anécdotas uno nunca puede aburrirse. A mis compis del team CICA, Martín y Lidia, con los que no he compartido tanto laboratorio, motivo por el que cuando nos visitáis siempre nos tenemos algo que contar, aunque a veces sea una canción que no nos podemos sacar de la cabeza en todo el día. Gracias por ayudarme con esos western cuando he sido yo el que ha ido de visita. A la más reciente incorporación de los "metagenómicos", Almudena, sin esfuerzo has encajado en el grupo como si llevaras aquí desde el principio. A los que pasaron por el laboratorio como alumnos con los que trabaje más directamente (¡y espero no olvidarme de nadie!), Alba, Natalí, Ariel, Diego, Tania, Jaime, Edu y Javier. A Ana, Nadia y Flor que pasaron poco tiempo por el laboratorio pero que espero se llevaran tan buen recuerdo como el que dejaron. A Ángela, por tu amistad, por ese verano en Galicia y por los viajes a Murcia e Italia. A todas las personas que más directa o indirectamente me han ayudado en la facultad de ciencias, Rosa, Kike y Elia. A mis compañeros de Ourense encabezados por la doctora Mª Luisa Rua, y a los de Madrid por los doctores José Berenguer y Aurelio Hidalgo, por los cursos sobre termófilos, reuniones de la REDEX, y los consejos y observaciones que acabaron plasmados también en este libro.

I would also like to thank fellow researchers Moussa, Ouassila, Amine and Maya that visited us from Argelia resulting in very fruitful collaborations and friendships. I thank also Kamila who stayed in the

laboratory and provided me with so much insight into metagenomics and the bioinformatic analysis, and for all the good memories we shared outside work. Thamer, we did not get to share much time mostly due to the pandemic and the different workplace, although we did have quite a number of conversations over the phone (mostly revolving on Angel's whereabouts). I would like to also thank doctor Liisa van Vliet at the University of Cambridge, for her workshop and for the fun we had in our short visit there. I also have only gratitude for Rob Edwards at San Diego State University and his talented team, specially Geni who helped me so much with everything, Adrian and his bioinformatics course, Daniel, and the friends I made in California.

A mis amigos, que me han apoyado siempre, a la OlimoBand de Coruña y a la BlackList de León. Ya podéis decir ¡por fin! A mi familia, mis primos y primas y tíos y tías, vuestro apoyo y cariño me ha motivado para seguir y mejorar. Por escuchar esas explicaciones "en términos que yo me entere de lo que estás haciendo".

A Aida, son demasiadas cosas como para hacer una lista, pero sabes que te lo agradezco todo. Gracias por una amistad como si nos conociéramos desde niños.

A mi abuela Luisa, a quien adoro y admiro, gracias se me queda pequeño.

A mi abuela Angélica, que sé que está orgullosa de su nieto.

A mi hermano, por los años viviendo juntos y por inventar eso de "Tales todo el día", por compartir desayunos espaciales y las visitas de fin de semana. Por aguantar a tu hermano cuando no hay quien le aguante, y poder decir eso de "esto... esto... no te guardo rencor". Gracias Víctor.

Por último, a mis padres, a Cande y a Juan, a los que os debo todo. Tenéis todas las cualidades que admiro y siempre me habéis demostrado todo vuestro apoyo y cariño. Gracias por todo lo que habéis hecho por mí todos estos años.

ABSTRACT

Hot spring metagenomes are described as the DNA of whole microbial communities inhabiting high temperature inland water habitats. The focus of the present work was to search for thermozymes (that maintain their activity at high temperatures) encoded in such metagenomes, as they find biotechnological applications in all kinds of industries. Additional information like the metabolic interactions between community members have been studied as well. The first step of the metagenomic study consisted in the sampling of two hot springs: As Burgas and Muiño da Veiga. The metagenomic DNA from these samples was extracted and used to construct two metagenomic libraries. These libraries were screened to detect potentially relevant enzymatic activities using a wide range of substrates, resulting in the discovery of a novel thermozyme with lipolytic activity (LipB12_A11) which was biochemically characterized. The metagenomic DNA from one of the hot springs was also sequenced with a shotgun sequencing strategy. This led to the description of the microbial community from the hot spring, based on sequence reads abundance and taxonomical and functional assignment. Reads assembly into larger sequences allowed the prediction of several putative proteins with biotechnological interest. Of these, a novel endoglucanase (Cel776) was expressed and biochemically characterized.

RESUMEN

Los metagenomas de aguas termales consisten en el conjunto del ADN de comunidades microbianas procedentes de hábitats de aguas continentales a altas temperaturas. El objetivo de este trabajo es la búsqueda en estos metagenomas de termozimas (que mantienen su actividad a altas temperaturas) codificadas en ellos, ya que tienen aplicaciones biotecnológicas en todo tipo de industrias. Además, se han estudiado las interacciones metabólicas que existen entre los miembros de la comunidad microbiana. El primer paso en el estudio metagenómico ha sido el muestreo de dos fuentes termales: As Burgas y Muiño da Veiga. Se extrajo el ADN metagenómico de estas muestras y se construyeron dos metagenotecas. Se realizó una búsqueda funcional usando un abanico de sustratos en las metagenotecas para encontrar actividades enzimáticas potencialmente relevantes, y se caracterizó bioquímicamente una nueva termozima con actividad lipolítica (LipB12_A11). Además, se secuenció el ADN metagenómico de una de las fuentes termales con una estrategia *shotgun*. En base a la abundancia de las lecturas y su asignación taxonómica y funcional se describió la comunidad microbiana de la fuente termal. Las lecturas se ensamblaron en secuencias más largas para predecir proteínas putativas con interés biotecnológico. De ellas, se expresó y caracterizó una nueva endoglucanasa (Cel776).

RESUMO

Os metaxenomas de augas termais consisten no conxunto do ADN de comunidades microbianas de hábitats de augas continentais a alta temperatura. O obxectivo deste traballo é a procura nestes metaxenomas de termoencimas (que manteñen a súa actividade a altas temperaturas) que neles se codifican, xa que teñen aplicacións biotecnolóxicas en todo tipo de industrias. Ademáis, estudáronse as interaccións metabólicas que existen entre os membros da comunidade microbiana. O primeiro paso do estudo metaxenómico foi a toma de mostras de dúas fontes termais: As Burgas e Muiño da Veiga. Destas mostras foi extraído o ADN metaxenómico e construíronse dúas metaxenotecas. Realizouse unha busca funcional empregando unha serie de substratos nas metaxenotecas para atopar actividades encimáticas potencialmente relevantes e caracterizouse bioquímicamente unha nova termocima con actividade lipolítica (LipB12_A11). Ademais, o ADN metaxenómico dunha das augas termais secuenciouse mediante unha estratexia de *shotgun*. Baseado na abundancia de lecturas e na súa asignación taxonómica e funcional, describimos a comunidade microbiana da fonte termal. As lecturas reuníronse en secuencias máis longas para predicir proteínas con suposto interese biotecnolóxico. Destas, expresouse e caracterizouse unha nova endoglucanasa (Cel776).

PREFACE

PREFACE

On thermophiles and industrial application of thermozymes

Extremophiles are microorganisms adapted to life in habitats where a certain physico-chemical characteristic is considered extreme, meaning beyond the typical limits found in most of Earth biosphere. Thermophiles are a particular group within extremophilic microorganisms that can thrive at high temperatures (above 50°C, or in the case of hyperthermophiles above 80°C) (1). In contrast, organisms that can survive these high temperatures but naturally live at lower temperatures are considered thermotolerant mesophiles.

These organisms are found in a variety of natural environments, including continental hot springs (2), submarine hydrothermal vents (3), deserts and volcanic fumaroles and solfatara fields, among others; and habitats of human origin or construction, such as compost and artificial hot water streams (4).

Cellular components from thermophiles, including their enzymes (named thermozymes), are adapted to maintain their functions at high temperatures, and in fact they often perform optimally in such conditions. In addition to their activity at high temperatures, a number of other extremophilic characteristics are commonly found in thermozymes, such as activity in organic solvents or extreme pH tolerance. Nevertheless, the mechanisms underlying enzymatic thermostability are hard to generalize and usually are described in a case to case basis (5).

The characteristics of thermozymes make them interesting from a biotechnological standpoint, as all sort of human activities where enzymes are involved require operation at high temperature. This feature is often translated in a reduction of processes costs. This is mainly due to the reduced power consumption that is normally required to lower or increase the temperature of a liquid medium in order to advance from one process to the next. Other advantages

related to operations at high temperatures include the reduction of contamination risks, better solubility and diffusion of substrates and products (5) and better mass transfer (6), and less viscosity of liquid media (that in turn also helps in reducing the cost of operation) (7).

Another advantage of employing thermozymes is the possibility to reduce the use of environmentally harsher chemical compounds, as in some industrial processes they can be completely replaced for this environmentally friendly alternative (8). Furthermore, in regard to shifting towards a "greener" industry, examples like the valorization of biomass for the production of biofuels or other desirable products are achievable using thermozymes. Besides the biofuel production, thermozymes are applicable in all sorts of other industries like the pulp and paper industry (for processes like bleaching and deinking), in food industries (for example, in the clarification of juices), and the detergent industry, among others (9).

On metagenomics

A metagenome comprises the complete DNA sequence (genome) of all microorganisms present in an environment and can be studied in order to circumvent the limitations of pure culture when the species to study are difficult to grow in laboratory conditions. This is especially useful in extreme habitats, as many extremophilic species are deemed unculturable. Metagenomics, the study of the metagenome, comprises the molecular biology techniques that were first developed in order to have access to all the biotechnological potential of the uncultured microorganisms' metabolism (10). These bioprospecting techniques have allowed scientists to overcome the Great Plate Count anomaly (the discrepancy between the number of viable plate counts and the number of microorganisms observed directly under the microscope) (11), by providing means to study unculturable or yet to be cultured microorganisms. These studies can be oriented to the description of a microbial community and their metabolism and interactions, and towards the discovery of novel

biological compounds derived from them that can hold biotechnological value.

Regarding the methodology employed, there are two possibilities to assess the metagenome: functional metagenomics involves cloning the DNA into a suitable host to stablish a metagenomic library and then perform activity screenings to find desirable metabolites or enzymes, whereas sequence-based metagenomics relies on whole-genome sequencing to obtain the DNA sequence of the metagenome and then study its contents.

In their inception, metagenomic studies adopted the function-driven analysis strategy, resulting in metagenomic libraries with the potential to find novel gene functions and protein families different from those previously found by other more direct approaches. This functional metagenomic strategy was hindered by factors related to the host system compatibility towards the foreign genes and gene products and the methodological limitations for the analysis of metagenomic libraries, which usually comprises a very large number of clones. Problems like over-representation of abundant species within the habitat and redundancies are also to be expected when dealing with metagenomic libraries (12–15). New research protocols and refinements to existing ones have been developed in order to assess such problems, including the development of new vector systems that allow for bigger insert sizes, replication and expression in several host systems and regulation of the gene expression, and high-throughput screening methods (4,16,17).

The advent of the high-throughput DNA sequencing era allowed researchers to develop a new sequence-driven approach to metagenomic studies, by allowing the search for genes of interest directly in the metagenomes' sequences. This method is based on the comparison to previously known gene products, and thus is limited by the information available on databases that can be scant or even imprecise, but in turn overcomes the need for a suitable host system and other bottlenecks of the heterologous gene expression method

(18). If enough sequence coverage is achieved, problems like species representation and sequence redundancy may also be overcome. Nevertheless, a possible hindrance that is important to consider is the high computational power required to analyse the big volume of data generated by these studies. The development of new generations of sequencing technologies have improved the quality of these analyses and greatly reduced their cost, and bioinformaticians continuously develop new tools that demand fewer resources and shorten the required time to perform them, in order to make the sequence-based approach more universal (19–21).

Even though both techniques share the same starting material (the metagenomic DNA), they can target gene products that could potentially go undetected using one or the other. For this reason, a combined analysis using the two approaches can be a desirable methodology in order to maximize the discovery of novel products of interest, complementing each other shortcomings (22).

The present doctoral thesis involves the search for novel enzymes from two different metagenomes from hot springs located in the city of Ourense (Galicia, Spain) and is based in both a functional metagenomics study and a sequence-driven metagenomics analysis, in order to maximize our gene product discovery reach. In fact, this combined approach has allowed us to discover novel enzymes that have potential interest in industrial processes. The main objectives were the identification and the biochemical characterization of thermozymes encoded in the metagenomes, and to study the microbial community composition and its metabolic network in one of the hot springs.

References

1. Wagner ID, Wiegel J. Diversity of Thermophilic Anaerobes. Ann N Y Acad Sci [Internet]. 2008 Mar 26;1125(1):1–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18378585

2. Kaushal G, Kumar J, Sangwan RS, Singh SP. Metagenomic analysis of geothermal water reservoir sites exploring carbohydrate-related thermozymes. Int J Biol Macromol [Internet]. 2018 Nov;119:882–95. Available from: https://doi.org/10.1016/j.ijbiomac.2018.07.196

3. Dalmaso G, Ferreira D, Vermelho A. Marine Extremophiles: A Source of Hydrolases for Biotechnological Applications. Mar Drugs [Internet]. 2015 Apr 3;13(4):1925–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25854643

4. DeCastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol [Internet]. 2016 Sep 27;7(SEP):1–21. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01521/abstract

5. Adams MW., Kelly RM. Finding and using hyperthermophilic enzymes. Trends Biotechnol [Internet]. 1998 Dec;16(8):329–32. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779998011937

6. Bruins ME, Janssen AEM, Boom RM. Thermozymes and Their Applications. Appl Biochem Biotechnol. 2001;90:155–86.

7. Bhalla A, Bansal N, Kumar S, Bischoff KM, Sani RK. Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes. Bioresour Technol [Internet]. 2013 Jan;128:751–9. Available from: http://dx.doi.org/10.1016/j.biortech.2012.10.145

8. Krüger A, Schäfers C, Schröder C, Antranikian G. Towards a sustainable biobased industry – Highlighting the impact of extremophiles. N Biotechnol [Internet]. 2018 Jan;40:144–53. Available from: http://dx.doi.org/10.1016/j.nbt.2017.05.002

9. Elleuche S, Schäfers C, Blank S, Schröder C, Antranikian G. Exploration of extremophiles for high temperature biotechnological processes. Curr Opin Microbiol [Internet].

2015 Jun;25:113–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26066287

10. Mardanov A V., Kadnikov V V., Ravin N V. Metagenomics: A Paradigm Shift in Microbiology. In: Metagenomics [Internet]. Elsevier; 2018. p. 1–13. Available from: https://doi.org/10.1016/B978-0-08-102268-9.00001-X

11. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol [Internet]. 1985;39:321–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3904603

12. Ekkers DM, Cretoiu MS, Kielak AM, van Elsas JD. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl Microbiol Biotechnol [Internet]. 2012 Feb 22;93(3):1005–20. Available from: http://link.springer.com/10.1007/s00253-011-3804-3

13. Ferrer M, Martínez-Martínez M, Bargiela R, Streit WR, Golyshina O V, Golyshin PN. Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microb Biotechnol [Internet]. 2016 Jan;9(1):22–34. Available from: http://doi.wiley.com/10.1111/1751-7915.12309

14. Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC. Current and future resources for functional metagenomics. Front Microbiol [Internet]. 2015 Oct 29;6(October):1196. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.01196/abstract

15. Felczykowska A, Krajewska A, Zielińska S, Łoś JM, Bloch SK, Nejman-Faleńczyk B. The most widespread problems in the function-based microbial metagenomics. Acta Biochim Pol [Internet]. 2015 Feb 13;62(1):161–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25680374

16. Taupp M, Mewis K, Hallam SJ. The art and design of functional

metagenomic screens. Curr Opin Biotechnol [Internet]. 2011 Jun;22(3):465–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166911000371

17. Escuder-Rodríguez J-J, DeCastro M-E, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Advances of Functional Metagenomics in Harnessing Thermozymes. Metagenomics: Perspectives, Methods, and Applications. 2017.

18. Chistoserdova L. Functional Metagenomics: Recent Advances and Future Challenges. Biotechnol Genet Eng Rev [Internet]. 2009 Jan;26(1):335–52. Available from: http://www.tandfonline.com/doi/abs/10.5661/bger-26-335

19. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front Genet [Internet]. 2017 Mar 6;8(MAR):1–10. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2017.00023/full

20. Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, et al. EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D595-603. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1195

21. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics [Internet]. 2008 Dec 19;9(1):386. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-386

22. Wohlgemuth R, Littlechild J, Monti D, Schnorr K, van Rossum T, Siebers B, et al. Discovering novel hydrolases from hot environments. Biotechnol Adv [Internet]. 2018

Dec;36(8):2077–100. Available from: https://doi.org/10.1016/j.biotechadv.2018.09.004

OBJECTIVES

1. To construct metagenomic libraries from environmental DNA extracted from samples obtained in two hot springs located in the Ourense region: As Burgas and Muiño da Veiga.

2. To perform functional screens on the metagenomic libraries in order to identify relevant enzymatic activities using specific substrates.

3. To sequence, clone, express and purify a lipolytic activity enzyme to make predictions based on sequence data and to perform a biochemical characterization.

4. To sequence the metagenomic DNA from the Muiño da Veiga sample to perform a bioinformatic analysis to describe the microbial community and its metabolic interaction networks.

5. To analyze the sequenced metagenome to identify thermozymes that can be cloned and expressed in the laboratory.

6. To clone, express and purify a sequence-based analysis predicted β-xylosidase and endoglucanase to perform their biochemical characterization and confirm their biotechnological potential.

TABLE OF CONTENTS

FIGURES AND TABLES INDEX

Chapter 1: Metagenomic libraries construction from two hot springs

Chapter 2: Bioprospecting for thermozymes of biotechnological interest and biochemical characterization of a novel lipolytic thermozyme belonging to the SGNH/GDSL family of hydrolases

Chapter 3: Microbial community inhabiting the Muiño da Veiga hot spring described by sequence-based metagenomics and in-silico analysis

Chapter 4: Sequence analysis, cloning, purification and characterization of a novel thermostable endoglucanase and sequence analysis and cloning of a predicted β-xylosidase found by sequence-based metagenomics

ABBREVIATIONS

°C: Celsius degrees

µg: micrograms

µL: microlitres

µm: micrometres

A_{260}/A_{280}: absorbance at 260 nanometres wavelength divided by absorbance at 280 nanometres wavelength

AZCL-HE-Cellulose: Azurine Cross-Linked Hydroxyethyl Cellulose

AZCL-Xylan: Azurine Cross-Linked Xylan

BAC: bacterial artificial chromosome

BLAST: Basic Local Alignment Search Tool

BLASTn: Nucleotide BLAST

BLASTp: Protein BLAST

bp: base pairs

BSA: Bovine Serum Albumin

CAZy: Carbohydrate Active Enzymes

CHAPS: cholamidopropyl dimethylammonium propanesulfonate

cm: centimetres

CMC: carboxymethylcellulose

CTAB: cetyltrimethylammonium bromide

Da: daltons

DMSO: Dimethiy sulfoxide

DNA: Desoxyribonucleic acid

DNS: Dinitrosalicylic acid

DTT: Dithiothreitol

EC number: Enzyme Commision number

EDTA: Ethylenediaminetetraacetic acid

EMBL-EBI: European Molecular Biology Laboratory - European Bioinformatics Institute

GC: Guanine-Cytosine

GH: Glycosyl hydrolase

IPTG: Isopropyl β-D-1-thiogalactopyranoside

kb: kilobase (1000 base pairs)

kDa: kilodaltons

KEGG: Kyoto Encyclopedia of Genes and Genomes

K_M: Michaelis constant

KO: KEGG Orthology

L: litres

LB: Lysogenic broth

M: molar

mA: milliamperes

mg: milligrams

Miniprep: minipreparation of DNA

mL: millilitres

mM: millimolar

MPC: MasterPure™ Complete

mRNA: messenger ribonucleic acid

MUTMAC: 4-methylumbelliferyl-trimethylammonium cinnamate chloride

MUX: 4-methylumbelliferyl-β-D-xylopyranoside

N: North

NCBI: National Center for Biotechnology Information (USA)

ng: nanograms

NGS: Next Generation Sequencing

nm: nanometres

NR/NT: non-redundant nucleotide

NR: non-redundant

OD_{600}: optical density at a 600 nanometres wavelength

ORF: Open Reading Frame

OTU: Operational Taxonomic Unit

PCR: Polymerase Chain Reaction

pNP: paranitrophenol

pNPX: para-nitrophenil xylopyranoside

psi: pounds per square inch

QC: Quality Control

R^2: coefficient of determination

RNA: ribonucleic acid

RNAse: ribonuclease

rpm: revolutions per minute

rRNA: Ribosomal ribonucleic acid

SDS: sodium dodecyl sulphate

SDS-PAGE: sodium dodecyl sulphate - polyacrylamide gel electrophoresis

SGNH/GDSL: Serine, Glycine, Asparagine, Histidine / Glycine, Aspartic acid, Serine, Leucine

SMTL: Swiss Model Template Library

TBS: Tris Buffered Saline

TCA: Tricarboxilic Acid Cycle

TE: Tris-EDTA

tRNA: transfer ribonucleic acid

U: Unit of enzymatic activity

UK: United Kingdom

USA: United States of America

UV: ultraviolet

v/v: volume per volume

V: volts

$\mathbf{V_{max}}$: maximum reaction rate

w/v: weight per volume

W: West

CHAPTER 1

Metagenomic libraries construction from two hot springs

Abstract

Construction of metagenomic libraries is a fundamental first step in order to assess the vast amount of DNA found in a microbial community. In this chapter, we describe the high-molecular weight metagenomic DNA extraction and construction of two metagenomic libraries from two different hot spring sources from the Galicia region (northwest of Spain). The protocols from sample collection to the transformation of the host microorganism and its long-term preservation in a 96-well plate format allowing fast and high-throughput functional screenings are described in detail. We also developed a protocol to induce the fosmid vector used to construct the libraries from low copy number to high copy number, which proves to be useful for DNA purification steps and for the expression of heterologous genes.

Introduction

Ourense hot springs

The city of Ourense, located in the Galicia region in the northwest of Spain, has been since Roman times known for its hot springs, and in the last two decades these geothermally-heated waters had experienced a rekindled interest as an economic factor that draws thermalism tourism to the region (1). Unlike many other well-studied hydrothermal systems as Yellowstone National Park in USA (2), and some regions in Iceland (3) and Japan (4), to mention a few; the hot springs in Ourense have a non-volcanic origin. The geological setting of the region has been described previously (1,5). In brief, the city lies at the bottom of a hollow formed by the erosion process of Miño river and its main tributaries Barbaña and Lonia. From a lithological point of view, two groups of granite crystalline rocks are contacted by intrusion along both sides of the

Miño river, delimiting a north and a south side, and are all covered by alluvial soils. The granite rock presents almost no permeability, which makes the water flow only along faults and fractures. This in turn makes the water follow a deep circuit of percolation at higher elevations, storage and circulation with heat transfer through main fault systems (6), and emergence at lower elevations (1,5). The process involves the interaction of water with minerals, adding solutes to its composition. As the water reaches greater depths, rock temperature increases and heat transfer takes place, explaining the higher occurrence of hot springs in the lower elevations of the valley, where water has circulated deeper and for longer amounts of time. Aside from the high basal heat flow of the region, radioactive decay reactions in igneous rocks are also a possible additional source for heat transfer to the water (6). In this study we focused our interest on two hot springs in the city of Ourense area, namely the Muiño da Veiga hot spring and the As Burgas hot spring. The Muiño da Veiga hot spring site is located on the north side of the Miño river, whereas the As Burgas hot springs are located on the south side.

Metagenomic DNA extraction and construction of metagenomic libraries

Construction of metagenomic libraries requires the extraction of the metagenomic DNA, and depending on the size of the cloned DNA there is a distinction between small-insert libraries, plasmid libraries with insert size lower than 10kb; and large-insert libraries, cosmid or fosmid libraries up to 40kb or Bacterial Artificial Chromosome (BAC) libraries beyond that size (7). Small-insert libraries have the potential to contain single genes or small operons that can be identified afterwards with a screening method but lack the capability of large-insert libraries to contain more complex pathways and to represent significant contiguous genomic regions. In fact, large-insert libraries have been proven useful not only in the bioprospecting of metagenomes, but also in the study of rare

species of a microbial community, the study of unknown metabolic pathways and the closure of genomes from more abundant species in reconstruction of microbial genomes (8). Of these, the fosmid vector large-insert libraries have gained popularity due to their capabilities derived from cosmids and BACs. From the BAC vectors they take the F-(fertility) factor that ensures low copy number per cell, in order to minimize intermolecular recombination events, and from the cosmid vectors they take the cos sites that allow packaging inside empty bacteriophage lambda heads, significantly improving the transformation efficiency and allowing a more even distribution of insert sizes (8). The DNA extraction step is critical, as high-quality (as in pure enough) and in high yields metagenomic DNA must be recovered to ensure the construction of large enough libraries covering the microbial diversity of the ecosystem, and in the case of large-insert libraries it also needs to be relatively unsheared. This four parameters (purity, yield, size and representativeness) are unfortunately often in competition as the extraction method will impact them in different ways, favouring ones over others (9). For purity, the protocol design must take into consideration the nature of the sample selected (10,11). In the case of terrestrial hot springs, it is known that microbial cells tend to attach to particulate materials and that clay minerals must be removed from the sample as they can inhibit downstream processes of the library generation protocols (8). DNA yield from hot springs is also a bottleneck for library construction, as the biomass in this habitat is lower than other more populated ecosystems (12). Moreover, it is known that metagenomic libraries introduce some degree of bias to species representation of the whole community, although this is deemed of little consequence for the purposes of bioprospecting (13), as the apparent functional diversity seems not to be compromised by the DNA extraction method (9).

After the extraction, the DNA is cloned in a suitable host, that in most cases is the model organism *Escherichia coli* (7), which is estimated to be able to correctly express around 40% of genes with

a sequence deposited in public databases (12). As a potential contingency for functional screenings based on metagenomic libraries using exclusively this host, the need for broader host-range shuttle vectors is highlighted (12). Also, and as mentioned, a low copy number vector has the advantage of improved genomic stability, but in some instances, it is advantageous to have a high number of copies of the vector per cell. For this reason, inducible copy number vectors are advisable (8).

In this chapter, we describe the construction of two metagenomic libraries from the hot springs mentioned above, from the DNA extraction protocol to the transformation of the bacterial host, and its long-term storage for later use in functional screenings.

Materials and Methods

Metagenomic DNA sources: site description and sample collection

Water was collected from two different hot spring sources in the northwest of Spain, specifically in the city of Ourense, which is located in the Ourense province in the Galicia region. A detailed view of the sampling sites is provided in Figure 1, with images generated using the Google Maps service (14).

The hot spring As Burgas (42°20'04.6"N 7°51'55.1"W) is located inside the city of Ourense. The site consists in two artificial fountains with a continuous water flow and a hot water pool located above the fountains where public bathing is allowed. Water samples were collected from the continuous flow from one of the fountains (the one to the left when facing them) in January 2014.

The hot spring Muiño da Veiga (42°21'05.3"N 7°54'36.3"W) is located next to the Miño river in the outskirts of Ourense. The water springs from a manually operated pump, and the site also consists in four hot-water pools open to the public. The water was pumped several times before collection, and the date was December 2015.

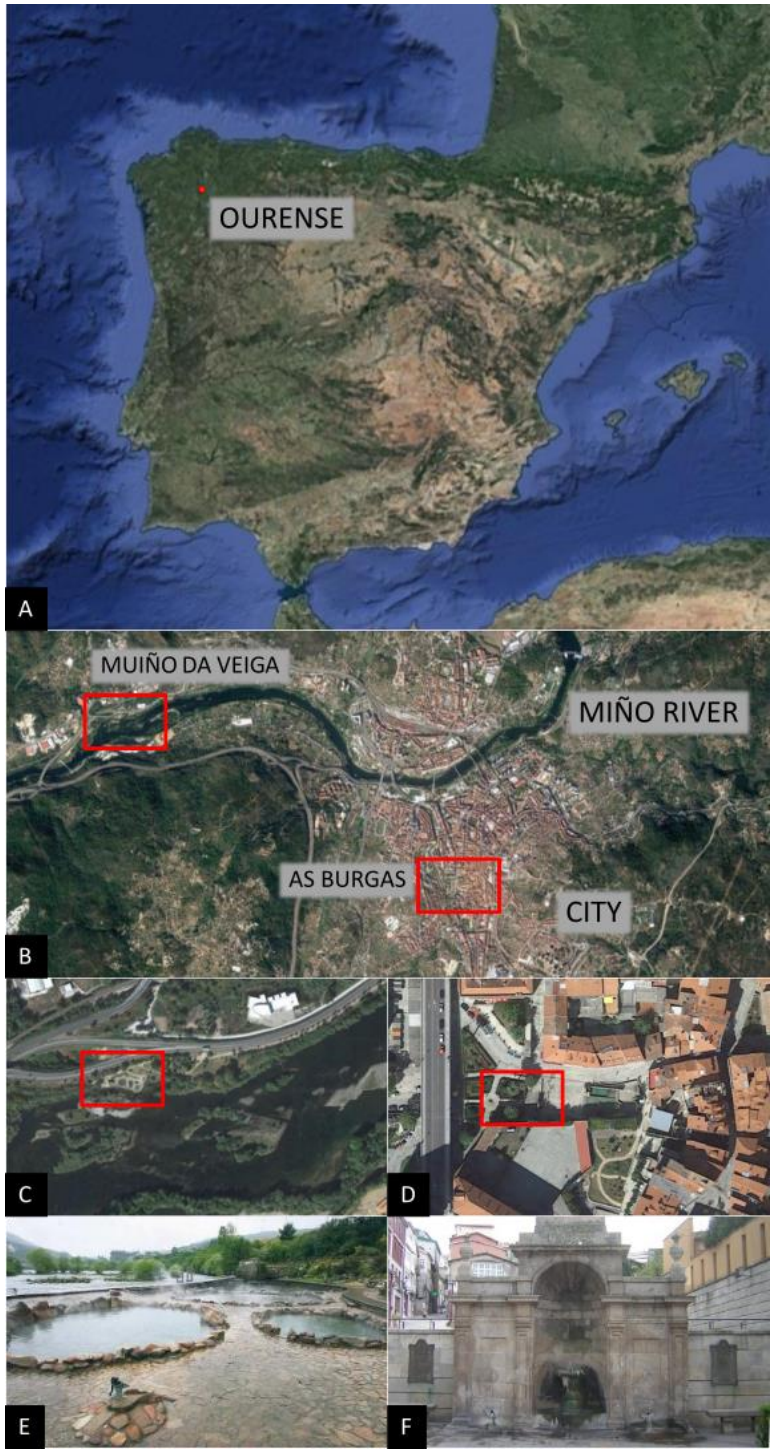

OURENSE
A
MUIÑO DA VEIGA
MIÑO RIVER
AS BURGAS
CITY
B
C
D
E
F

Figure 1. Sampling sites locations. (A) Location of the city of Ourense in the Iberian Peninsula, marked as a red dot. (B) Close-up view of the city of Ourense and the Miño river north of it. Inside red squares are the sampling sites, Muiño da Veiga to the left and As Burgas to the right of the image. (C) Close-up view of the red square from the previous picture corresponding to the Muiño da Veiga hot spring. Another red square marks the precise location of the hot spring. (D) Close-up view of the red square from Figure 1B that corresponds to the As Burgas hot spring. Another red square marks the exact location of the As Burgas hot spring fountains. (E) Muiño da Veiga hot spring, on the center of the image the four main pools can be viewed, in the lower part of the image is the fountain where water was collected, in the upper left corner the Miño river can be seen. (F) As Burgas hot spring, in the lower part of the image two fountains are present. Water was collected from the left fountain.

Before sample collection we measured temperature and pH on each site using a thermometer and pH strips. A total of 125L of water were collected from each source in ethanol-washed plastic bottles to minimize microbial contamination and were washed several times on-site with the hot spring water before final sample collection. The samples were transported to the laboratory on the date of collection for processing in the shortest amount of time possible. In both cases processing of the samples including the DNA extraction step was done within the same week as the collection of the samples, starting the next day after collection.

DNA extraction

The protocol outlined in the commercial kit “Metagenomic DNA isolation kit for water” (Epicentre, USA) was adopted with slight modifications in order to extract the DNA from the microorganisms present in the sample (the metagenomic DNA). This method is based in a combination of chemical and enzymatic treatments for the extraction, allowing the recovery of high molecular weight DNA randomly sheared in fragments of approximately 40kb, which is the optimal size for ligation to the fosmid vector and viral-assisted transformation.

Water from the As Burgas source was pre-filtered with a 5µm membrane filter in order to remove big particulate material that would block smaller filters. We found this step not to be necessary for the water from the Muiño da Veiga source as big particulate matter was not present. Water was then filtered with a 0.2µm membrane filter (vacuum was achieved with a vacuum pump, always working below 15psi to avoid damaging the membrane filter) in order to trap the microorganisms present in the sample. For every 25L of water filtered (the volume of one plastic bottle) from 2 to 4 filter membranes were obtained, and two of these filters were used in each extraction procedure. Filters were manipulated and cut into four pieces with ethanol-soaked instruments (scissors and clamps) and placed face-up at the bottom of a conic-shaped 50mL tube. The filters were washed with 1mL of Filter Wash Buffer from the kit supplemented with Tween 20 (final concentration 0.2% v/v) by slow vortexing to soak the membranes and a 2 minutes long high-speed step with brief stops in order to detach the microorganisms from the filter membranes.

The cell suspension was transferred to a new 1.5mL tube where it was pelleted at 14,000rpm for 2 minutes. This pellet was resuspended in 300µL TE buffer (10mM Tris-HCl pH 7.5, 1mM EDTA) and treated with 2µL Ready-Lyse Lysozyme and 1µL RNAse A included in the kit following the manufacturer instructions (a brief vortex followed by a 30 minutes incubation at 37°C). After this treatment, 300µL of Meta-Lysis Solution and 1µL of Proteinase K from the kit were added, mixed by vortexing and followed by a brief centrifugation pulse, and incubated at 65°C for 15 minutes. The mixture was then cooled on ice for at least 5 minutes. 350µL of MPC Protein Precipitation Reagent were added and mixed by vortexing, and the precipitate was pelleted by centrifugation at 14,000rpm for 10 minutes in a refrigerated centrifuge (4°C). The supernatant was then transferred to a new clean tube, were 570µL of isopropanol were added and mixed by inverting the tube several times. DNA was then precipitated by centrifugation at 14,000rpm for 10 minutes in a

refrigerated centrifuge (4°C). The supernatant was pipetted out and a brief centrifugation pulse was performed in order to remove any remaining liquid on the walls. The pellet was washed with 500µL of 70% ethanol, and the previous steps of high-speed cold centrifugation, supernatant removal and wall-attached liquid removal were repeated. The pellet was air-dried for at least 8 minutes at room temperature. Finally, the DNA was resuspended in 30µL Tris Buffer 10mM, and the extractions from the same source were pooled together.

Library construction

A modified procedure for the commercial kit "CopyControl Fosmid Library Production Kit" (Epicentre, USA) was adopted for the construction of both metagenomic libraries.

End-repair and size-selection of metagenomic DNA

Ends from the extracted DNA were repaired (blunt-ended and 5' phosphorylated) following the protocol described by the manufacturer. For size selection, DNA was loaded on a 1% 20cm agarose gel and electrophoresis was carried out on ice at 40V overnight with the GeneRuler High Range DNA Ladder (Thermo Fisher Scientific, USA) as molecular weight marker and the Fosmid Control DNA as reference. External lanes from the gel (including the molecular weight marker, the Fosmid Control DNA and a small portion of the metagenomic DNA) were cut and stained with a 3x GelGreen (Biotium, USA) solution for 30 minutes in order to visualize the DNA under UV light without exposing the samples to it. The DNA that co-migrated with the 40kb band was then cut from the gel and the "GeneJET Gel Extraction Kit" (Thermo Fisher Scientific, USA) was used to recover the DNA. The agarose plugs were weighed and a 1:1 proportion (volume:weight) of Binding Buffer was added. Agarose was melted at 60°C for 10 minutes, and the solution was loaded to a GeneJET purification column. After centrifugation for 1 minute at 14,000rpm the flow-through was

discarded and 700µL of Wash buffer (containing ethanol) was added. The tube was centrifuged twice (the second time without load for removal of residual ethanol) for 1 minute at 14,000rpm, and both flow-throughs were discarded. The DNA was eluted in 20µL of pre-heated water (65°C) by placing the column in a new tube and centrifuging for 1 minute at 14,000rpm.

Vector cloning

The fosmid vector pCC1FOS from the kit was replaced with the pCT3FK fosmid (15), a shuttle vector for *E. coli* and *Thermus thermophilus* derived from pCC1FOS, kindly provided by author Dr. Angelov. A vector map is provided as Figure 2 generated with the Serial Cloner software (16). These vectors contain a single copy replication origin but can be induced to high copy number by arabinose induction mediated by the *trfA* gene product provided by the host strain and feature a chloramphenicol resistance gene for *E. coli*. The modified vector also adds a thermostable kanamycin resistance gene cassette and sequences from the *T. thermophilus* genome for positive selection and homologous recombination in that host. In order to obtain enough vector to clone the metagenomic DNA it was necessary to clone it first in a bacterial host, induce it to high-copy numbers and extract it following a minipreparation protocol.

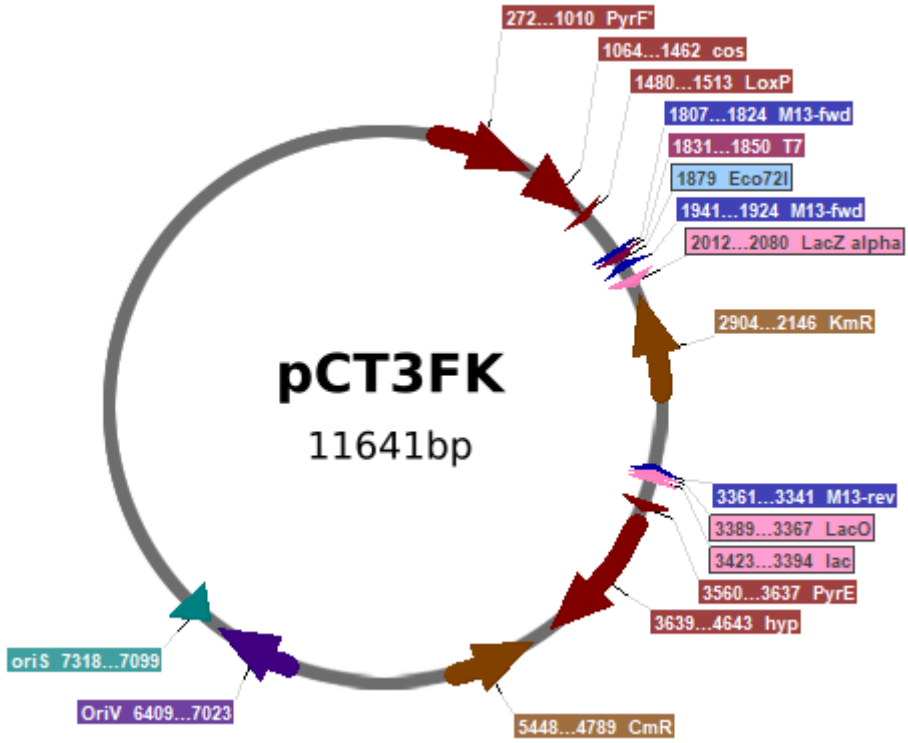


Figure 2. pCT3FK fosmid vector map, highlighting some of its features, including PyrF' and PyrE for *T. thermophilus* recombination, Km^R for kanamycin resistance in that host, restriction site *Eco*72I, Cm^R for chloramphenicol resistance in *E. coli*, origins of replication OriS and OriV for the bacterial F plasmid and cos site for lambda phage recognition.

The *E. coli* strain EPI300 ™ $T1^R$ (F^- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80d*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara*, *leu*)7697 *gal*U galK λ^- *rps*L *nup*G *trf*A *ton*A *dhfr*) was induced to competence following the classic calcium chloride chemical method (17). A single colony grown overnight at 37°C on LB medium without glucose (18) (1% Bacto™ tryptone (BD, USA), 0.5% Bacto™ yeast extract (BD, USA), 0.5% NaCl, 1.5% Bacto™ Agar (BD, USA)) was used to inoculate 100mL of LB medium in a 1L flask, and growth was monitored at 37°C for 3 hours by reading the OD_{600} of aliquots taken

at regular intervals. When OD_{600} reached 0.4 the culture was transferred to 50mL ice-cold conic tubes and was kept on ice for at least 10 minutes. Afterwards, the cells were pelleted by centrifugation at 2,700rpm for 10 minutes in cold (4°C) and resuspended in 30mL of an ice-cold $MgCl_2$-$CaCl_2$ solution (80mM $MgCl_2$, 20mM $CaCl_2$) by vortexing. Cells were again pelleted under the same conditions and resuspended in 2mL of ice-cold 0.1M $CaCl_2$. For long-term storage, cells were kept as a glycerol stock mixing them with LB supplemented with glycerol to a final concentration of 10% (v/v).

The competent cells were then transformed by the thermal shock method (17). 200µL of cell suspension were mixed with 50ng of pCT3FK vector and kept on ice for 30 minutes. The mix was incubated at 42°C for 90 seconds and transferred immediately to ice where it was kept for at least 2 minutes. Cells were recovered and allowed to express the antibiotic resistance gene by incubation in 800µL LB medium for 45 minutes at 37°C with agitation. Excess medium was discarded by centrifugation and around 200µL of the cell suspension were grown on a Petri dish containing LB medium supplemented with chloramphenicol (12.5µg/mL) at 37°C overnight.

A single colony from this Petri dish (harbouring the fosmid pCT3FK) was employed to inoculate 100µL of LB medium supplemented with chloramphenicol and 0.02% arabinose to induce a high copy number. The empty fosmid was extracted following the "GeneJET Plasmid Miniprep" kit (Thermo Fisher Scientific, USA). The culture was transferred to 50mL conical tubes and centrifuged at 5,000rpm for 5 minutes. The pelleted cells were resuspended in the Resuspension solution (containing RNAse A) and transferred to a new 1.5mL tube. 250µL of Lysis solution were added to this suspension and mixed by inversion until it became clear. Then, 350µL of Neutralization solution were added and mixed by inversion, forming a white precipitate. This pellet was removed by centrifugation at 14,000rpm for 5 minutes, and the supernatant was

transferred to the GeneJet Spin Column. The column was centrifuged at 14,000rpm for 1 minute and the flow-through was discarded. The column was washed twice with 500µL of Wash solution (containing ethanol) and dried by centrifugation of the non-loaded column for 2 minutes. The vector was eluted in 50µL of water by placing the column in a new tube and centrifuging for 2 minutes at 14,000rpm, and the concentration was found to be 591.7ng/µL with a 1.85 A_{260}/A_{280} ratio.

Ligation and packaging of the metagenomic DNA

The ligation of the metagenomic DNA with the empty pCT3FK vector was carried out following the manufacturer's indications, but volumes were adjusted to allow for the maximum amount of metagenomic DNA due to its low concentration compared to the one suggested in the protocol. Incubation was performed at 16°C overnight, and the Fast-Link DNA Ligase was heat-inactivated at 70°C for 10 minutes. This reaction mixture was used directly to package in Lambda phage extracts. 10µL of the ligation reaction were mixed with 25µL of MaxPlax Packaging Extract (thawed on ice) and incubated at 30°C for 2 hours. Another 25µL of the phage extract was added and the mix was incubated for another 2 hours at 30°C. The solution was dissolved with 500µL of Phage Dilution Buffer and 25µL of chloroform. A tittering step was performed testing four serial dilutions (original dilution, 1:10, 1:100 and 1:1000) and it was found that the original dilution produced the desired number of colonies per plate.

0.5mL of an overnight culture of *E. coli* EPI300 $T1^R$ grown on 50mL of LB supplemented with 10mM $MgSO_4$ and 0.2% Maltose (to induce the expression of the membrane maltose receptor used by the phage to infect the cell) at 37°C was used to inoculate 50mL of the same medium. This fresh culture was grown to an OD_{600} of 0.8 and stored at 4°C until the infection reaction.

For the infection, 10µL of diluted packaged phage particles were mixed with 100µL of the culture described above and incubated for 1 hour at 37°C (with no agitation) to allow for the phages to infect the cells. The cells were then spread on LB plates supplemented with chloramphenicol (12.5µg/mL) and grown overnight at 37°C.

Individual colonies were picked from these plates and transferred to 96-well plates containing 100µL of LB supplemented with chloramphenicol and glycerol to a 10% (v/v) final concentration in order to store them at -80°C (long term storage plates). The cultures were grown overnight at 37°C with agitation, and duplicates from these plates (working plates) were obtained transferring 5µL of each well to new plates with the same conditions for growth and storage as the original ones.

For clarity, a schematic overview of the metagenomic library construction process is provided in Figure 3.

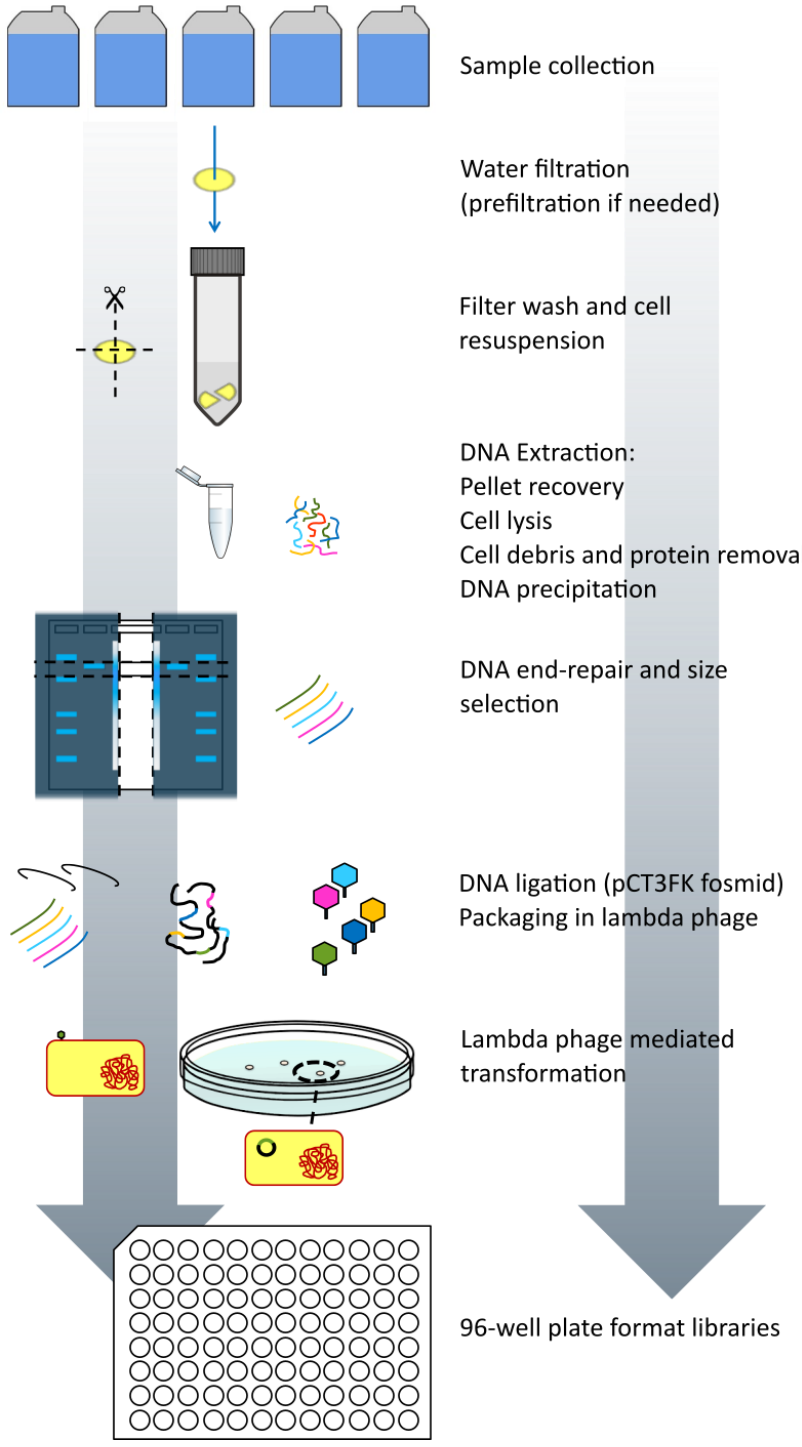

Sample collection
Water filtration
(prefiltration if needed)
Filter wash and cell
resuspension
DNA Extraction:
Pellet recovery
Cell lysis
Cell debris and protein removal
DNA precipitation
DNA end-repair and size
selection
DNA ligation (pCT3FK fosmid)
Packaging in lambda phage
Lambda phage mediated
transformation
96-well plate format libraries

Figure 3. Overview of the metagenomic library production protocols outlined in this chapter, from sample collection to the final 96-well plate format.

Arabinose induction test

In order to assess the possibility of employing arabinose to replace the autoinduction solution provided in the commercial kits, we performed a test to check our ability to recover DNA from similarly grown cultures (same OD_{600} measurements) with a no induction control group, an autoinduction solution (from the Epicentre CopyControl Fosmid Library Production kit) induced group and an 0.02% arabinose induced group. All groups comprised triplicates of the same *E. coli* EPI300 $T1^R$ clone harbouring an empty pCT3FK vector. DNA was recovered following the same protocol described in the "vector cloning" step using the GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific, USA).

Results and discussion

Water sampling

We measured the temperature and pH at which the water springs from the As Burgas fountains, finding them to be 67°C and 7.5 respectively. For Muiño da Veiga the temperature and pH were also read on site and were found to be 68°C and 7.0 respectively. Previous reports from the same sites reported very similar yet slightly lower temperatures (66.3°C for As Burgas and 66°C for Muiño da Veiga) (5,6), similar or identical pH values for As Burgas (7.56 (5) or 7.5 (6)) and varying pH values for Muiño (7.6 (6) or 8.1 (5) compared to our reading of 7.0). Thus, both hot springs can be described as thermophilic and circumneutral to slightly alkaline. Composition analysis previously conducted by other researchers have concluded that both waters are sodium-bicarbonated due to the higher content of this compounds relative to others (5).

Extraction of high molecular weight DNA

For the As Burgas source, a final concentration of 1,733.3ng/µL was obtained, with a A_{260}/A_{280} ratio of 1.83 in a volume of 20µL (34,666ng of total DNA). The metagenomic DNA for Muiño da Veiga was extracted at a final concentration of 1,371.8ng/µL with a A_{260}/A_{280} of 2.5 in a final volume of 45µL (61,731ng of total DNA). Agarose gel electrophoresis results are given for the As Burgas and Muiño da Veiga DNA extractions in Figure 4. The DNA in both cases migrated mostly as a discernible band co-migrating with the 40kb DNA controls, thus avoiding the need to further shear the metagenomic DNA for accommodation to the required insert size for the fosmid vector. An aliquot of metagenomic DNA from Muiño da Veiga hot spring was saved to be used for shotgun sequencing as later described in Chapter 3.

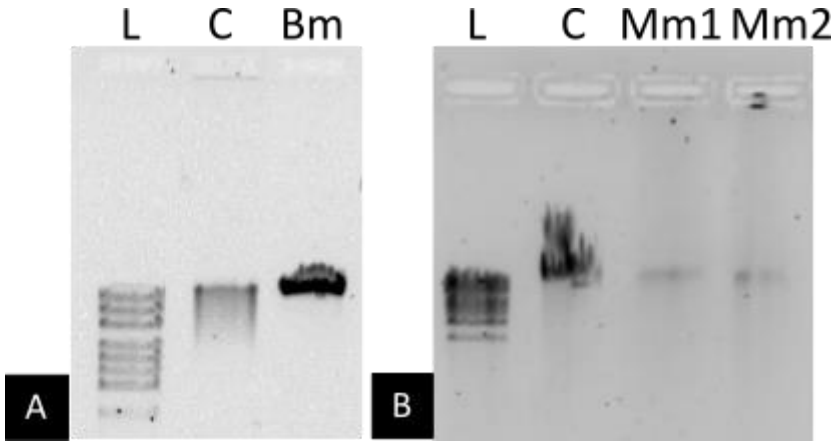


Figure 4. Agarose gel electrophoresis of metagenomic DNA extracted from the water samples from (A) As Burgas and (B) Muiño da Veiga. L: DNA Ladder; C: Control 40kb DNA; Bm: As Burgas metagenomic DNA; Mm1 and Mm2: Muiño da Veiga metagenomic DNA (two different extractions previous to pooling them together).

Library construction

The DNA concentration was measured after the gel extraction and was found to be 76.8ng/µL for the As Burgas source, in a volume of 50µL, being the total DNA recovered 3,840ng. For the Muiño da Veiga source, two different extractions were obtained but were pooled together for library construction, with 422.3ng/µL and 634.5ng/µL concentrations respectively, in a volume of 50µL in each case, the amount of DNA recovered after pooling them together being 52,840ng.

After cloning, the final library size was of 27,898 clones for the As Burgas source, with clones distributed in 16 different 96-well plates, averaging 20 clones per well. For the Muiño da Veiga source, we obtained 4,399 individual colonies, that were distributed in 5 different 96-well plates, averaging 10 clones per well. Other libraries using the similar commercial fosmid vectors pCC1FOS or pCC2FOS have been reported to yield similar number of clones (11), from as low as 5,000 with the pCC2FOS vector (similar to our Muiño da Veiga library) to as big as 96,000 with the pCC1FOS fosmid, averaging 37,575 which is slightly above our As Burgas library size. Moreover, using the same vector pCT3FK similar sized (if not lower) libraries have been reported (19) with 6,048 clones from hot spring samples collected in Portugal and 1,920 clones from a compost sampled in Germany. Our two libraries were stored for long-term conservation, and duplicated libraries were generated in order to conduct functional-based screenings as later explained in Chapter 2 of the present work.

Arabinose induction test

It was found that higher amounts of fosmid DNA were able to be retrieved from the arabinose induced group compared to the amount obtained with the autoinduction solution group. Figure 5 shows the results of DNA concentration recovered depending on the method of high-copy induction used. Following this result, we

adopted this protocol for all subsequent induction to high-copy number steps of our studies when the pCT3FK vector system was involved (specifically, before the extraction of fosmid DNA from individual clones, for the functional screenings of the metagenomic libraries and for expression of genes from selected clones to perform enzymatic tests). The original pCC1FOS vector, upon which the pCT3FK is based on, carries the origin of replication *oriV* from the RK2 plasmid requiring the *trfA* gene product, and not only induces the vector to high-copy number, but also makes it compatible with several *Proteobacteria* hosts (13). As the *E. coli* strain EPI300 $T1^R$ provides the *trfA* gene in trans under control of an inducible promoter (13), it is likely that the L-arabinose controlled *araC-PBAD* system (20) drives the induction to high copy number in our metagenomic libraries. Inducible vectors provide a means to ensure high genomic stability and also reduce the need for high yields of biomass (thus reducing research costs) in downstream experiments (8).

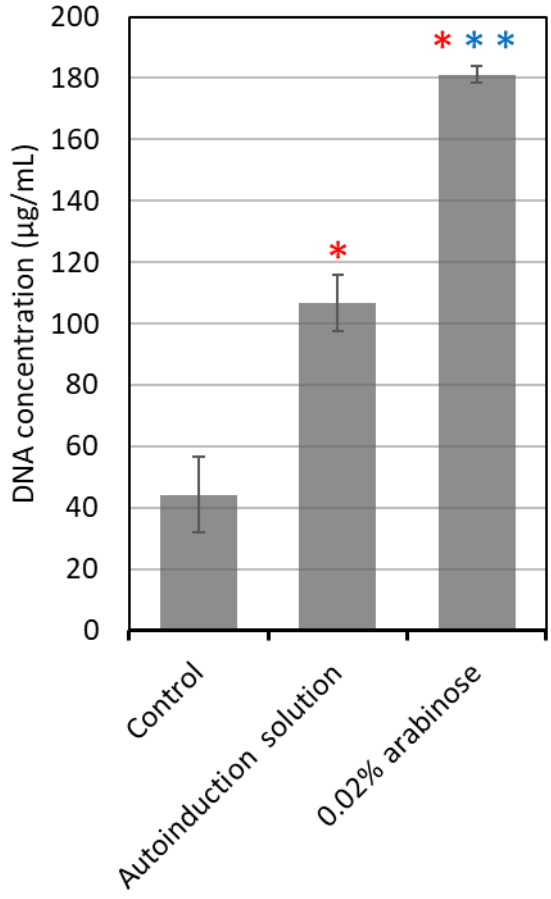


Figure 5. DNA concentration obtained after miniprep recovery of empty fosmids from same OD_{600} cultures with different methods of induction for high-copy number or no treatment for the control group. Red asterisks represent a p-value lower than 0.05 in a t-student test for two samples assuming equal variances for comparisons between the control group and each of the induction treatments. Blue asterisks represent a p-value lower than 0.005 in a t-student test for two samples assuming equal variances for a comparison between the autoinduction treatment group and the arabinose treatment group.

Conclusions

We obtained two metagenomic libraries with sizes comparable to other successful bioprospecting studies based on functional screenings for biocatalysts of biotechnological relevance. These libraries are resources of value in the search for new thermozymes, as they circumvent the need to have previous knowledge on the sequence of the gene products. Nevertheless, metagenomic libraries require robust screening methods to overcome some of the bottlenecks associated with the heterologous expression in laboratory host strains, as discussed in the next chapter.

References

1. González-Barreiro C, Cancho-Grande B, Araujo-Nespereira P, Cid-Fernández JA, Simal-Gándara J. Occurrence of soluble organic compounds in thermal waters by ion trap mass detection. Chemosphere [Internet]. 2009 Mar;75(1):34–47. Available from: http://dx.doi.org/10.1016/j.chemosphere.2008.11.067

2. Colman DR, Lindsay MR, Boyd ES. Mixing of meteoric and geothermal fluids supports hyperdiverse chemosynthetic hydrothermal communities. Nat Commun [Internet]. 2019 Dec 8;10(1):681. Available from: http://dx.doi.org/10.1038/s41467-019-08499-1

3. Cousins CR, Fogel M, Bowden R, Crawford I, Boyce A, Cockell C, et al. Biogeochemical probing of microbial communities in a basalt-hosted hot spring at Kverkfjöll volcano, Iceland. Geobiology. 2018;16(5):507–21.

4. Kubo K, Knittel K, Amann R, Fukui M, Matsuura K. Sulfur-metabolizing bacterial populations in microbial mats of the Nakabusa hot spring, Japan. Syst Appl Microbiol [Internet]. 2011 Jun;34(4):293–302. Available from: http://dx.doi.org/10.1016/j.syapm.2010.12.002

5. Delgado-Outeiriño I, Araujo-Nespereira P, Cid-Fernández JA, Mejuto JC, Martínez-Carballo E, Simal-Gándara J. Behaviour

of thermal waters through granite rocks based on residence time and inorganic pattern. J Hydrol [Internet]. 2009 Jul;373(3–4):329–36. Available from: http://dx.doi.org/10.1016/j.jhydrol.2009.04.028

6. López DL, Araujo PA, Outeiriño ID, Cid JA, Astray G. Geochemical signatures of the groundwaters from Ourense thermal springs, Galicia, Spain. Sustain Water Resour Manag [Internet]. 2019 Mar 5;5(1):103–16. Available from: http://dx.doi.org/10.1007/s40899-018-0239-3

7. Simon C, Daniel R. Construction of Small-Insert and Large-Insert Metagenomic Libraries. In: Streit WR, Daniel R, editors. Metagenomics [Internet]. New York, NY: Springer New York; 2017. p. 1–12. (Methods in Molecular Biology; vol. 1539). Available from: http://link.springer.com/10.1007/978-1-4939-6691-2

8. Reigstad LJ, Bartossek R, Schleper C. Preparation of High-Molecular Weight DNA and Metagenomic Libraries from Soils and Hot Springs. In: Methods in Enzymology [Internet]. 1st ed. Elsevier Inc.; 2011. p. 319–44. Available from: http://dx.doi.org/10.1016/B978-0-12-386489-5.00013-0

9. Ekkers DM, Cretoiu MS, Kielak AM, van Elsas JD. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl Microbiol Biotechnol [Internet]. 2012 Feb 22;93(3):1005–20. Available from: http://link.springer.com/10.1007/s00253-011-3804-3

10. Gurgui C, Piel J. Metagenomic Approaches to Identify and Isolate Bioactive Natural Products from Microbiota of Marine Sponges. In: Streit WR, Daniel R, editors. Metagenomics [Internet]. Totowa, NJ: Humana Press; 2010. p. 247–64. (Methods in Molecular Biology; vol. 668). Available from: http://link.springer.com/10.1007/978-1-60761-823-2

11. DeCastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol [Internet]. 2016 Sep 27;7(SEP):1–21. Available from:

http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01521/abstract

12. Ferrer M, Golyshina O, Beloqui A, Golyshin PN. Mining enzymes from extreme environments. Curr Opin Microbiol [Internet]. 2007 Jun;10(3):207–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369527407000513

13. Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC. Current and future resources for functional metagenomics. Front Microbiol [Internet]. 2015 Oct 29;6(October):1196. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.01196/abstract

14. Google Maps Imagery ©2021 TerraMetrics, Map data ©2021 GeoBasis-DE/BKG (©2009) CNES / Airbus, Landsat / Copernicus, Maxar Technologies, Map data ©2021 Instituto Geográfico Nacional Spain [Internet]. 2021. Available from: https://www.google.com/maps/

15. Angelov A, Mientus M, Liebl S, Liebl W. A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. Syst Appl Microbiol [Internet]. 2009 May;32(3):177–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0723202009000137

16. Serial Cloner [Internet]. [cited 2021 Jul 28]. Available from: http://serialbasics.free.fr/Serial_Cloner.html

17. Sambrook J, Russell DW. Molecular cloning: A laboratory manual. 3rd ed. NY, USA: Cold Spring Harbor Laboratory Press; 2001.

18. Bertani G. Studies on lysogenesis. J Bacteriol [Internet]. 1951 Sep;62(3):293–300. Available from: https://journals.asm.org/doi/10.1128/jb.62.3.293-300.1951

19. Leis B, Angelov A, Mientus M, Li H, Pham VTT, Lauinger B, et al. Identification of novel esterase-active enzymes from hot

environments by use of the host bacterium *Thermus thermophilus*. Front Microbiol [Internet]. 2015 Apr 8;6(April):275. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25904908

20. Newman JR, Fuqua C. Broad-host-range expression vectors that carry the l-arabinose-inducible *Escherichia coli araBAD* promoter and the *araC* regulator. Gene [Internet]. 1999 Feb;227(2):197–203. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378111998006015

CHAPTER 2

Bioprospecting for thermozymes of biotechnological interest and biochemical characterization of a novel lipolytic thermozyme belonging to the SGNH/GDSL family of hydrolases

Abstract

Functional screenings were conducted on two metagenomic libraries from hot springs in order to find novel thermozymes with potential biotechnological applications. These included enzymes acting on plant cell walls like endoglucanases and exoglucanases, β-glucosidases, xylanases and β-xylosidases, and broad application enzymes like proteases and lipolytic hydrolases. Of all the enzymes found by this bioprospection, we selected a novel lipolytic enzyme for further characterization. The protein was found to belong to the SGNH/GDSL family of hydrolases. It was purified and its biochemical parameters determined. We found that the enzyme was most active at 60°C and pH 9 using pNP-laurate as substrate and was highly thermostable. It also showed preference for short-chained substrates and activation with temperature and with certain detergents like Tween 80. Proteins of this family of hydrolases are relevant for their broad substrate specificity, that coupled with this protein high temperature optima, broad pH range and thermostability further highlights its biotechnological potential.

Introduction

The functional screening approach to assess metagenomes

The function-driven approach to study metagenomes consists in the cloning of metagenomic DNA in a suitable host to construct a library, that is later used to perform screenings in which the specific activity desired is assayed. This bioprospecting tool has been extensively reviewed (1–9) often comparing it to the sequence-based approach where the metagenomic DNA is sequenced with Next-Generation

Sequencing techniques. Due to the need for high computational power and its processing costs, as well as Whole Genome Sequencing costs and requirement for specific equipment, functional metagenomics can be regarded as a more affordable (yet still expensive and labour-demanding) method for assessing the metagenome. Moreover, the functional approach has some other advantages, such as the lack of requirement for previous knowledge on the gene product amino acid sequence, thus holding the potential to discover truly novel enzymes harbouring new protein domains or belonging to not yet described protein families. Nevertheless, both methods are proven to successfully retrieve previously uncharacterized biocatalysts (10,11) and studies where both strategies can be adopted may make the most from the same starting material as they complement the information that can be retrieved (12,13).

Many authors have identified main bottlenecks that functional screening studies usually face (1,11,14). First, and as discussed in the previous Chapter of this work, the isolation of contaminant-free environmental DNA of sufficient quality and integrity, in large-enough amounts and with minimal sampling bias is a challenging task, especially when the source material typically yields low biomass as it is the case for thermal springs and other extreme habitats (3).

Particular to the function-driven strategy, the selection of a suitable heterologous host and vector system and of the screening method are key choices to successfully find sought activities (1,14). The most obvious (but not to be overlooked) decision is to pick a host strain lacking the desired activity that would be screened (either naturally or by genetically engineering knock-out strains), as background signal can prevent discrimination between positive and negative hits (5). Some of the problems associated with the gene expression in a heterologous host include: (i) Internal cell accumulation, which refers to situations where the biomolecule is accumulated inside the library host and thus avoids detection, which may be partially solved if a mild

lysis protocol is adopted such as the use of non-ionic detergents or for thermozymes the use of high temperatures to disrupt the membrane of mesophilic hosts (1). (ii) Codon usage bias, if the heterologous host has different preferences in codon usage, it may lead to situations where the host has to overexpress genes with an unusually high number of rare codons which in turn prevents the correct biosynthesis of the gene product. These situations could be alleviated by either inserting the tRNA genes required in the expression vector, or by trying alternative hosts systems for the metagenomic libraries using broad-range shuttle vectors (1). (iii) Failure in promoter recognition, that happens in the transcription step of gene expression, where the RNA polymerase from the host may present low affinity for the metagenomic DNA promoter regions. Solutions to this issue include the co-expression of heterologous sigma factors with broad promoter recognition specificities (15), or the use of a high-efficiency expression systems based on viral components (7,11). (iv) Failure in recognition of Ribosome Binding Sites during translation, limiting the expression levels of the heterologous gene product, that may be improved by co-expressing proteins related to the mRNA recognition step (16). (v) Protein improper folding or lack of processing steps (including cleavage, secretion and peptide modifications) by the heterologous host, in some instances can be partially addressed by co-expression of chaperons (for proper folding) or processing proteins (7). (vi) Toxicity of the gene product, that can be addressed by using inducible systems for gene expression, such as a lambda phage lytic gene cassette under the control of a UV-inducible promoter (8). Other factors that are proposed to affect the heterologous gene expression include (vii) formation of inclusion bodies, (viii) lack of required co-factors and (ix) accelerated enzymatic breakdown by proteases from the host (14). It is generally accepted that the use of broad range shuttle vectors is desirable for the construction of metagenomic libraries as it may circumvent problems related to the heterologous host system by completely changing the cellular framework in which the gene products are being expressed (7,14–17). The most widely

used host for metagenomic libraries is *E. coli* (6,18), which has been tested for its expression potential against a set of different genomes, revealing a wide range that goes as low as 7% to as high as 73% of expressed genes (19).

In regards to the screening method, three main strategies can be adopted, although often the protocol involves a combination of two or more of these approaches: (i) phenotypic insert detection, where the expression of the heterologous gene product is directly detected in positive hits; (ii) modulated detection, where selective conditions are complemented by the heterologous gene product allowing growth of the library host or by the use of co-expressed reporter genes (20); and (iii) substrate induction, used mainly for the detection of catabolic gene products, usually inducible by metabolites of the catalytic gene products (14). Although it possesses lower sensitivity with a higher detection threshold, the direct phenotypic detection is the most widespread method for functional screening of metagenomic libraries (14), using chromogenic or fluorescent substrates to detect enzymatic activities of interest, or substrates that leave a clear halo around the colonies producing the desired gene product (7). The development of novel enzymatic assays has been driven by the need of more sensitive and targeting a broader range of biotechnologically interesting activities (11,21).

Thermozymes and their applications

Since their discovery, thermophilic microorganisms that thrive in high temperature environments such as hot springs have been considered an important source for thermozymes, as they have evolved and adapted their metabolism in harsh environmental conditions (22). Even more, thermozymes with higher thermal stability and optimal temperatures than the optimal growth temperature of source thermophilic microorganisms are common place (23). Many industrial applications could benefit from the use of thermozymes. Advantages such as lower operational costs and increased selectivity with a reduction of undesired by-products can be achieved by the use

of thermally stable catalysts (22), replacing chemical processes or mesophilic variants. The mechanisms underlying the improved thermal stability and optimal activity rates at higher temperatures had been studied (23–25), and other resistances and tolerances against other extreme conditions are often associated with thermozymes, such as activity at extreme pH values, and resistance to chemical denaturing agents and non-aqueous solvents (26) as well as proteolytic resistance (22,24) which may align with the requirements of the industrial processes. Operation at high temperatures has other associated benefits, such as lowering the risk of contamination from microorganisms (most pathogenic and food-rottening microorganisms cannot survive temperatures above 70°C) (24), lowering the viscosity of fluids allowing to reduce costs of processes like pumping and filtration and increasing the heat and mass transfer rates as well as solubility of substrates (allowing for equilibriums shifts towards reaction products) (23,24).

Metagenomic surveys have successfully retrieved several thermozymes from high temperature habitats including hot springs (11), and comprehensive lists can be found elsewhere showing the continuous interest in these bioprospecting studies for the past two decades (3,6,9,18,27,28). Many of the thermozymes sought are plant-cell wall degradation enzymes, acting on cellulose, the most abundant polymer on Earth, and other complex polymers like hemicellulose. These include cellulases, β-glucanases, xylanases and β-xylanases as well as other accessory enzymes like certain esterases (6,29–31); but there is interest in many other glycosidases (22,23) like β-galactosidases (10); amylases and chitinases (9); pullulanases (24); pectinases (32) as well as other non-carbohydrate related enzymes like proteases (28), phosphatases and oxidorreductases (6), among others. Potential applications for these thermozymes are very varied, cellulases (endoglucanases, exoglucanases and β-glucanases) and hemicellulases (xylanases and β-xylosidases) are important in the process of biofuel (bioethanol) production from lignocellulosic biomass (6,28,33,34) and some cellulases are used in the formulation

of detergents, among many other applications including paper and pulp, textile, and food processing industries (35,36); lipases can be used in biodiesel production, production of cosmetics, processing in leather and pulp industries, synthesis of drugs and fine chemicals and bioremediation (6,28,33,34,37) while some esterases are accessory enzymes in the lignocellulosic biomass degradation process (29); and proteases are components for detergents, and can be used for processing in food, leather, pharmaceutical and textile industries (6,28,33,34).

In the present work, we centered our efforts on screening two metagenomic libraries for a range of different biotechnologically relevant activities that mainly relate to plant biomass valorization (29,30,32). These include cellulases and β-glucosidades, xylanases, β-xylosidases and feruloyl esterases. We also included more substrate-broad hydrolases, namely lipolytic enzymes and proteases. As the sources for the metagenomic DNA for both libraries were hot springs, it was expected that the targeted gene products would have high temperature optima and thermostability, among potentially other beneficial characteristics from an industrial applicability standpoint. After detecting positive hits for some of these activities, we proceeded to the purification and biochemical characterization of one of the lipolytic enzymes found.

Materials and Methods

Two metagenomic libraries were constructed with metagenomic DNA from two different hot spring sources, as described in the previous Chapter of this work. Copies of the original two libraries were then used to conduct a series of functional screening protocols for a variety of biotechnologically interesting enzymes. In the case of plates using a fluorescent substrate and liquid media-based screenings, in order to reduce costs, we constructed an additional screening plate pooling together wells in the same position across all the plates from the same library, resulting in two pooled plates for high-throughput screening.

Plate assay screening methods

Metagenomic libraries were plated on LB medium without glucose (38) consisting in 1% Bacto™ tryptone (BD, USA), 0.5% Bacto™ yeast extract (BD, USA), 0.5% NaCl and 1.5% Bacto™ Agar (BD,USA) supplemented with chloramphenicol (12.5µg/mL) for positive selection of clones harbouring the fosmid vector pCT3FK, and a collection of substrates for the detection of specific enzymatic activities. To maintain the same format as the 96-well plates the libraries were built in, we used a 48-pin replica-plate stamp soaked in ethanol and sterilized on the flame, resulting in two Petri dishes for every 96-well plate and for each of the conditions assayed. Cells were grown overnight at 37°C and checked for their activity, and they were kept for several days (up to five days) in these conditions to detect activity that could require expression over extended periods of time. Additionally, incubation at 65°C for 3 hours after the initial growth overnight was also tested to detect activities from thermozymes that may not manifest at lower temperatures.

Cellulase screening methods on agar plates

The wells from the libraries were replicated on Petri dishes with modified LB media containing 0.5% carboxymethylcellulose (CMC) (endoglucanase) or AVICEL (exoglucanase) and chloramphenicol (12.5µg/mL). To induce the vector to high copy number (in order to increase the number of copies of the gene so that gene expression was maximized) 0.02% (w/v) arabinose was also added to the medium. The activity was assayed following three different plated-media protocols using these substrates to ensure minimal false positive or false negative hits, and the methods were tested with a positive control using a commercial endoglucanase solution (Cellulase from *Aspergillus niger*, Sigma-Aldrich, USA). An additional insoluble substrate was also tested for endoglucanase activity and another additional liquid-assay fluorescence-based protocol for the detection of endoglucanases was also employed as described in the next sections.

Congo red staining protocol for CMC and AVICEL substrates

After letting clones grow overnight at 37°C the plates were incubated at 65°C for 3 hours and then dyed with a 0.1% Congo Red solution for 15 minutes (39). As both the temperature and the dyeing treatments are lethal to the cells, a replica plate with only LB media supplemented with antibiotic was also cultured. Excess of stain was removed incubating with 1M NaCl solution for 15 minutes and removing the solution afterwards. Positive hits were expected to form an orange halo around colonies (39–43).

Trypan blue staining protocol for CMC and AVICEL substrates

The modified LB media containing the appropriate substrate and antibiotic was also supplemented with 0.01% trypan blue dye (44). Blue-coloured plates were expected to screen with clear halos around positive hit colonies (44,45) after an overnight growth period at 37°C and incubation at 65°C for 3 hours, without further staining necessary. Replicas of the plates were also made as the thermal treatment rendered the cells unviable.

Gram's Iodine staining protocol for CMC and AVICEL substrates

The clones were let to grow overnight at 37°C and then they were incubated at 65°C for 3 hours. The plates were then stained with a Gram's Iodine solution (0.67% KI, 0.33% Iodine) for 5 minutes (46). Positive hits were expected to have clear halos around the colonies after the removal of the staining solution. As before, replicas of the plates were necessary due to the treatments performed affecting cell viability.

Azurine Cross-Linked Hydroxyethyl Cellulose (AZCL-HE-Cellulose) substrate protocol

LB media containing 0.02% arabinose and 12.5µg/mL chloramphenicol was supplemented with the insoluble substrate 0.1% AZCL-HE-Cellulose (MegaZyme, Ireland) (47) and was added as

top agar over LB plates to allow for a better distribution (48). The activity was expected to be visualized by the formation of blue halos around the colonies.

Screening for xylanase activity

Azurine Cross-Linked Xylan (AZCL-Xylan) substrate protocol

Similarly to endoglucanase activity screening, LB media containing 0.1% insoluble AZCL-Xylan (MegaZyme, Ireland) and supplemented with 0.02% arabinose and 12.5µg/mL chloramphenicol was used to test the clones for xylanase activity. To maximize the insoluble substrate diffusion in the most homogeneous way possible, this screening media was poured over a previously solidified LB media plate as top agar (49). For this method, blue haloes around positive clones were expected.

Screening for β-xylosidase activity

4-methylumbelliferyl-β-D-xylopyranoside (MUX) substrate protocol

LB media containing the specific substrate 0.04% 4-methylumbelliferyl-β-D-xylopyranoside and supplemented with 0.02% arabinose and 12.5µg/mL chloramphenicol were employed for the screening of β-xylosidase activity in the metagenomic libraries (50). Examination under UV light would reveal positive hits due to the release of the fluorophore on the plates, requiring a replica plate as this exposure rendered the cells not viable.

Screening for lipolytic activity

For the detection of lipolytic activity, a medium containing 0.5% Bacto™ peptone (DB, USA), 0.3% Bacto™ yeast extract (DB, USA), 1% arabic gum (Acros Organics, USA), 1% (v/v) glyceryl tributyrate (Sigma-Aldrich, USA) and 1.3% Bacto™ agar (DB, USA) was used (51–54), supplemented with 0.02% arabinose and 12.5µg/mL chloramphenicol. All components except glyceryl tributyrate and

agar were dissolved in water, and after addition of glyceryl tributyrate an emulsion was obtained by using a blender for 5 minutes. Agar was added to the emulsion and the medium was autoclaved, adding independently sterilized arabinose and chloramphenicol afterwards. Positive hits would form clear haloes against an opaque background.

Screening for feruloyl esterases

As described in other functional screening studies, a second round of screening for feruloyl esterases was conducted with lipolytic-positive clones from the previous screening (51,55,56). Clones were plated on LB medium supplemented with 12.5µg/mL chloramphenicol and 0.02% arabinose after autoclaving and overlayed with 0.7% agar containing 20µg/mL of the substrate 4-methylumbelliferyl-trimethylammonium cinnamate chloride (MUTMAC) (Sigma-Aldrich, USA) to further search for feruloyl esterases (55). Activity was detected under examination with a UV lamp. Positive activity was expected to be detected as fluorescence centered around colonies.

Liquid assay screening methods

β-glucosidase activity screening

For the detection of β-glucosidase activity, liquid LB media containing 1mM para-nitrophenyl-(β-D)-glucopyranoside (Sigma-Aldrich, USA), 12.5µg/mL chloramphenicol and 0.02% arabinose was used. 700µL of media were dispensed in 96-deep-well plates and each well was inoculated with 10µL of clones from the working plates of the libraries. Growth at 37°C and agitation was sustained for 5 days, cells were precipitated by centrifugation and 100µL of supernatant were used to read absorbance at 400nm wavelength. Activity was expected to manifest as an increase of absorbance at $\lambda = 400nm$, and we adopted the threshold for positives hits as values above the mean absorbance across all wells plus two times the standard deviation (57).

Cellulase screening with fluorescent EnzChek Cellulase substrate

The EnzChek blue fluorescent Cellulase Substrate (Invitrogen, USA) was used to screen cellulase activity in both metagenomic libraries. A 384-well plate format was adopted, requiring volume adjustments to the suggested protocol from the supplier, using 10µL of enzyme crude extract and 10µL of substrate solution. The substrate was prepared according to the suggested manufacturer's protocol, using 50% DMSO to prepare a substrate solution. The digestion buffer used was 100mM sodium acetate, pH 5. Clones from the libraries were grown at 37°C for five days to allow natural lysis (57) in 96-well plates containing LB media supplemented with 12.5µg/mL chloramphenicol and 0.02% arabinose. Cells were centrifuged at maximum speed in a refrigerated (4°C) centrifuge, and supernatants were employed as crude extracts and transferred to opaque black 384-well plates for fluorescence reading with excitation and emission wavelengths of 360 and 460nm respectively. Assay time was 30 minutes. Similarly to the β-glucosidase activity screening, a threshold for positive hits was adopted requiring readings in the emission wavelength of relative fluorescence units at least above the mean across all wells plus two times the standard deviation (57).

Protease screening with Bodipy Casein fluorescent substrate

The EnzChek green fluorescent Bodipy Casein Substrate (Invitrogen, USA) was used to screen protease activities in the two metagenomic libraries. We followed the manufacturer's protocol adapting the format from 96-well plates to 384-well plates (10µL of working solution consisting in reconstituted substrate in Phosphate Buffer Saline and digestion buffer, and 10µL of enzyme crude extract). As with cellulase screening using the EnzChek subtrate, growth of clones was maintained for 5 days in LB media supplemented with 12.5µg/mL chloramphenicol and 0.02% arabinose and crude extracts were obtained by centrifugation in a refrigerated centrifuge. Assay time was 1 hour and excitation and emission wavelengths were 485 and 530nm respectively. The same threshold of mean across all wells plus

two times the standard deviation was adopted for positive hits determination (57).

Subcloning of selected clones and sequencing

After the identification of positive hits through functional screening, clones were selected based on factors like clarity and time to develop the positive signal. In total, four selected clones were further studied: LipB12_A11, LipM1_3D, XyloB5_D5 and GluBp_F7. Individual colonies were isolated by streaking them in the appropriate screening media. DNA was obtained from fresh cultures of LB media containing 12.5µg/mL chloramphenicol and induced to high copy number with 0.02% arabinose, following the procedure of the FosmidMAX DNA purification kit (Epicentre, USA) for 1.5mL cultures. First, a single colony was used to inoculate the media, and cells were grown overnight (OD_{600} ≈ 3). The culture was centrifuged at 13,000rpm for 3 minutes, and the cellular pellet was resuspended in 200µL cold FosmidMAX solution 1 (resuspension buffer). Then, 400µL of FosmidMAX solution 2 (lysis solution) were added, followed by 300µL of cold FosmidMAX solution 3 (precipitation reagent). The mixture was placed on ice for 15 minutes and then the cellular debris was pelleted by centrifugation at 4°C at 13,000rpm for 15 minutes. 540µL of isopropanol were added to the supernatant and a centrifugation at 4°C at 13,000rpm for 15 minutes allowed the precipitation of nucleic acids. The isopropanol supernatant was carefully removed, and the pellet was air-dried for 5 minutes before resuspension in 250µL of TE Buffer. 250µL of cold FosmidMAX solution 4 were added and residual impurities were removed by centrifugation at 4°C at 13,000rpm for 15 minutes. 1mL of ethanol was added to the recovered supernatant, and the nucleic acids were again precipitated following a centrifugation at 4°C at 13,000rpm for 15 minutes. The ethanol-containing supernatant was removed, and the pellet was air-dried for 5 minutes. The pellet was resuspended in 25µL of TE Buffer and 1µL of a previously prepared RNAse Mix were added to remove RNA that co-purifies with the fosmid (acting as a

carrier) by incubation at 37°C for 30 minutes. For DNA fragmentation, three different strategies were adopted: digestion with a set of restriction enzymes (*Eco*RV*, Mlu*I, *Hpa*I, *Aat*II, *Nhe*I, *Pvu*I, *Hind*III, *Bam*HI and *Cla*I), shearing using g-TUBEs (Covaris, USA) in a centrifuge for 30 seconds at 11,000 rpm or by sonication with a setting of 10 seconds with 30% amplitude and pulses of 1 second ON and 1 second OFF. Subcloning was performed using the CloneJET PCR Cloning kit (Thermo Fisher Scientific, USA) using the pJET1.2 blunt vector and following the manufacturer's instructions. For transformation, competent *E. coli* XL-1Blue (*rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *sup*E44 *rel*A1 *lac* [F´ *pro*AB *lacI*^{q}ZΔM15 *Tn*10 (Tet^{r})]) was used. Successful subcloning was checked using the same screening methods as described for each enzymatic activity, replacing the chloramphenicol antibiotic in culture media with ampicillin and with no arabinose.

Fosmids that harboured selected clones LipM1_3D, XyloB5_D5 and GluBp_F7 were sequenced by an external service provided by Novogene (Novogene, UK). The sequencing files were uploaded to the KBASE web service (58) for analysis. Reads were assembled using metaSPAdes (59) to generate contigs that would represent the cloned fragments sequences. Contigs were binned with MetaBAT2 (60) and the assembly was annotated using Prokka (61) to find the putative ORFs responsible for the detected enzymatic activities. Contigs were downloaded and further analysis were performed. First, they were aligned to the NR/NT nucleotide collection database from NCBI using the BLAST algorithm (62) on the default settings optimized for somewhat similar sequences (BLASTn with match score = 2, mismatch score = -3, gap existence cost = 5 and gap extension cost = 2). Then ORFs were predicted using ExPASy Translate Tool (63) and aligned to the NR protein sequences database from NCBI using the BLASTp algorithm on the default settings (protein-protein BLAST with BLOSUM62 matrix, cost of gap existence = 11 and of gap extension = 1).

Subcloning of a DNA fragment containing a lipolytic activity-conferring gene

Digestion with the restriction enzyme *Eco*RV was performed on a DNA extraction of a selected clone from the As Burgas metagenomic library (LipB12_A11) that showed lipolytic activity in the functional screening as described in the previous section. 1µg of fosmid DNA was used in a restriction reaction using *Eco*RV (Roche, Switzerland) with SuRE Cut Buffer B at 37°C for 1 hour. The resulting fragments from the digestion were used for ligation on the pJET1.2 vector following the CloneJET PCR Cloning kit (Thermo Fisher Scientific, USA) user manual. Transformation of the ligation reaction was performed on chemically competent cells, *E. coli* XL1Blue, by the heat-shock method and colonies were cultured on the media outlined for the screening of lipolytic activity, replacing the antibiotic chloramphenicol with ampicillin for positive selection of the new vector and without arabinose. Single colonies showing clear halos on the selection media were picked for further studies.

Sequencing and primer walking of a DNA fragment containing a lipolytic activity-conferring gene

Sanger sequencing was performed on DNA extracted with the NZYMiniPrep (NZYTech, Portugal) from lipolytic activity-positive transformants from the subcloning step (sequencing services were provided by Sistemas Genómicos, Valencia, Spain). Primer design for primer walking along the sequence was performed using the NetPrimer analysis software (64) (Premier Biosoft, USA). A list of the primers used is provided in Table 1. Primer walking was performed until the forward and reverse sequences overlapped with each other.

Table 1. Sequencing primers used in the primer walking sequencing process for the lipolytic activity conferring DNA cloned fragment.

Primer	**Sequence (5' – 3')**
pJET1.2 Forward	CGACTCACTATAGGGAGAGCGGC
pJET1.2 Reverse	AAGAACATCGATTTTCCATGGCAG
PF1	GGATTTACTGTATGACGGGATACCT
PR1	AGTCGCACTACACGGTCGTTCTGTT
PF2	GGAGCGGCTTTATTTTACCCACTTT
PR2	CAGAAACAAGGTTGGGACAAGCACT
PF3	CATCCGTTGTCTTTGGCTAACTGAA
PR3	TCTCAACATAAGGAAAGCGAAACG
PF4	ATGACTGGAAAAGCAAGGGGAATCT
PR4	GATACCGAAACGATGACAGCCCTTA
PF5	ATTAGGGGATTCATCAGGGCACAGT
PR5	ATCGCTTGTAAGACGGAGGGAATA

Sequence analysis

The sequence of the ORF was predicted using ExPASy Translate Tool (63) and aligned to the NR/NT nucleotide collection database from NCBI using the BLAST algorithm (62) on the default settings optimized for somewhat similar sequences, and to the NR protein sequences database from NCBI using the BLASTp algorithm on the default settings. The protein sequence structure was modelled with the SWISS-MODEL (65) web tool, using the best matching template

provided by the database. The model was then visualized using the PyMol software (Schrödinger LCC, USA) (66). The protein parameters were estimated using the PROTPARAM tool from ExPASy (67). Multiple sequence alignments were performed using CLUSTAL OMEGA (68) from EMBL-EBI with sequences retrieved from the UniProt Knowledgebase database (69). Conserved amino acids were visualized using the WEB LOGO service (70) for protein sequences.

Purification

For expression purposes, the LipB12_A12 insert subcloned in the pJET1.2 plasmid was transferred from *E. coli* XL1-Blue to *E. coli* T7 Express. DNA was extracted from fresh overnight-grown cultures using the NZYMiniPrep (NZYTech, Portugal) kit and used for transformation by the heat-shock method following protocols outlined previously.

Test were conducted to find optimal induction conditions using varying times, temperatures and concentrations of IPTG as well as optimal purification steps conditions for differential thermal precipitation (times and temperatures of incubation) and ammonium sulphate precipitation (concentration), allowing us to stablish a expression and purification protocol outlined below.

A pre-inoculum was prepared by picking a single colony from an overnight grown plate of LB media supplemented with ampicillin and transferring to a 10mL LB medium supplemented with ampicillin in a 50mL flask. The pre-inoculum was grown overnight and was used to start a culture at initial OD_{600} of 0.2. This culture was grown to an OD_{600} of 0.8 and then was induced with 0.04mM IPTG (final concentration) for 2 hours. The cells were pelleted in a refrigerated centrifuge (4°C) and washed twice using cold milliQ water. The pellet was resuspended in ice-cold sonication buffer 50mM Tris HCl pH 8.0, EDTA 25mM, NaCl 25mM and sonicated immersed in ice-cold water using a VCX130 Vibra-Cell sonicator (Sonics & Materials INC., USA) with a setting of 10 minutes total time, with pulses of 2 seconds ON

and 8 seconds OFF to avoid excessive heating and an amplitude of 100%. The cell debris was precipitated using a refrigerated centrifuge (4°C) and the crude extract was recovered. This crude extract was treated at 60°C for 20 minutes for an initial step of purification to precipitate mesophilic proteins from the host and the vector. The thermally treated extract was again centrifuged at 4°C and the supernatant was recovered. An ammonium sulphate incubation at final concentration 30% (v/v) for 1 hour in a rotatory wheel was performed to remove a fraction of proteins. After a centrifugation at 4°C the recovered supernatant was incubated with ammonium sulphate at a final 50% concentration (v/v) which made the protein of interest to precipitate. Centrifugation at 4°C allowed to remove the supernatant and recover the pellet, that was resuspended in buffer 0.1M Tris HCl pH 9; 0.1M NaCl and 1mM DTT. The same resuspension buffer was used to perform a dialysis with a SnakeSkin Dialysis Tubing membrane with a 3.5 Molecular Weight cut-off (Thermo Fisher Scientific, USA) overnight at 4°C to remove the ammonium sulphate, substituting the 1mM DTT with 1mM β-mercaptoethanol. Lastly a molecular exclusion chromatography was performed using a HiLoad 16/60 Superdex^{TM} 75 prep grade column (Cytiva, USA) with a 1mL/minute flow rate collecting 1mL fractions. Fractions showing lipolytic activity and without contaminants visualized in SDS-PAGE were combined to perform the biochemical characterization. The pooled fractions were concentrated using a Pierce Concentrator column with 10K Molecular Weight cut-off (Thermo Fisher Scientific, USA).

Activity assays

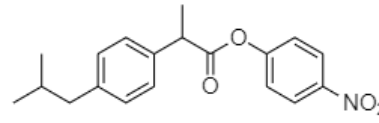
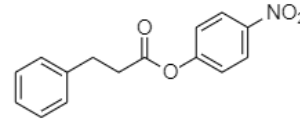
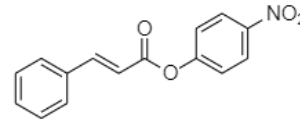
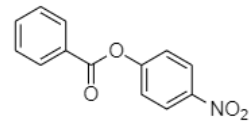
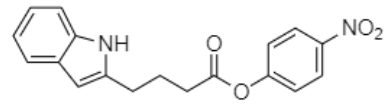
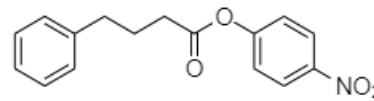
All activity assays were performed in triplicate. Lipolytic activity was tested using 25mM pNP-Laurate (C12) as the substrate and 0.1M Sodium Phosphate Buffer pH 7 (at 60°C) as reaction buffer. An enzymatic unit is defined as the micromols of p-nitrophenol released per minute and per microlitre of enzyme in the reaction conditions. 20µL of purified enzyme were mixed with 730µL reaction buffer and

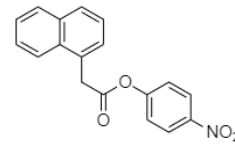
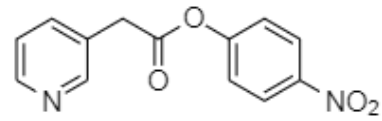
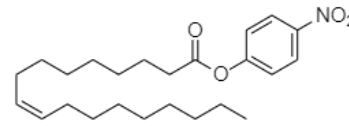
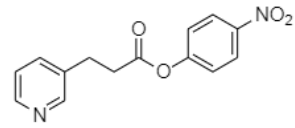
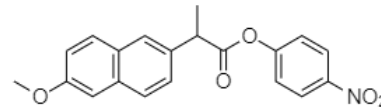
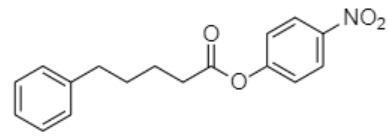
100µL substrate solution and incubated for 30 minutes at the reaction temperature (60°C unless indicated otherwise). The reaction was stopped using 250µL 0.1M ice-cold $NaCO_3$ and transferring to ice. Enzymatic activity was estimated as release of p-nitrophenol which was measured as increments the absorbance at 400nm wavelength. The extinction coefficient was determined elsewhere as 17.215 mM/cm (71).

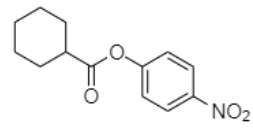
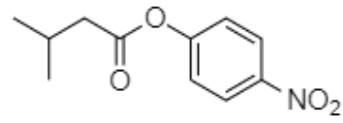
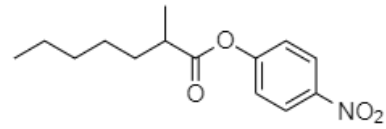
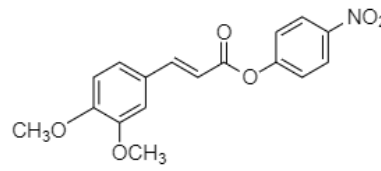
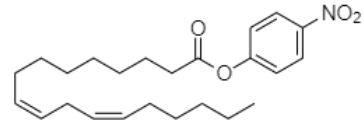
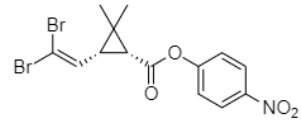
For temperature optima tests, the reaction temperature was changed accordingly for the range between 40 to 90°C. For pH optima tests, the reaction buffer was 0.1M Sodium Phosphate Buffer for the range 6 to 8, 0.1M Sodium Acetate Buffer for the range 4 to 6 and 0.1M Tris HCl Buffer for the range 8 to 9.5. Thermostability was assessed by incubating the purified enzyme at a specified temperature and for varying amounts of time before performing the activity test at optimal pH and temperature.

A library of synthetic pNP ester substrates (72) was kindly provided by the authors and was also assayed, each substrate at a final concentration of 0.2mM using the reaction buffer Tris HCl pH 9 (at 60°C) supplemented with 0.1% Arabic gum and 1% CHAPS. Reaction volumes were adjusted to 5µL of purified enzyme, 8µL of substrate and 387µL reaction buffer. Table 2 shows the chemical formula for this synthetic compounds.

Table 2. Synthetic substrates used in this study.

Code	Substrate	Chemical formula
76	4-nitrophenyl 2-(4-isobutylphenyl)propanoate)	
77	4-nitrophenyl 3-phenylpropanoate	
78	4-nitrophenyl cinnamate	
79	4-nitrophenyl benzoate	
80	4-nitrophenyl 4-(1H-indol-2-yl)butanoate	
81	4-nitrophenyl 4-phenylbutanoate	

82	4-nitrophenyl 2-(naphthalen-1-yl)acetate	
83	4-nitrophenyl 2-(pyridin-3-yl)acetate	
84	4-nitrophenyl oleate	
85	4-nitrophenyl 3-(pyridin-3-yl)propanoate	
86	4-nitrophenyl 2-(6-methoxynaphthalen-2-yl)propanoate	
87	4-nitrophenyl 5-phenylpentanoate	

88	4-nitrophenyl cyclohexanecarboxylate	
90	4-nitrophenyl 3-methylbutanoate	
91	4-nitrophenyl 2-methylheptanoate	
92	4-nitrophenyl (E)-3-(3,4-dimethoxyphenyl)acrylate	
94	4-nitrophenyl linoleate	
100	(1R,3R)-4-nitrophenyl 3-(2,2-dibromovinyl)-2,2-dimethylcyclopropanecarboxylate	

Enzyme kinetics were tested using the following pNP-laurate concentrations: 2mM, 0.5mM, 0.25mM, 0.1mM, 0.02mM and 0.002mM. Enzyme volume was reduced to 10µL instead of the 20µL used in standard assays.

The effect of different additives was tested at final concentrations 5mM: NaCl, KCl, $CaCl_2$, $ZnSO_4$, $CuSO_4$, $AgNO_3$, $FeCl_3$, $FeCl_2$, $MgCl_2$, $MnCl_2$, $NiCl_2$, SDS, Triton X-100, Tween 20, Tween 80, CHAPS, CTAB and EDTA.

Results

Screening for plant cell wall degrading activities

No positive exoglucanase clones were found using the screening protocols with AVICEL as the substrate in the two libraries. Neither did we found any xylanase activity conferring gene by screening the two metagenomic libraries using plates containing the AZCL-Xylan substrate.

No positive hits were found using the methods to find endoglucanases with CMC or AZCL-HE-Cellulose as substrates (solid media plate chromogenic substrate assays), but 2 As Burgas clones and 4 Muiño da Veiga clones were found positive for endoglucanase activity using the EnzChek fluorescent substrate (liquid media fluorogenic assays). We used a naming convention based on the library where the clone was found (B for As Burgas and M for Muiño da Veiga) and the plate and well within the library (p was used instead of a number when the plate used was a pooled library), and thus the clones were labelled CelBp_H8, CelBp_H10, CelMp_G4, CelMp_H4, CelMp_D6 and CelMp_F12. Differences in the substrates' detection threshold are expected between chromogenic and fluorogenic molecules, especially when low amounts of enzymes are produced. The format of the screening changing from colonies on agar plates to 384-well plates can also improve the sensitivity of the method (8,16).

Screening for β-xylosidase activity using the fluorogenic substrate MUX revealed 2 positive hits from the As Burgas metagenomic library but did not reveal any clone with that activity in the Muiño da Veiga library. These clones were named XyloB5_D5 and XyloB5_E5 following the same naming convention stablished before. For β-glucosidase activity using pNP-glucopyranoside as substrate we found 5 positive hits for the As Burgas library and only 1 for the Muiño da Veiga library. Clones were named accordingly GluBp_G1, GluBp_F2, GluBp_G5, GluBp_F7 and GluBp_F10 for As Burgas positive hits and GluMp_H3 for the Muiño da Veiga library.

Screening for protease activity

4 positive hits were found using Bodipy Casein fluorescent substrate in the As Burgas metagenomic library and 6 in the library from Muiño da Veiga. Again, these positive clones were named based on their activity profile, source metagenomic library and well within the library as follows: ProBp_D1, ProBp_H1, ProBp_E9, ProBp_F12, ProMp_H1, ProMp_A10, ProMp_A11, ProMp_A12, ProMp_C12 and ProMp_G12.

Screening for lipolytic and feruloyl esterase activities

The screening in the As Burgas metagenomic library using the glyceryl tributyrate substrate gave a single positive hit labelled LipB12_A11. Additionally, we found 8 different clones harbouring lipolytic activities in the Muiño da Veiga library, named LipM1_D3, LipM2_C2, LipM2_H2, LipM3_G1, LipM3_F5, LipM5_H7, LipM5_A8 and LipM5_E8. Subsequent streak cultures on screening media allowed the isolation of single colonies harbouring the lipolytic activity. Subsequent screening for feruloyl esterase activity of each of these clones using the fluorescent substrate MUTMAC did not yield any positive hit.

An overview of all the functional screening results is given in Table 3. Examples of positive hit identification in the functional screenings are provided in Figure 1.

Table 3. Overview of all the functional screenings performed on the two metagenomic libraries and the number of positive hits detected for each specific substrate.

Enzymatic activity	Substrate	Method	Library	
			As Burgas	Muiño da Veiga
Endoglucanase	CMC	Halo	0	0
	AZCL-HE-Celulose	Colorimetric	0	0
	EnzChek cellulase sustrate	Fluorimetric	2	4
Exoglucanase	AVICEL	Halo	0	0
β-glucosidase	pNP-glucopyranoside	Colorimetric	5	1
Xylanase	AZCL-Xylan	Colorimetric	0	0
β-xylosidase	MUX	Fluorimetric	2	0
Protease	EnzChek Casein	Fluorimetric	4	6
Lipolytic	Tributyrin	Halo	1	8
Feruloyl esterase	MUTMAC	Fluorimetric	0	0

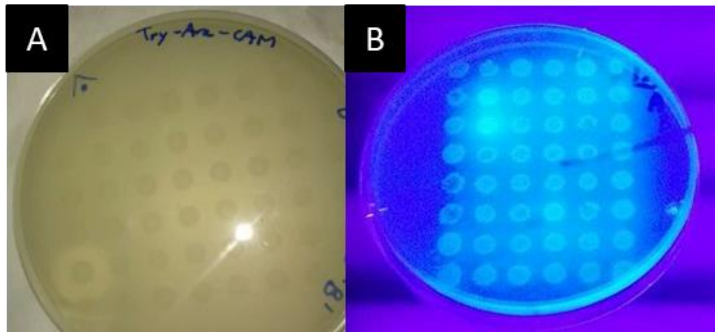


Figure 1. (A) Lipolytic activity screening with a positive hit from the As Burgas metagenomic library using glyceryl tributyrate as the substrate. Direct observation of haloes surrounding the colonies possessing the activity allows to identify positive hits. (B) β-Xylosidase activity screening with two positive hits from the As Burgas Metagenomic library using MUX as the substrate. Observation under UV-light for fluorescence around the colonies is required to identify positive hits.

Subcloning of selected positive hits and sequence analysis

Despite our best efforts, no activity was observed in any of the subclones obtained through restriction enzyme digestion nor through random shearing methods for LipM1_3D, XyloB5_D5 and GluBp_F7. LipB12_A11 was suscefully subcloned using the *Eco*RV restriction enzyme and its analysis is described in the next section.

Analysis of these three clones fosmid sequences revealed that LipM1_3D fosmid had a 58% query cover and 98.26% identity with *Thermus oshimai* JL-2, complete genome (Accession CP003249.1), GluBp_F7 a query cover of 75% and 99.99% identity with *Acinetobacter johnsonii* strain M19 chromosome, complete genome (Accession CP037424.1) and XyloB5_D5 a 74% cover and 99.99% identity with *Acinetobacter johnsonii* XBB1, complete genome (Accession CP010350.1). Annotation using the Prokka and the BLAST algorithms did not reveal annotated sequences in the database that aligned to the contigs and matched the observed activity from the screenings.

Sequencing of a subcloned metagenomic library DNA insert conferring lipolytic activity and sequence analysis

The DNA sequence of the subcloned LipB12_A11 lipolytic activity enzyme obtained by primer walking and sequencing and the translated protein sequence using ExPASy translate tool are given in Figure 2A. The predicted protein model using Lipolytic protein G-D-S-L family from *Desulfitobacterium hafniense* DCB-2 (SMTL ID 4rsh.1.A) as template is given in Figure 2B. Its Global Model Quality Estimation (GMQE) score was the higher out of all other templates at 0.62 and its QMEAN (73) quality estimation score was of -3.12. Expasy ProtParam predicted a molecular weight of 22241.45 Da and a theorical isoelectric point of 5.64.

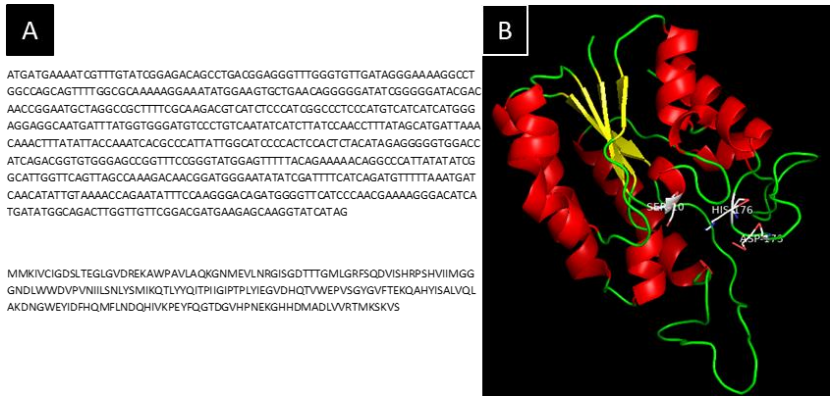
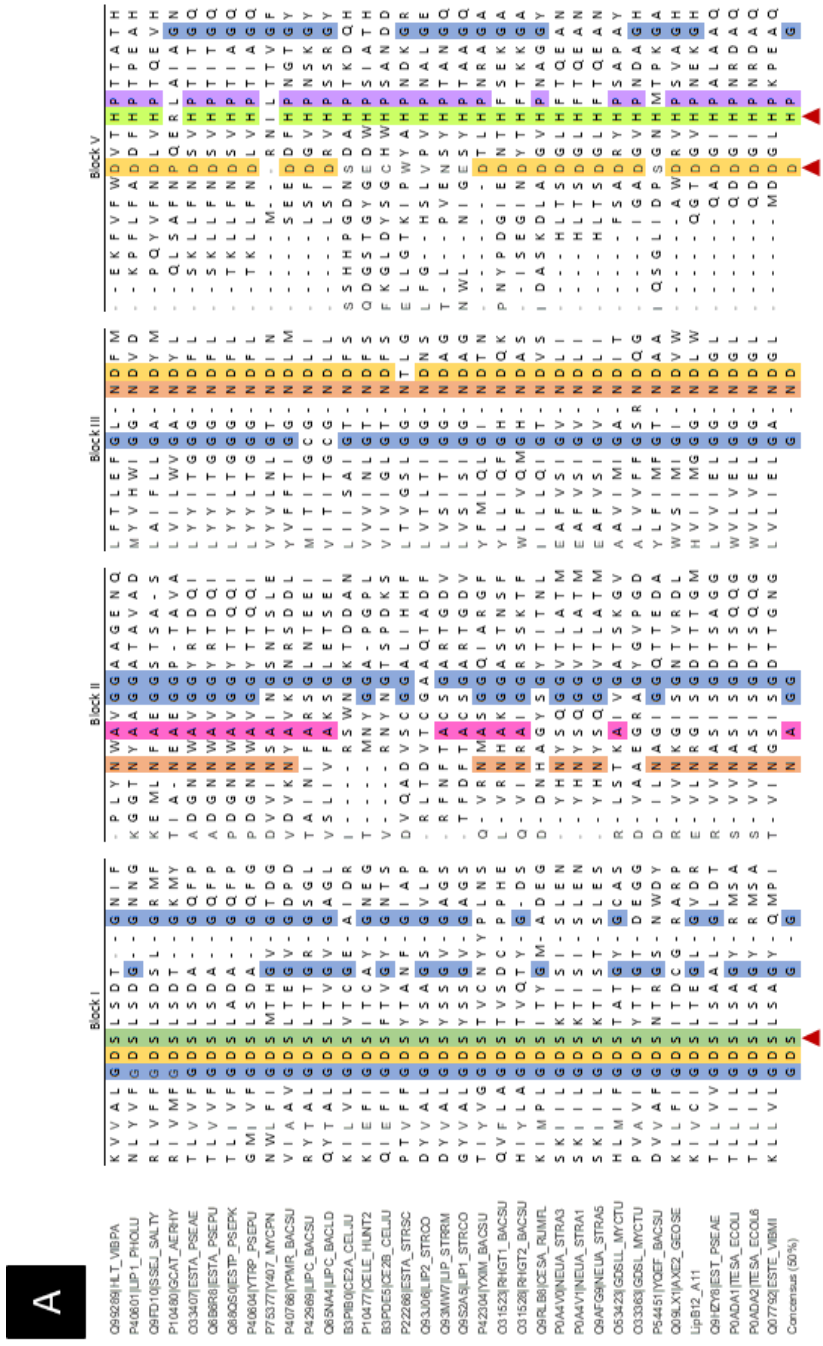


Figure 2. (A) DNA sequence and translated protein sequence of the DNA conferring lipolytic activity. (B) Predicted model of the lipolytic protein LipB12_A11 generated using SWISS-MODEL and visualized with PyMOL using a putative lipolytic protein of G-D-S-L family from *Desulfitobacterium hafniense* (SMTL ID 4rsh.1.A) as template. Secondary structures are highlighted in colour with α-helices in red, β-sheets in yellow and ribbons in green. The catalytic triad residues are highlighted (SER10, ASP173 and HIS176).

Alignment results using NCBI databases for DNA and protein sequences

The best match for the DNA sequence in the NR/NT database was the chromosome sequence of *Bacillus citotoxycus* (accession code CP024120.1) with an E-value of 2E-33. The query cover was of 94% and the percentage of identity of 66.84% with a Total Score of 156. For the protein database, the best match was a SGNH/GDSL hydrolase family protein from *Bacillus* sp. (accession code WP_028398363.1) with an E-value of 6E-84, a total score of 256, query cover of 95% and identity percentage of 58.73%. BLASTp identified the protein domain architecture as SGNH/GDSL hydrolase family protein: “hydrolytic enzyme such as an esterase or lipase; may have multifunctional properties including broad substrate specificity and regiospecificity”. Indeed, known functions of this protein family are broad and potentially useful in a variety of biotechnological applications (74).

A search for reviewed proteins in the UNIPROT database with the term “GDSL” and with “bacteria” as a taxonomy filter gave 43 results. Manual review of each entry allowed to remove entries that were not true GDSL family proteins (lacking the conserved block of amino acids) and reduced this number to 34 proteins containing the conserved motif. These were submitted for multiple sequence alignment along the lipolytic enzyme found in this study using the Clustal Omega web service. Clustal Omega results of multiple sequence alignments revealed previously described conserved blocks of amino acids in the lipolytic enzyme (74), as shown in Figure 3A. Results were compared to previously described conserved blocks and consensus sequences (74), revealing four conserved blocks previously reported but slight differences in the consensus sequence at a 50% threshold of amino acid occurrences in the same position across all sequences. Moreover, amino acids frequencies in these four blocks were better visualized using the Web Logo tool as shown in Figure 3B.


A
Block I
Block II
Block III
Block V
Q9929B|HLT_VIBPA
P40601|LIP1_PHOLU
Q9FD10|SSEJ_SALTY
P10480|SCAT_AERHY
O33407|ESTA_PSEAE
Q6BBR8|ESTA_PSEPU
Q8GQ50|ESTP_PSEPK
P40604|YTRP_PSEPU
P75377|Y407_MYCPN
P40768|YPMR_BACSU
P42988|LIPC_BACSU
Q65NA4|LIPC_BACLD
B3PIB0|CE2A_CELJU
P10477|CELE_HUNT2
B3PDE5|CE2B_CELJU
P22286|ESTA_STRSC
Q9JU06|LIP2_STRCO
Q93MW7|LIP_STRRM
Q9S2A5|LIP1_STRCO
P42304|YXIM_BACSU
O31523|RHGT1_BACSU
O31528|RHGT2_BACSU
Q9RLB8|CESA_RUMFL
P0A4V0|NEUA_STRA3
P0A4V1|NEUA_STRA1
Q9AFG9|NEUA_STRA5
O53423|GDSL1_MYCTU
O33363|GDSL_MYCTU
P54451|YQEF_BACSU
Q09LX1|AXE2_GEOSE
LipB12_A11
Q9HZY0|EST_PSEAE
P0ADA1|TESA_ECOLI
P0ADA2|TESA_ECOL6
Q07792|ESTE_VIBMI
Consensus (50%)

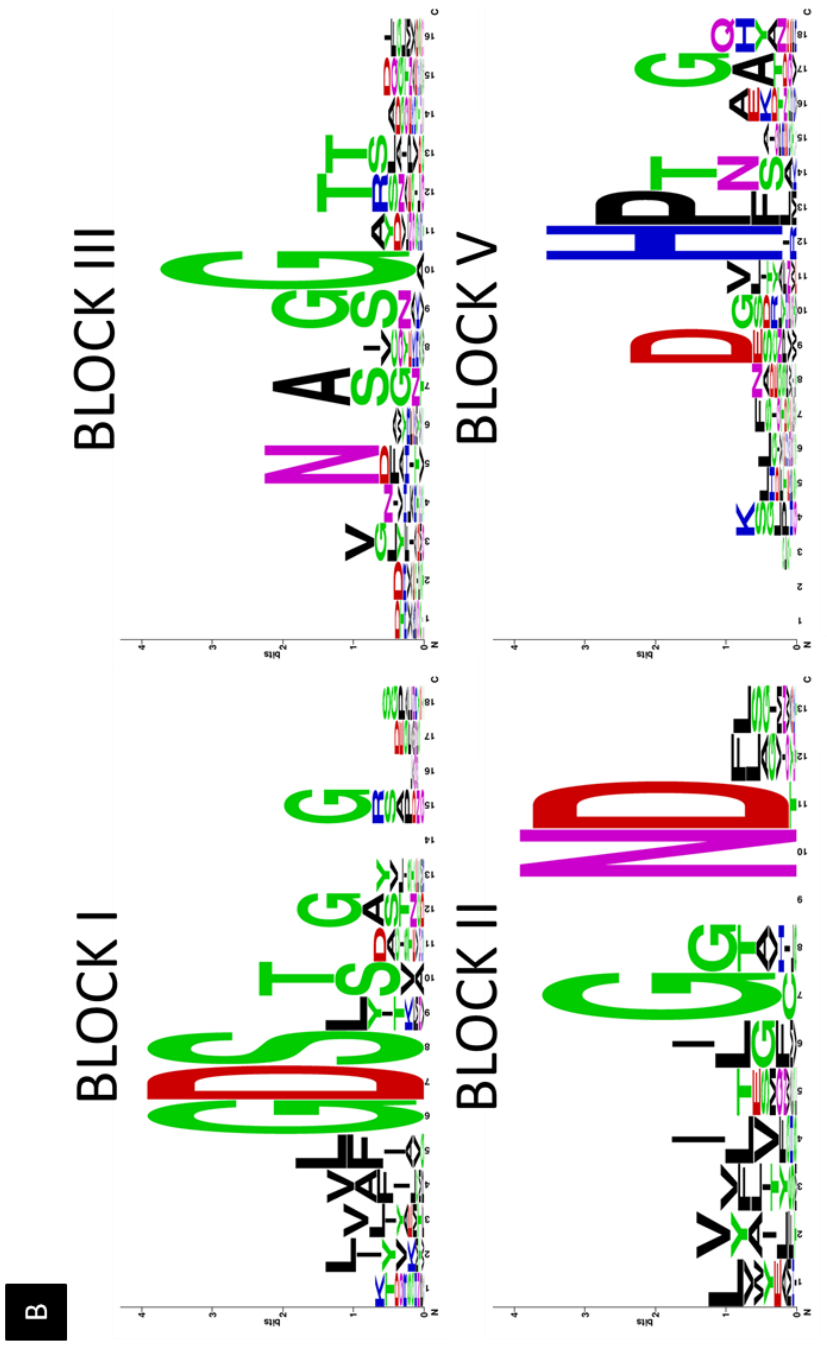

B
BLOCK I
BLOCK II
BLOCK III
BLOCK V

Figure 3. (A) Multiple sequence alignment of GDSL family lipases of bacterial origin reviewed and deposited in the UniProt database, including the lipolytic activity enzyme found in this work. Four conserved blocks are represented. A consensus sequence of at least 50% conserved residues among all sequences is given, and the three residues from the catalytic triad are highlighted as reported elsewhere (74). (B) Protein web logos resulting from the alignments of the amino acids from the four blocks.

Purification and biochemical characterization of the Lipolytic Activity Enzyme LipB12_A11

The enzyme was purified before performing the biochemical characterization. An overview of the purification steps is given in Figure 4 with a SDS-PAGE of the different fractions and as Table 4 with the protein yields after each step.

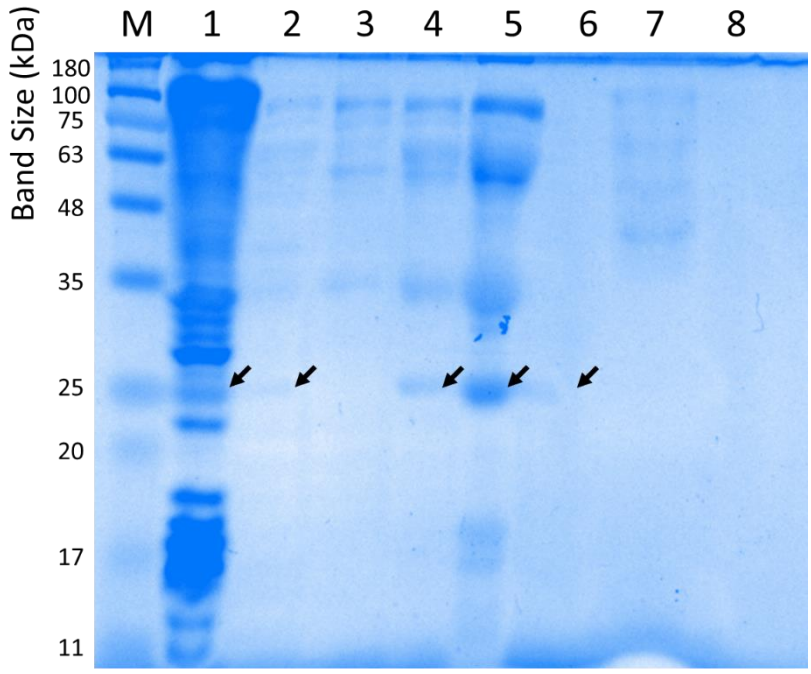


Figure 4. SDS-PAGE of the purification steps for LipB12_A11. (M) NZYColour Protein Marker II (NZYTech, Portugal), (1) crude extract, (2) supernatant after thermal treatment (60°C, 20 minutes), (3) pellet from 30% ammonium

sulphate precipitation, (4) pellet from 50% ammonium sulphate precipitation, (5) dialyzed and concentrated fraction, (6) pooled molecular exclusion fractions showing lipolytic activity, (7) supernatant from the 50% ammonium sulphate precipitation, (8) concentration column flow-through. Black arrows represent the LipB12_A11 enzyme migration band.

Table 4. Purification steps for LipB12_A11.

Purification step	Volume (mL)	Total Activity (U/µL)	Total protein (mg)	Specific Activity (U/mg)	Yield (%)	Purification Fold
Crude extract	16	0.000201	17.01	1,1807E-08	100	1
Differential thermal precipitation	8	0.000155	10.88	1,4287E-08	77.42	1.21
Ammonium sulphate precipitation	8	0.000149	4.33	3,4641E-08	74.69	2.93
Molecular exclusion chromatography	3	0.000145	2.70	5,383E-08	72.49	4.56

The enzyme was found to have an optimal temperature of 60°C with the substrate pNP-laurate (C12). The optimal pH was 9 but maintained its activity in the range between 7 and 9.5. The enzyme was found to be thermostable with a half-life of 11.9 hours at 60°C, 1.7 hours at 70°C and 1.1 hours at 80°C. Results from this biochemical characterization are represented in figure 5.

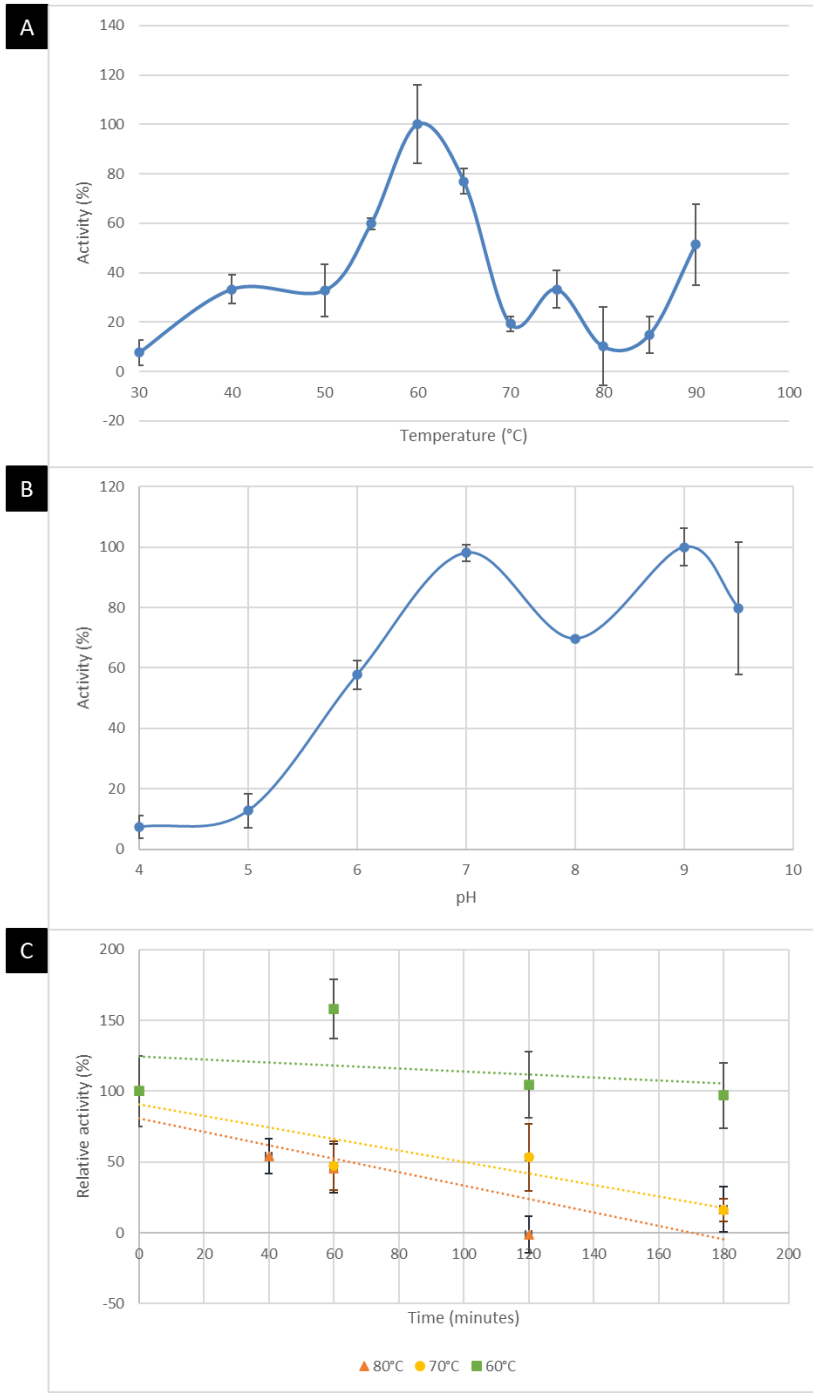

A
140
120
100
80
60
40
20
0
-20
Activity (%)
30
40
50
60
70
80
90
100
Temperature (°C)
B
120
100
80
60
40
20
0
Activity (%)
4
5
6
7
8
9
10
pH
C
200
150
100
50
0
-50
Relative activity (%)
0
20
40
60
80
100
120
140
160
180
200
Time (minutes)
80°C
70°C
60°C

Figure 5. (A) Optimal temperature for LipB12_A11. 100% activity is 1.58E-5 U/µL. (B) Optimal pH for LipB12_A11. 100% activity is 8.70E-6 U/µL. (C) Thermostability of LipB12_A11. 100% activity is 9,42E-6 U/µL.

Activity was measured using a range of different substrates. The results are summarized in Figure 6. The enzyme was most active using substrate pNP-octanoate (C8), pNP-laurate (C12) and 4-nitrophenyl 5-phenylpentanoate (substrate 87). Some degree of activity was observed towards pNP-hexanoate (C6) and compounds 78, 79, 80, 81, 82, 83, 85, 86 and 94. Longer esters like pNP-estearate (C18) and the other synthetic compounds assayed were not used as substrates by LipB12_A11.

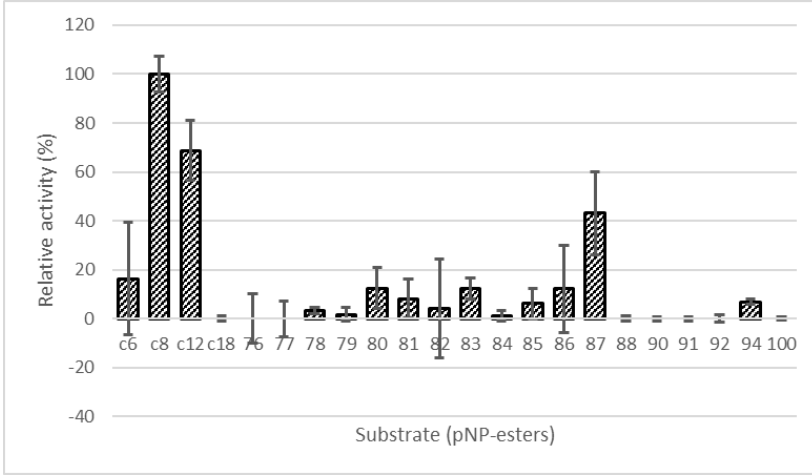


Figure 6. Use of various substrates for LipB12_A11 in relation with its activity towards pNP-octanoate. C6: pNP-hexanoate; C8: pNP-octanoate; C12: pNP-dodecanoate; C18: pNP-stearate; numbers represent synthethic substrates following the naming convention as represented in Table 2. 100% activity is 3.90E-05 U/µL.

Effect of additives on enzymatic activity

The effect of several additives was tested on the LipB12_A11 enzymatic activity. All metal ions decreased the enzymatic activity compared to the no additive control except Na^+ and K^+, but activity was also lowered when using the chelating agent EDTA. Ag^+ and Mg^{2+} completely inhibited the enzymatic activity. Surfactant Tween 80 almost had a 10-fold increase effect on the enzymatic activity. CHAPS also increased the enzymatic activity. On the contrary, Triton X100 inhibited all activity. Tween 20 and SDS had no significant effect on the enzymatic activity. Results on the relative activity of the enzyme with the additives are shown in Figure 7.

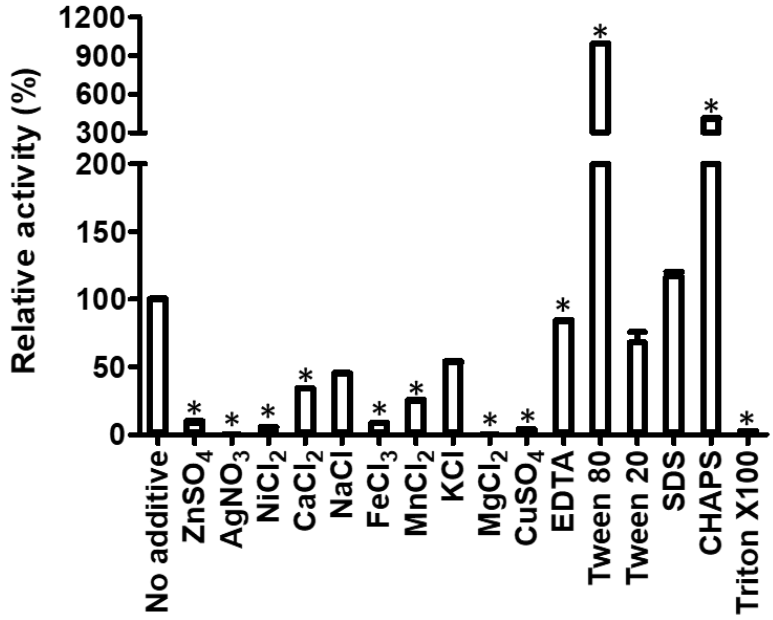


Figure 7. Effect of additives on LipB12_A11 lipolytic activity towards pNP-Laurate relative to the no additive control. Asterisks represent a p-value <0.05 for the t-test statistical analysis. 100% activity is 1.35E-5 U/µL.

Enzyme kinetics

The enzymatic parameters were calculated for LipB12_A11 using varying concentrations of substrate. Results are represented in Figure 8 showing the Michaelis-Menten fit of the enzyme kinetics and the non-linear regression. The enzyme kinetic parameters with the Michaelis-Menten model were V_{max} = 3.487E-005 ± 1,091E-006 U/µL and K_M = 0.01731 ± 0.003506 mM, with a coefficient of R^2 = 0.9425.

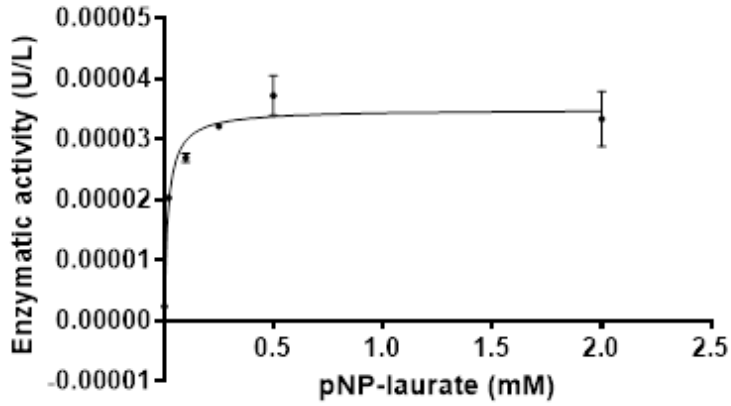


Figure 8. Enzyme kinetics for LipB12_A11 using varying concentrations of the substrate pNP-Laurate at 60°C and pH 9. Michaelis-Menten non-linear regression was calculated using the GraphPad Prism software.

Discussion

Functional screenings revealed several positive clones for a diverse range of enzymatic activities that hold biotechnological potential, and that may be thermozymes. Although the scope of the screening allowed this identification, practical limitations dictated the need to select only some of the clones for a more thorough analysis. Of the selected clones, only the lipolytic enzyme LipB12_A11 was successfully subcloned and its ORF identified, as it aligned to a known function protein. On the other hand, neither LipM1_3D, GluBp_F7

nor XyloB5_D5 could be subcloned, highlighting the difficulty to work with unknown sequences as it forces the use of random methods for shearing the DNA. Although we obtained the full sequence of the clones to try to overcome this problem, none of the predicted ORFs could be identified as responsible for the activities found in the screening. Nonetheless, the DNA inserts were aligned with high identity to *Acinetobacter johnsonii* for both β-glucosidase GlupBp_F7 and β-xylosidase XyloB5_D5. *Acinetobacter* isolates have been indeed described to possess the two activities (75,76). Similarly, the sequence of the fosmid harbouring the lipolytic enzyme LipM1_3D had high identity with the *Thermus oshimai* JL-2 genome, in which a carboxylic ester hydrolase and an esterase/lipase had been annotated (77). More studies are needed to confirm if these enzymes found by functional screenings represent novel enzymes and indeed research in this regard is in progress.

The lypolitic enzyme LipB12_A11 was identified, and its sequence allowed it classification as a SGNH/GDSL hydrolase. Proteins belonging to the SGNH/GDSL hydrolase family (74) contain the conserved GDS(L) aminoacids including the active site serine, located near the N-terminus. In the LipB12_A11 sequence this motif and residue was located in serine 10. Also, in this protein family the conserved residues serine, glycine, asparagine and histidine (SGNH) are present in four conserved blocks numbered I, II, III and V and play important roles in the hydrolase function (74). The active site serine is the nucleophile and together with glycine from block II and asparagine from block III (residues Gly42 and Asn71 in LipB12_A11) they act as proton donors for the oxyanion hole. Aspartic acid and histidine from block V are the other two members of the catalytic triad (Asp173 and His176). This histidine residue makes the active site serine more nucleophilic via hydroxyl group deprotonation (74). One of the key characteristics of this family of hydrolases is the wide range of substrates they can accept. Of the 34 GDSL family proteins of bacterial origin reviewed in UniProt, two belonged to thermophilic organisms, namely Cellulase/esterase (CelE) from *Clostridium*

thermocellum and Acetylxylan esterase (Axe2) from *Geobacillus stearothermophilus*. Axe2 had a temperature and pH optimum between 50 and 60°C and 7.1 and 9.2 respectively using p-nitrophenyl acetate and 2-naphthyl acetate as substrates (78). The biochemical characterization of CelE is not documented but assays had been performed at temperatures up to 60°C using CMC and Xylan as substrates at pH 7.0 (79), and at 37°C using substrates 4-nitrophenyl acetate and acetylated glucomannan at pH 6.5 (80). Although not included in the reviewed UniProt database, other thermostable GDSL family proteins have been reported, including the GDSL family esterase from *Geobacillus thermodenitrificans* T2 EstL5 with temperature and pH optima at 60°C and 8.0 respectively with pNP-butyrate as the substrate (81); an acetyl xylan esterase from *Caldicellulosiruptor bescii* Cbes-AcXE2 with an optimum temperature of 70°C and optimum pH of 7.0 using pNP-acetate as the substrate (82); and a GDSL-type lipase from *Geobacillus thermocatenulatus* Lip29 that had optimal activity at 50°C and pH 9.5 using pNP-palmitate as the substrate (83). Our results show that the enzyme LipB12_A11 found in the functional screening of the metagenomic libraries from hot springs possess comparable temperature optimum to other identified thermostable enzymes belonging to the SGNH/GDSL hydrolase family with an optimum of 60°C and pH 9 using pNP-laurate as the substrate. Activity in a broad pH range is observed and also in agreement with other enzymes of the family (78). Both substrate preference and presence of an interfacial activation mechanism are instrumental for classification of lipolytic enzymes as either lipases or esterases (83–85). The preference for short substrates like pNP-octanoate (C8) and pNP-laurate (C12) and lack of activity towards longer chain substrates like pNP-estearate (C18), coupled with the enzyme kinetics not showing an interfacial activation mechanism typical of true lipases suggest that the SGNH/GDSL hydrolase found in this study is an esterase. The enzyme was found to have some activity towards other alternative synthetic substrates such as 4-nitrophenyl 4-phenylbutanoate, 4-nitrophenyl 2-(pyridin-3-yl)acetate, 4-nitrophenyl linoleate and specially towards

4-nitrophenyl 5-phenylpentanoate. GDSL family hydrolases are known for their broad substrate specificity and stereoselectivity, associated with a highly flexible active site environment (74), thus, activity towards many structurally varied nitrophenyl esters was expected. Regarding its thermostability, the enzyme was highly thermostable at temperature 60°C with no significant loss of activity after a 3 hours incubation period, similarly to other reported thermostable GDSL family esterases (83). Moreover, a thermal activation effect was observed at that temperature, not unlike what have been observed with other proteins in the family (81). Activity was heavily affected by the presence of metallic ions, which also has been reported for members of the GDSL family (83), although not to the same extent and with as many metal species as reported here. Lastly, detergents also impacted the activity of the enzyme: CHAPS and most notably Tween 80 enhanced the activity whereas Triton X100 decreased it.

Conclusions

This study consisted in the bioprospecting of two metagenomic libraries through functional screening in order to find potentially industrially relevant enzymatic activities. Several positive hits were found for both libraries for some of the activities tested. Further studies were conducted in a selected clone that harboured lipolytic activity, which allowed the classification as novel esterase in the SGNH/GDSL family of hydrolases, with 60°C and pH 9 optima using pNP-laurate as the substrate. The enzyme was found to be highly thermostable at 60°C and its activity can be enhanced with detergents like Tween 80. As enzymes of this family have numerous potential biotechnological applications, these parameters should be all considered when deciding the best fit for the enzyme applicability.

References

1. Felczykowska A, Krajewska A, Zielińska S, Łoś JM, Bloch SK, Nejman-Faleńczyk B. The most widespread problems in the function-based microbial metagenomics. Acta Biochim Pol [Internet]. 2015 Feb 13;62(1):161–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25680374

2. Escuder-Rodríguez J-J, DeCastro M-E, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Advances of Functional Metagenomics in Harnessing Thermozymes. Metagenomics: Perspectives, Methods, and Applications. 2017.

3. Ferrer M, Golyshina O, Beloqui A, Golyshin PN. Mining enzymes from extreme environments. Curr Opin Microbiol [Internet]. 2007 Jun;10(3):207–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369527407000513

4. Mardanov A V., Kadnikov V V., Ravin N V. Metagenomics: A Paradigm Shift in Microbiology. In: Metagenomics [Internet]. Elsevier; 2018. p. 1–13. Available from: https://doi.org/10.1016/B978-0-08-102268-9.00001-X

5. Armstrong Z, Rahfeld P, Withers SG. Discovery of New Glycosidases From Metagenomic Libraries. In: Methods in Enzymology [Internet]. 1st ed. Elsevier Inc.; 2017. p. 3–23. Available from: http://dx.doi.org/10.1016/bs.mie.2017.06.001

6. Decastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol. 2016;7(September):1521.

7. Guazzaroni ME, Silva-Rocha R, Ward RJ. Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. Microb Biotechnol. 2015;8(1):52–64.

8. Taupp M, Mewis K, Hallam SJ. The art and design of functional metagenomic screens. Curr Opin Biotechnol [Internet]. 2011

Jun;22(3):465–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166911000371

9. Berini F, Casciello C, Marcone GL, Marinelli F. Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett [Internet]. 2017 Nov 15;364(21). Available from: https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnx211/4329276

10. Decastro M-E, Escuder-Rodríguez J-J, Cerdán M-E, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Heat-loving β-galactosidases from cultured and uncultured microorganisms. Curr Protein Pept Sci. 2018;19(12).

11. Ferrer M, Martínez-Martínez M, Bargiela R, Streit WR, Golyshina O V, Golyshin PN. Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microb Biotechnol [Internet]. 2016 Jan;9(1):22–34. Available from: http://doi.wiley.com/10.1111/1751-7915.12309

12. Lorenz P, Liebeton K, Niehaus F, Eck J. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Curr Opin Biotechnol [Internet]. 2002 Dec;13(6):572–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12482516

13. Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC. Current and future resources for functional metagenomics. Front Microbiol [Internet]. 2015 Oct 29;6(October):1196. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.01196/abstract

14. Ekkers DM, Cretoiu MS, Kielak AM, van Elsas JD. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl Microbiol Biotechnol [Internet]. 2012 Feb 22;93(3):1005–20. Available from:

http://link.springer.com/10.1007/s00253-011-3804-3

15. van der Helm E, Genee HJ, Sommer MOA. The evolving interface between synthetic biology and functional metagenomics. Nat Chem Biol [Internet]. 2018 Aug 16;14(8):752–9. Available from: http://dx.doi.org/10.1038/s41589-018-0100-x

16. Uchiyama T, Miyazaki K. Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr Opin Biotechnol [Internet]. 2009 Dec 1 [cited 2018 May 3];20(6):616–22. Available from: https://www.sciencedirect.com/science/article/pii/S0958166909001244

17. Craig JW, Chang F-Y, Kim JH, Obiajulu SC, Brady SF. Expanding Small-Molecule Functional Metagenomics through Parallel Screening of Broad-Host-Range Cosmid Environmental DNA Libraries in Diverse Proteobacteria. Appl Environ Microbiol [Internet]. 2010 Mar 1;76(5):1633–41. Available from: https://aem.asm.org/content/76/5/1633

18. Lorenz P, Eck J. Metagenomics and industrial applications. Nat Rev Microbiol. 2005;3(June):510–6.

19. Gabor EM, Alkema WBL, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ Microbiol [Internet]. 2004 Sep;6(9):879–86. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2004.00640.x

20. Choi SL, Rha E, Lee SJ, Kim H, Kwon K, Jeong YS, et al. Toward a generalized and high-throughput enzyme screening system based on artificial genetic circuits. ACS Synth Biol. 2014;3(3):163–71.

21. Goddard J-P, Reymond J-L. Recent advances in enzyme assays. Trends Biotechnol [Internet]. 2004 Jul;22(7):363–70. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779904001106

22. Demirjian DC, Morís-varas F, Cassidy CS. Enzymes from extremophiles. Curr Opin Chem Biol. 2001;5:144–51.

23. Haki GD, Rakshit SK. Developments in industrially important thermostable enzymes: A review. Bioresour Technol. 2003;89:17–34.

24. Bruins ME, Janssen AEM, Boom RM. Thermozymes and Their Applications. Appl Biochem Biotechnol. 2001;90:155–86.

25. Kumar S, Dangi AK, Shukla P, Baishya D, Khare SK. Thermozymes: Adaptive strategies and tools for their biotechnological applications. Bioresour Technol [Internet]. 2019 Apr;278(November 2018):372–82. Available from: https://doi.org/10.1016/j.biortech.2019.01.088

26. Chen G-Q, Jiang X-R. Next generation industrial biotechnology based on extremophilic bacteria. Curr Opin Biotechnol [Internet]. 2018 Apr;50:94–100. Available from: https://doi.org/10.1016/j.copbio.2017.11.016

27. Escuder-Rodríguez J-J, DeCastro M-E, Rodríguez-Belmonte E, Becerra M, María-Isabel G-S. Hot Springs Thermophilic Microbiomes: Biodiversity and Biotechnological Applications. In: Yadav AN, Rastegari AA, Yadav N, editors. Microbiomes of Extreme Environments: Biodiversity and Biotechnological Applications [Internet]. 1st ed. First edition. | Boca Raton : CRC Press, 2021.: CRC Press; 2021. p. 87–105. Available from: https://www.taylorfrancis.com/books/9781000734911

28. Prayogo FA, Budiharjo A, Kusumaningrum HP, Wijanarka W, Suprihadi A, Nurhayati N. Metagenomic applications in exploration and development of novel enzymes from nature: a review. J Genet Eng Biotechnol [Internet]. 2020 Dec 4;18(1):39. Available from: https://jgeb.springeropen.com/articles/10.1186/s43141-020-00043-9

29. Escuder-Rodríguez J-J, López-López O, Becerra M, Cerdán M-E, González-Siso M-I. Extremophilic esterases for bioprocessing of lignocellulosic feedstocks. Extremophilic

Enzymatic Processing of Lignocellulosic Feedstocks to Bioenergy. 2017.

30. Escuder-Rodríguez J-J, DeCastro M-E, Cerdán M-E, Rodríguez-Belmonte E, Becerra M, González-Siso M-I. Cellulases from Thermophiles Found by Metagenomics. Microorganisms [Internet]. 2018 Jul 10;6(3):66. Available from: http://www.mdpi.com/2076-2607/6/3/66

31. Juturu V, Wu JC. Insight into microbial hemicellulases other than xylanases: A review. J Chem Technol Biotechnol. 2013;88(3):353–63.

32. Barcelos MCS, Ramos CL, Kuddus M, Rodriguez-Couto S, Srivastava N, Ramteke PW, et al. Enzymatic potential for the valorization of agro-industrial by-products. Biotechnol Lett [Internet]. 2020 Oct 9;42(10):1799–827. Available from: https://doi.org/10.1007/s10529-020-02957-3

33. Elleuche S, Schröder C, Sahm K, Antranikian G. Extremozymes—biocatalysts with unique properties from extremophilic microorganisms. Curr Opin Biotechnol [Internet]. 2014 Oct;29(1):116–23. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166914000755

34. Wohlgemuth R, Littlechild J, Monti D, Schnorr K, van Rossum T, Siebers B, et al. Discovering novel hydrolases from hot environments. Biotechnol Adv [Internet]. 2018 Dec;36(8):2077–100. Available from: https://doi.org/10.1016/j.biotechadv.2018.09.004

35. Dadwal A, Sharma S, Satyanarayana T. Progress in Ameliorating Beneficial Characteristics of Microbial Cellulases by Genetic Engineering Approaches for Cellulose Saccharification. Front Microbiol [Internet]. 2020 Jun 24;11(June):1–18. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.01387/full

36. Sadhu S, Maiti TK. Cellulase Production by Bacteria : A Review.

Br Microbiol Res J [Internet]. 2013;3(3):235–58. Available from: www.sciencedomain.org

37. Urbieta MS, Donati ER, Chan K-G, Shahar S, Sin LL, Goh KM. Thermophiles in the genomic era: Biodiversity, science, and applications. Biotechnol Adv [Internet]. 2015 Nov;33(6):633–47. Available from: http://dx.doi.org/10.1016/j.biotechadv.2015.04.007

38. Bertani G. Studies on lysogenesis. J Bacteriol [Internet]. 1951 Sep;62(3):293–300. Available from: https://journals.asm.org/doi/10.1128/jb.62.3.293-300.1951

39. Pandey S, Singh S, Yadav AN, Nain L, Saxena AK. Phylogenetic diversity and characterization of novel and efficient cellulase producing bacterial isolates from various extreme environments. Biosci Biotechnol Biochem [Internet]. 2013;77(7):1474–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23832366

40. Farrow MF, Arnold FH. High Throughput Screening of Fungal Endoglucanase Activity in *Escherichia coli*. J Vis Exp. 2011;21(54):1–4.

41. Zhang L, Fan Y, Zheng H, Du F, Zhang KQ, Huang X, et al. Isolation and characterization of a novel endoglucanase from a *Bursaphelenchus xylophilus* metagenomic library. PLoS One. 2013;8(12):1–8.

42. Berlemont R, Delsaute M, Pipers D, D'Amico S, Feller G, Galleni M, et al. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. ISME J [Internet]. 2009;3(9):1070–81. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19458657&retmode=ref&cmd=prlinks%5Cnpapers2://publication/doi/10.1038/ismej.2009.48

43. Ilmberger N, Streit WR. Metagenomics [Internet]. Streit WR, Daniel R, editors. Metagenomics: Methods and Protocols. Totowa, NJ: Humana Press; 2010. 177–188 p. (Methods in Molecular Biology; vol. 668). Available from:

http://link.springer.com/10.1007/978-1-60761-823-2

44. Berlemont R, Pipers D, Delsaute M, Angiono F, Feller G, Galleni M, et al. Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements. Rev Argent Microbiol. 2011;43(2):94–103.

45. Kim Y-K, Lee S-C, Cho Y-Y, Oh H-J, Ko YH. Isolation of Cellulolytic *Bacillus subtilis* Strains from Agricultural Environments. ISRN Microbiol. 2012;2012:1–9.

46. Kasana RC, Salwan R, Dhar H, Dutt S, Gulati A. A rapid and easy method for the detection of microbial cellulases on agar plates using Gram's iodine. Curr Microbiol. 2008;57(5):503–7.

47. Knapik K, Becerra M, González-Siso M-I. Microbial diversity analysis and screening for novel xylanase enzymes from the sediment of the Lobios Hot Spring in Spain. Sci Rep [Internet]. 2019 Dec 1;9(1):11195. Available from: http://www.nature.com/articles/s41598-019-47637-z

48. Anderson I, Abt B, Lykidis A, Klenk H-P, Kyrpides N, Ivanova N. Genomics of Aerobic Cellulose Utilization Systems in Actinobacteria. Ben-Jacob E, editor. PLoS One [Internet]. 2012 Jun 18;7(6):e39331. Available from: https://dx.plos.org/10.1371/journal.pone.0039331

49. Blank S, Schröder C, Schirrmacher G, Reisinger C AG. Biochemical Characterization of a Recombinant xylanase from *Thermus brockianus*, Suitable for Biofuel Production. JSM Biotechnol Bioeng. 2014;2(1):1027.

50. Pinheiro GL, Correa RF, Cunha RS, Cardoso AM, Chaia C, Clementino MM, et al. Isolation of aerobic cultivable cellulolytic bacteria from different regions of the gastrointestinal tract of giant land snail *Achatina fulica*. Front Microbiol [Internet]. 2015 Aug 20;6(AUG):1–15. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00860/abstract

51. Ohlhoff CW, Kirby BM, Van Zyl L, Mutepfa DLR, Casanueva A,

Huddy RJ, et al. An unusual feruloyl esterase belonging to family VIII esterases and displaying a broad substrate range. J Mol Catal B Enzym [Internet]. 2015 Aug;118:79–88. Available from: http://dx.doi.org/10.1016/j.molcatb.2015.04.010

52. Hårdeman F, Sjöling S. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. FEMS Microbiol Ecol. 2007;59(2):524–34.

53. Leis B, Angelov A, Mientus M, Li H, Pham VTT, Lauinger B, et al. Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. Front Microbiol [Internet]. 2015 Apr 8;6(April):275. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25904908

54. Selvin J, Kennedy J, Lejon DPH, Kiran S, Dobson ADW. Isolation identification and biochemical characterization of a novel halo-tolerant lipase from the metagenome of the marine sponge *Haliclona simulans*. Microb Cell Fact [Internet]. 2012;11(1):72. Available from: Microbial Cell Factories

55. Sang SL, Li G, Hu XP, Liu YH. Molecular Cloning, Overexpression and Characterization of a Novel Feruloyl Esterase from a Soil Metagenomic Library. J Mol Microbiol Biotechnol [Internet]. 2011;20(4):196–203. Available from: https://www.karger.com/Article/FullText/329833

56. Andersen A, Svendsen A, Vind J, Lassen S., Hjort C, Borch K, et al. Studies on ferulic acid esterase activity in fungal lipases and cutinases. Colloids Surfaces B Biointerfaces [Internet]. 2002 Sep;26(1–2):47–55. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0927776501003125

57. Nyyssönen M, Tran HM, Karaoz U, Weihe C, Hadi MZ, Martiny JBH, et al. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. Front Microbiol [Internet]. 2013;4(SEP):1–14. Available from:

http://journal.frontiersin.org/article/10.3389/fmicb.2013.00282/abstract

58. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol [Internet]. 2018 Aug 6;36(7):566–9. Available from: http://www.nature.com/articles/nbt.4163

59. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res [Internet]. 2017 May;27(5):824–34. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.213959.116

60. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ [Internet]. 2019 Jul 26;7(7):e7359. Available from: https://peerj.com/articles/7359

61. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics [Internet]. 2014 Jul 15;30(14):2068–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu153

62. NCBI_Resource_Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D7–19. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1290

63. Gasteiger E. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res [Internet]. 2003 Jul 1;31(13):3784–8. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg563

64. BIOSOFT P. Net Primer [Internet]. Available from: http://www.premierbiosoft.com/netprimer/index.html

65. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–303.

66. SchrödingerLCC. The PyMOL Molecular Graphics System, Version 1.2r3pre [Internet]. Available from: https://pymol.org/2/#page-top

67. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook [Internet]. Totowa, NJ: Humana Press; 2005. p. 571–608. Available from: http://link.springer.com/10.1385/1592598900

68. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol [Internet]. 2011 Jan 11;7(1):539. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21988835

69. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D506–15. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

70. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res [Internet]. 2004 Jun;14(6):1188–90. Available from: ftp://ftp.ncbi.nih.gov/genomes/Bacteria

71. Fuciños P, Abadín CM, Sanromán A, Longo MA, Pastrana L, Rúa ML. Identification of extracellular lipases/esterases produced by *Thermus thermophilus* HB27: Partial purification and preliminary biochemical characterisation. J Biotechnol [Internet]. 2005 May;117(3):233–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168165605000581

72. Miguel-Ruano V, Rivera I, Rajkovic J, Knapik K, Torrado A,

Otero JM, et al. Biochemical and Structural Characterization of a novel thermophilic esterase EstD11 provide catalytic insights for the HSL family. Comput Struct Biotechnol J [Internet]. 2021;19:1214–32. Available from: https://doi.org/10.1016/j.csbj.2021.01.047

73. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics [Internet]. 2011 Feb 1;27(3):343–50. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq662

74. Akoh CC, Lee GC, Liaw YC, Huang TH, Shaw JF. GDSL family of serine esterases/lipases. Prog Lipid Res. 2004;43(6):534–52.

75. Kim MH, Hao OJ, Wang NS. *Acinetobacter* isolates from different activated sludge processes: characteristics and neural network identification. FEMS Microbiol Ecol [Internet]. 1997;23(3):217–27. Available from: https://academic.oup.com/femsec/article-lookup/doi/10.1111/j.1574-6941.1997.tb00404.x

76. Pulami D, Schauss T, Eisenberg T, Blom J, Schwengers O, Bender JK, et al. *Acinetobacter stercoris* sp. nov. isolated from output source of a mesophilic german biogas plant with anaerobic operating conditions. Antonie Van Leeuwenhoek [Internet]. 2021 Mar 16;114(3):235–51. Available from: http://link.springer.com/10.1007/s10482-021-01517-7

77. Murugapiran SK, Huntemann M, Wei C-L, Han J, Detter JC, Han CS, et al. Whole Genome Sequencing of *Thermus oshimai* JL-2 and *Thermus thermophilus* JL-18, Incomplete Denitrifiers from the United States Great Basin. Genome Announc [Internet]. 2013 Feb 28;1(1):2–3. Available from: https://journals.asm.org/doi/10.1128/genomeA.00106-12

78. Alalouf O, Balazs Y, Volkinshtein M, Grimpel Y, Shoham G, Shoham Y. A New Family of Carbohydrate Esterases Is Represented by a GDSL Hydrolase/Acetylxylan Esterase from *Geobacillus stearothermophilus*. J Biol Chem [Internet]. 2011 Dec;286(49):41993–2001. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S0021925820870902

79. Hall J, Hazlewood GP, Barker PJ, Gilbert HJ. Conserved reiterated domains in *Clostridium thermocellum* endoglucanases are not essential for catalytic activity. Gene. 1988;69(1):29–38.

80. Montanier C, Money VA, Pires VMR, Flint JE, Pinheiro BA, Goyal A, et al. The active site of a carbohydrate esterase displays divergent catalytic and noncatalytic binding functions. PLoS Biol. 2009;7(3):0687–97.

81. Yang Z, Zhang Y, Shen T, Xie Y, Mao Y, Ji C. Cloning, expression and biochemical characterization of a novel, moderately thermostable GDSL family esterase from *Geobacillus thermodenitrificans* T2. J Biosci Bioeng [Internet]. 2013 Feb;115(2):133–7. Available from: http://dx.doi.org/10.1016/j.jbiosc.2012.08.016

82. Soni S, Sathe SS, Odaneth AA, Lali AM, Chandrayan SK. SGNH hydrolase-type esterase domain containing Cbes-AcXE2: a novel and thermostable acetyl xylan esterase from *Caldicellulosiruptor bescii*. Extremophiles [Internet]. 2017 Jul 25;21(4):687–97. Available from: http://link.springer.com/10.1007/s00792-017-0934-2

83. Jo E, Kim J, Lee A, Moon K, Cha J. Identification and Characterization of a Novel Thermostable GDSL-Type Lipase from *Geobacillus thermocatenulatus*. J Microbiol Biotechnol [Internet]. 2021 Mar 28;31(3):483–91. Available from: http://www.jmb.or.kr/journal/view.html?doi=10.4014/jmb.2012.12036

84. Chahiniana H, Sarda L. Distinction Between Esterases and Lipases: Comparative Biochemical Properties of Sequence-Related Carboxylesterases. Protein Pept Lett [Internet]. 2009 Oct 1;16(10):1149–61. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8665&volume=16&issue=10&spage=1149

85. Fojan P, Jonson PH, Petersen MTN, Petersen SB. What distinguishes an esterase from a lipase: A novel structural approach. Biochimie. 2000;82(11):1033–41.

CHAPTER 3

Microbial community inhabiting the Muiño da Veiga hot spring described by sequence-based metagenomics and in-silico analysis

Abstract

The Muiño da Veiga hot spring metagenome was studied by shotgun sequencing using the Hi-seq Illumina Platform in order to characterize the microbial community composition and diversity, as well as to explore their metabolic connections and potential sources for thermozymes. The assessment of the sequencing data revealed a diverse microbial community dominated by Bacteria domain members, and in particular from the early-branching Aquificales group. The most abundant genus was *Sulfurihydrogenibium*, known for its implication in sulfur cycling and for forming mats that enable novel niches in the hot spring ecosystem. Most of the main metabolic pathways for cycling of elements were present in the metagenome, but some of the routes were attributed to less abundant members of the microbial community, reinforcing the idea that the rare biosphere may play important roles in the overall network of interactions present in an ecosystem, or act as genetic reservoirs for environments with fluctuating abiotic conditions. Our results also revealed the potential of the hot spring for bioprospecting studies to retrieve thermozymes of biotechnological interest.

Introduction

The very firsts studies on microbial identification were reliant on the ability to grow microorganisms to stablish pure cultures, that were further tested for defining phenotypical traits that allowed classification (1). Nonetheless, the notion of a great discrepancy between the metabolically active microbial community present in an environment and the total counts of colonies in laboratory media, named the great plate count anomaly (2) was known and

hindered our ability to fully understand microbial diversity and microbial communities, highlighting the need for molecular methods for classification (3). Although cloning of environmental DNA was deemed as a solution to this problem, not until the development of high-throughput sequencing was possible to meet the technical requirements to assess the metagenome (1,4). A wide range of technologies categorized as Next Generation Sequencing (NGS) are available, and have experienced a number of improvements overcoming some of the technical bottlenecks associated to obtaining sequencing data at a microbial community-level scale (4). Depending on the strategy employed, there is a distinction between the cost-effective targeted sequencing of biomarker genes (in most studies the variable regions of the 16S rRNA gene), focused on answering the "Who are they?" biodiversity question, and the shotgun metagenomic sequencing strategy that can answer the "What are they doing?" ecological question (5). To analyse the enormous amount of data generated from the NGS platforms, an extensive and ever-growing number of bioinformatic tools have been developed (6). Although metagenomics was first envisioned to explore the diversity of the soil microorganisms (7), it rapidly expanded to cover water environments, and in recent years the role of the human-associated microbiome in regards to health and disease has become the new focus of research and has driven the development of many bioinformatic tools (8).

Due to their biotechnological potential and unique conditions that can resemble primitive Earth environments and even potential extra-terrestrial ecosystems, hot springs are expected to host microbial communities that hold high value to both basic and applied science (8). In the present work, we conducted a shotgun metagenomic sequencing study of the Muiño da Veiga hot spring metagenome, in order to characterize the microbial community, its biotechnological potential and the main metabolic pathways involved in this thermophilic habitat.

Materials and Methods

The DNA from the environmental sample described in Chapter 1 (Muiño da Veiga hot spring, Ourense, Spain) was sequenced with an Illumina platform (Hi-seq) by the sequencing services provided by Health In Code (A Coruña, Spain). The sequencing resulted in two files in FastQ format (each containing a pair-ended read with the DNA sequence and its quality) with a mean length of 100bp for each sequence and a total of 104 million sequences per file. These two files were subjected to two different bioinformatic pipelines, with some common early steps. The common steps comprised basic quality filtering of the FastQ format raw sequence files, and pair-ended read merging. Briefly, the first pipeline operated with the merged reads directly to avoid changes in read counts, to estimate the amount of gene functions with a potential industrial application in the metagenome by alignment to two different databases of annotated gene functions (a custom SEED subsystems database and the non-redundant NCBI database). The second pipeline also used merged reads, but this time they were uploaded to a web-based metagenome analysis server (MG-RAST), to align the sequences to a database of annotated taxonomic features, with the purpose of understanding the composition of the microbial community. This analysis server also allowed to study KEGG pathways to analyse the possible metabolic interactions of the microbial community by alignment to two different hierarchical function-based databases (KEGG Orthology and SEED Subsystems).

Quality filtering

Sequence data was preprocessed with the prinseq-lite perl script (9) in order to assess quality control. Sequences were first trimmed using a quality score threshold of 25 by both right and left ends of the sequences, with around 303 thousand sequences trimmed only by the left side (none were trimmed by the right side). A minimum length of 60bp was required in order to pass the next filter, removing a total of 2.6 million sequences. The following step

involved the removal of sequences containing at least an unknown (N) basepair, resulting in the filtering of 24.5 million sequences. Sequence complexity was assessed by the entropy method (using a threshold of 70) and 1 million sequences were filtered. Lastly, sequences duplicates (including reverse complement duplicates) were also removed from the dataset, with a total of 7.8 million sequences removed. After the quality processing, a total of 6.2 million sequences from the first file and 5.4 million sequences for the second file were left as singletons (its pair-ended read could not pass the filter). These sequences were not used in the next steps of the analysis. A total of 79.6 million sequences passed the quality filters (76.52%) and were used for the bioinformatics analysis.

Read merging

Sequences corresponding to a pair-ended read were merged using the PEAR software (10) running on default settings. PEAR successfully merged 54 million sequences (67.85%). Sequences that could not be merged were discarded as they would alter the number of reads generating each count, thus providing false data when aligned to a database if they were used in conjunction of merged sequences.

Alignment to the SEED subsystems database

A database from the SEED subsystems was provided by Dr. Genivaldo Gueiros (Department of Computer Sciences, San Diego State University) obtained by clustering the subsystems with an identity of 98% as published in a previous work (11). This reduced database was used to align the merged pair-end reads using the open source algorithm DIAMOND (12) with a setting for e-value of 0.00001. The results of the alignment were further processed to only keep the best hits per sequence (those with a unique identifier and the highest e-value, but also keeping those with the same identifier and the same e-value) using a custom script.

Alignment to the NCBI nr database

The nr database from NCBI was retrieved from its server (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz) on the 28th of September of 2016 (13). This database was used to align the merged pair-end reads using DIAMOND (12) with a setting for e-value of 0.00001. As with the SEED subsystems database, only the results of the alignment with the highest e-value were kept for each sequence.

Sequence analysis on the MG-RAST web-based server

Sequences were uploaded to the MG-RAST (14) web-based server, that automatically applies its own bioinformatic analysis pipeline to give an overview of the metagenome. The pipeline options for the upload were as follows: dereplication "yes", screening "*H. sapiens*, NCBI v36", dynamic trimming "yes", minimum quality "15", and maximum low quality basepairs "5".

The step by step process for the automated upload pipeline is briefly described as follows:

1. Compute sequence statistics using DRISEE (15) and Jellyfish (16) and determine sequence information. This first step uses two sequence analysis algorithms to the detection of errors in the raw metagenomic sequence data.
2. Trimming of adapters using the Skewer software (17). In this step adapter sequences were detected and removed.
3. Denoising and normalization. Sequences were subjected to an additional process of quality control using fastq-mcf (18), using parameters for sequence length, number of ambiguous bases and base quality (this step had minimal impact due to previous sequence quality filtering in our pipeline).

4. Removal of sequencing artifacts. MG-RAST executes its own script to remove redundant technical replicate sequences by binning reads with identical first 50 base-pairs.
5. DNA contamination removal. Using Bowtie2 (19) the metagenomic sequences were aligned to the "*H. sapiens*, NCBI v36" reference genome in order to remove possible contamination of human DNA due to either source material contamination and/or sample manipulation contamination.
6. RNA feature identification (rRNA gene calling), clustering and similarity search. Potential ribosomal RNA sequences were identified with a 70% identity cut-off using SortMeRNA (20), and then clustered at 97% identity as same species using CD-HIT (21). The similarity search was performed using the BLAT algorithm (22) against MG-RAST reduced M5RNA database.
7. Identify putative protein-coding features (gene calling). Most likely open reading frames were predicted using FragGeneScan (23), translating nucleotide sequences to amino acid sequences. The putative protein features that overlap with rRNA features were then filtered using and algorithm from MG-RAST.
8. Amino acid sequence clustering and protein similarity search. Using CD-HIT amino acid sequences were clustered at 90% identity to reduce computational cost in the next step of similarity search. This was performed using BLAT against the M5NR protein database, using either clustered or singleton protein features from the previous step.
9. Mapping, indexing and expanding of annotations. Both rRNA and protein similarity annotations were expanded into various technical namespaces. The cluster annotations of rRNA and protein features were mapped back to the original sequences and indexed for fast access. The similarities were also expanded into various technical namespaces and indexed.

10. Abundance profiling. The observed abundance per feature (M5NR hit), per lowest common ancestor and per data source (e.g., RefSeq, Subsystems, etc.) were computed from the expanded similarities. This effectively changed the data structure from one line per similarity to one line per M5NR hit, lowest common ancestor or data source, respectively.
11. Features that were predicted by FragGeneScan but with no similarity hits with M5NR are extracted into a "darkmatter" sequence file.
12. The abundance profile data is computed into the Cassandra database for handling the data across the data centres and the cloud. The abundance profiles were computed from the expanded similarities and loaded into the database.
13. An overview of all the statistics is provided as a summary.

After the upload, a more in-depth analysis was performed in the analysis tab of the web server, using three different databases: RefSeq (24) for taxonomic assessment of the reads, and KEGG Ortology (KO) (25) and SEED Subsystems (26) for hierarchical functional assignments and KEGG mapping. The analysis parameters were as follows: e-value "5", %-ident "60", length "15", min. abundance "1", method "representative hit". The taxonomical categories at the order level were visualized using the KRONA web utility (27).

Results and discussion

Assessment of the microbial community potential for retrieving enzymes of biotechnological interest

Alignment against the SEED subsystems database

We obtained a total of 30,132,990 counts aligning to the reduced SEED subsystems database. We searched for enzymes that fall in the category of hydrolases (with EC number 3.-.-.-) and found 3,598,086 counts (11.94%). We further examined these counts in order to find

enzymes that have a potential interest as biotechnological products for various industrial applications. Such enzymes and their relative abundances are shown in Figure 1.

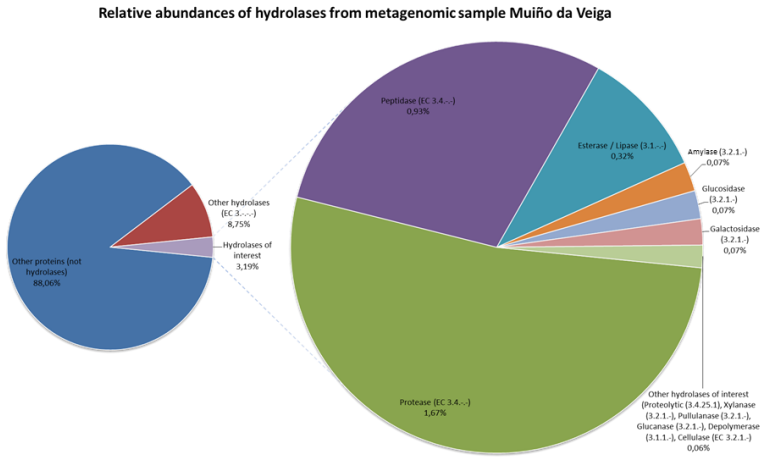


Figure 1. Relative abundances of hydrolases from the metagenomic sample obtained in Muiño da Veiga. Abundance was measured as number of counts obtained by the DIAMOND aligner using a SEED subsystems database clustered using an identity of 98% (in order to reduce the size of the original database). Out of 30,132,990 counts, 88.06% were assigned to a subsystem function different from hydrolases. 8.75% were found to be hydrolases with no direct biotechnological applications in industry. The remaining 3.19% were found to be hydrolases with possible interest in industrial processes: proteases, peptidases, esterases and lipases, amylases, glucosidases, galactosidases, proteolytic enzymes, xylanases, pullulanases, glucanases, depolymerases and cellulases. Relative abundances of each enzyme family are shown along with their EC number.

Alignment to the non-redundant database

A total of 43,613,726 counts aligning to the non-redundant (NR) database were obtained. We performed a keyword-based search for enzymes that fall in the category of hydrolases with possible industrial interest, and 1,276,574 counts (2,93%) were selected. Figure 2 shows those enzymes and their relative abundances.

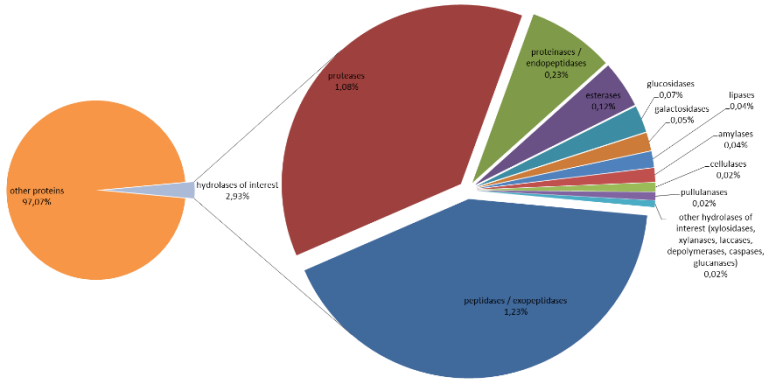


Figure 2. Relative abundances of hydrolases from the metagenomic sample obtained in Muiño da Veiga. Abundance was measured as number of counts obtained by the DIAMOND aligner using the NCBI NR database. The total amount of counts was 43,613,726. Of those, 1,276,574 counts (2.93%) were annotated as hydrolases with possible interest in industrial processes: proteases and peptidases (exopeptidases, endopeptidases, proteinases), esterases and lipases, amylases, cellulases, glucosidases, galactosidases, pullulanases, xylosidases, xylanases, laccases, depolymerases, caspases and glucanases. Relative abundances of each enzyme class are given as a percentage of the total number of counts obtained.

Our results showed that hydrolases were slightly over a tenth of the total proteins predicted by the bioinformatic analysis. Proteases and peptidases were the most abundant hydrolases with potential industrial application in both analyses. Although the number of enzymes with potential biotechnological application was relatively low compared to all other proteins predicted by alignment to either database, the presence of such enzymes in the metagenome was deemed interesting in order to conduct a bioprospecting study using the metagenomic data collected. Further studies in this regard are conducted in Chapter 4 of this work.

MG-RAST server metagenome overview and in-depth analysis

A summary of the metagenome statistics, as reported by the automated upload analysis of MG-RAST is presented in Table 1. This preliminary data shows that even with the previous quality filtering of sequences, more than a third of the sequences had to be further removed due to the quality criteria from the web-based analysis tool. A small (6.64% of total sequences) portion of the sequences could not be associated to any feature deposited in the databases, and of the predicted features, 21.12% (19.72% of total sequences) were annotated as proteins of unknown function.

Table 1. Statistics for the metagenome from Muiño da Veiga hot spring.

Sequences (uploaded)	27,113,937	sequences
Basepairs (uploaded)	3,968,584,153	bp
Mean sequence length (uploaded)	146 ± 24	bp
Mean GC content (uploaded)	44 ± 13	%
QC failed	9,290,572	sequences
	34.2649317	% of total
Artificial duplicate reads	4,807,111	sequences
	17.7292991	% of total
Sequences (post QC)	17,823,365	sequences
	65.7350683	% of total
Basepairs (post QC)	2,658,836,275	bp
	66.9970995	% of total
Mean sequence length (post QC)	149 ± 24	bp
Mean GC content (post QC)	46 ± 14	%
Predicted features	16,639,799	sequences
	93.3594694	% of total post QC
rRNA genes	72,722	sequences
	0.40801498	% of total post QC
Predicted protein with known function	13,052,391	sequences
	73.2319122	% of total post QC
	78.4407973	% of predicted features
Predicted protein with unknown function	3,514,686	sequences
	19.7195423	% of total post QC
	21.1221662	% of predicted features
Unknown features	1,183,566	sequences
	6.64053056	% of total post QC

Sequencing depth was assessed by plotting the number of species (unique OTUs) found against the total number or reads needed to achieve them, as represented in Figure 3. Generally, good sequencing depth is achieved when adding more reads to the analysis have little effect on the diversity of species found, meaning that further sampling will have little effect in the diversity of species.

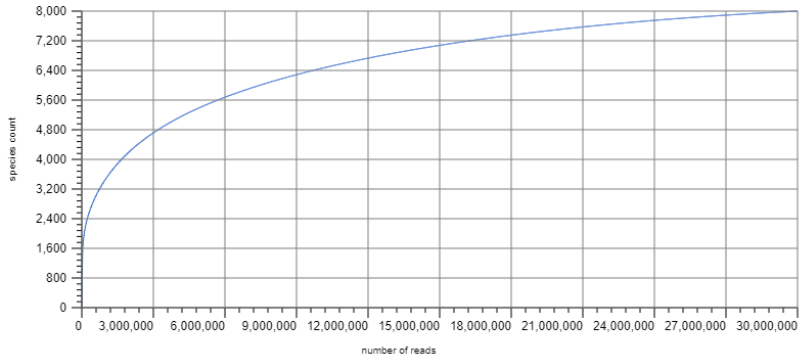


Figure 3. Rarefaction plot. Assessment of the sequencing depth by plotting species annotation counts against the number of reads analyzed.

Metabolism and interactions of the microbial community inferred from the metagenomic analysis

An in-depth analysis conducted in the MG-RAST server was employed to describe the microbial community composition in the hot spring and infer metabolic pathways that may explain the complex network of interactions between microorganisms inhabiting it. For the RefSeq database, a total of 15,369,172 hits were obtained for the taxonomical assessment of reads, whereas KO returned 7,135,862 hits and Subsystems 6,223,410 both for functional assessment of reads. Figure 4 shows the most abundant genera of the hot spring (that covered more than 50% of the total reads) found by alignment to the RefSeq database, and the

percentages of each taxonomical category at the order level represented used the KRONA web visualization tool.

Alignment to the KO and Subsystems databases allowed the function assignment of sequences, in order to reconstruct KEGG pathways and predict which metabolic routes were involved in the thermophilic microbial community, making possible the assessment of possible microbial interactions within the hot spring ecosystem. It is worth noting that this information ought to be supplemented with additional data published in literature on the most abundant genera, as sequence information alone is not considered enough on its own to fully describe such interactions and fully contextualize the complex metabolic network of a extremophilic microbial community (28). Moreover, the same need has been stablished for the rare microorganisms of a microbial community, as metagenomics alone is deemed to not always be sufficient to assess the full metabolic potential of the so-called rare biosphere (29). In this regard, an extensive search for the dominant and rare genera data on metabolism and core genomes and pangenomes was conducted to further understand the complex metabolic interactions in the Muiño da Veiga hot spring ecosystem, keeping in mind that these inferences are inherently biased (30), but nevertheless offer a framework that can be combined with functional assessment of reads for putative descriptions. The breakdown of reads assigned to functional categories in each database is represented in Figure 5.

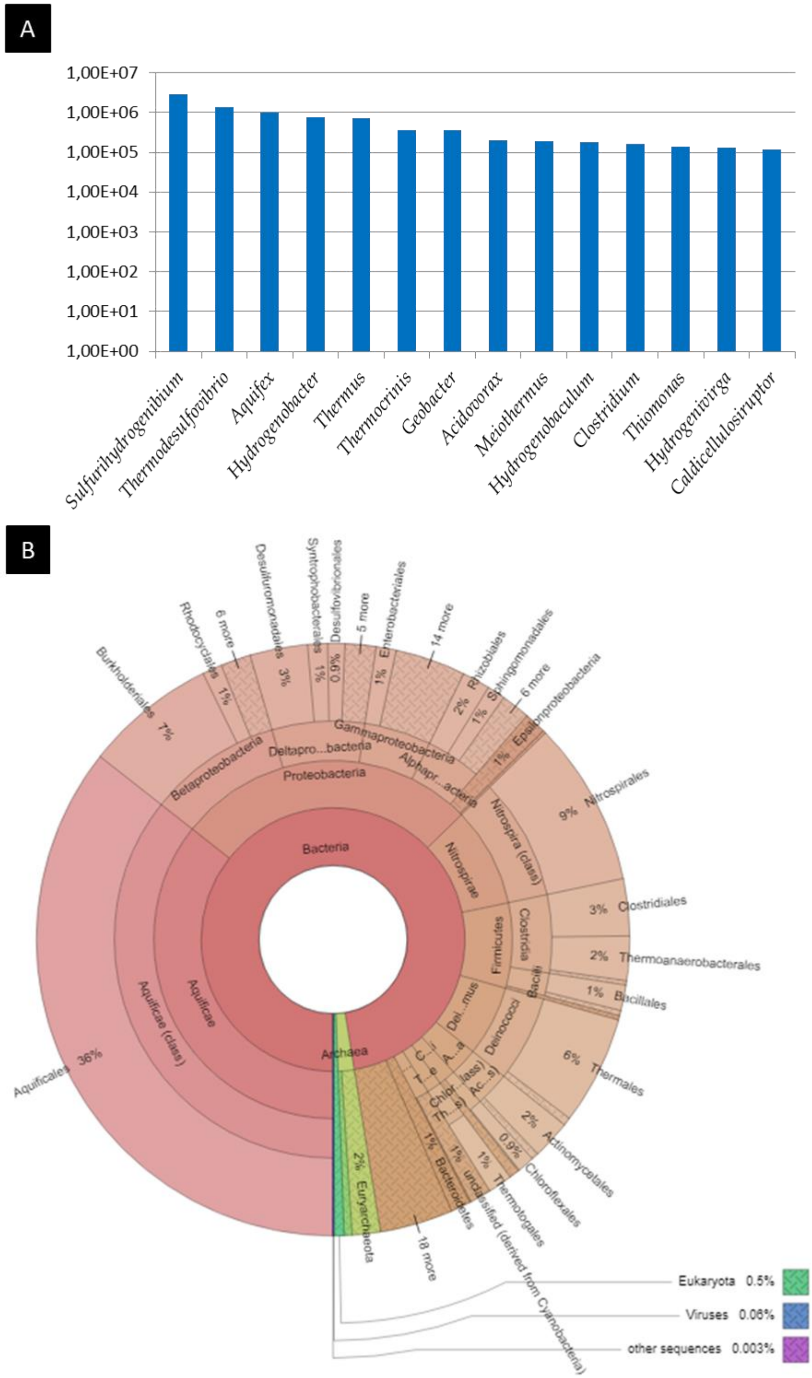

A
1,00E+07
1,00E+06
1,00E+05
1,00E+04
1,00E+03
1,00E+02
1,00E+01
1,00E+00
Sulfurihydrogenibium
Thermodesulfovibrio
Aquifex
Hydrogenobacter
Thermus
Thermocrinis
Geobacter
Acidovorax
Meiothermus
Hydrogenobaculum
Clostridium
Thiomonas
Hydrogenivirga
Caldicellulosiruptor
B
Bacteria
Archaea
Proteobacteria
Aquificae
Aquificae (class)
Aquificales 36%
Burkholderiales 7%
Betaproteobacteria
Deltapro...bacteria
Gammaproteobacteria
Alphapr...acteria
Nitrospirae
Nitrospira (class)
9% Nitrospirales
Firmicutes
Clostridia
3% Clostridiales
2% Thermoanaerobacterales
1% Bacillales
6% Thermales
2% Actinomycetales
0.9% Chloroflexales
1% Thermotogales
1% unclassified (derived from Cyanobacteria)
1% Bacteroidetes
18 more
2% Euryarchaeota
Eukaryota 0.5%
Viruses 0.06%
other sequences 0.003%

Figure 4A. Rank of the top 14 most abundant genus in the metagenome, accounting for 56.2% of the total reads. The Y-axis represents in logarithm scale the number of reads for each genus. Genus outside this ranking accounted individually for less than 0.75% of the total reads, whereas the most abundant genus (*Sulfurihydrogenibium*) accounted for 19.3% of the total reads by itself, and the 8 most abundant genus accounted for more than half of the total reads. All the top-ranking genus belonged to the Bacteria domain. Figure 4B. Krona pie chart representation of the taxonomical read assignments at the order level. Aquificales was the most abundant order with 36% of the reads, followed by Nitrospirales with 9% of the reads and Burkholderiales with 7%. Bacteria was the dominant domain with 97% of the reads, Archaea represented only 2% of the reads and the remaining reads belonged to Eukaryota, Viruses and unassigned sequences.

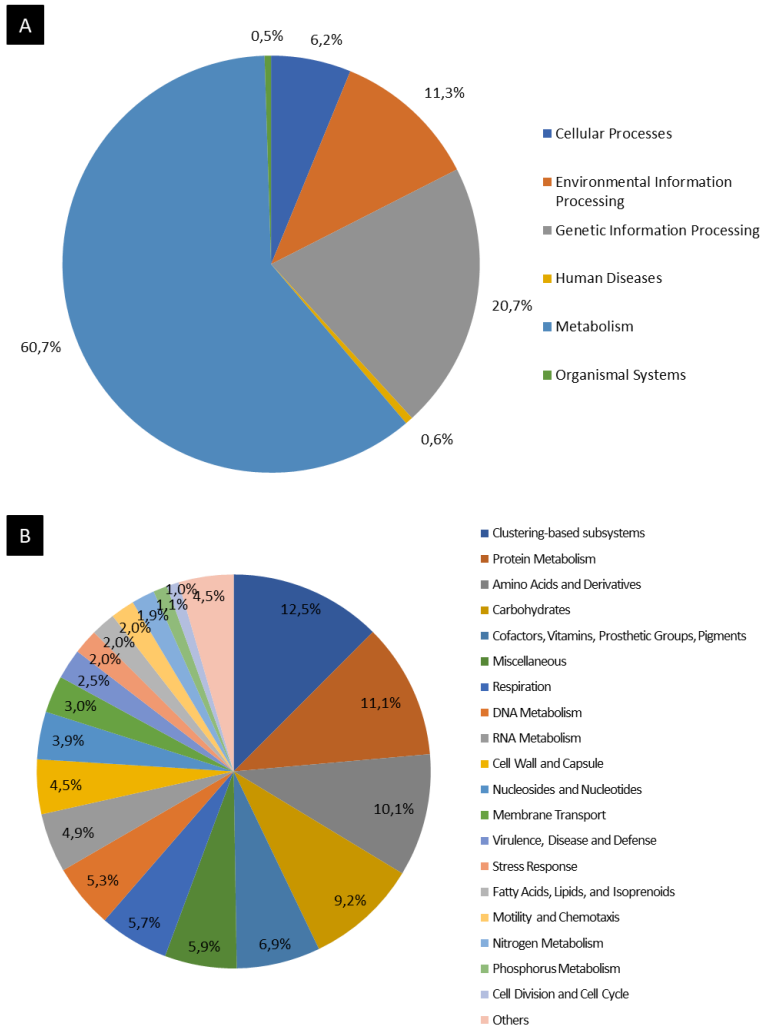


Figure 5. Pie charts of the functional categorization of reads based on alignment to either the (A) KO or (B) Subsystems databases. For the KO database 60.7% of the features were assigned to the Metabolism category, 20.7% to Genetic Information Processing and 11.3% to Environmental Information Processing. For the Subsystems database 12.5% of the features were categorized as Clustering-based subsystems, 11.1% as Protein Metabolism, 10.1% as Amino Acids and Derivatives, 9.2% as Carbohydrates, 6.9% as Cofactors, Vitamins, Prosthetic Groups and Pigments, 5.9% were assigned to the Miscellaneous category and 5.7% to Respiration.

Dominant genera and main metabolic pathways of the Muiño da Veiga hot spring microbial community

The presence of several sulphur reducing genera as part of the dominant genera of the microbial community pointed to the importance of the role of this metabolic pathway in the ecosystem. Indeed, the dominant genus was *Sulfurihydrogenibium* (19.3% of total reads), that contains neutrophilic, thermophilic and anaerobic to microaerobic species (31). Species in this genus are chemolitoautotrophs and facultatively heterotrophs. They are able to employ sulphur compounds as electron donors and molecular oxygen as the electron acceptor, but some species are known to use other donors (like molecular hydrogen) and acceptors (nitrate, Fe(III) or arsenate, for example) as well (32). We also found *Thermodesulfovibrio* (8.78%), *Aquifex* (6.60%), *Hydrogenobacter* (4.92%), *Thermus* (4.64%), *Thermocrinis* (2.33%), *Geobacter* (2.33%), *Hydrogenobaculum* (1.15%), *Thiomonas* (0.90%) and *Hydrogenivirga* (0.83%) as prevalent genera related to this metabolism. Sulphur metabolism was also confirmed to be present as revealed by the KEGG mapping assessment, as several key enzymes for the reduction of fixation of this element had assigned reads by alignment to the KO and Subsystems databases (Figure 6). *Sulfurihydrogenibium*, *Aquifex*, *Hydrogenobacter*, *Thermocrinis*, *Hydrogenobaculum*, *Thiomonas* and *Hydrogenivirga* are sulphur oxidizing (sulphur compounds are used as electron donors), whereas *Thermodesulfovibrio*, and some *Thermus* and *Geobacter* species have a sulfate-reducing metabolism (sulphur compounds act as the electron acceptor). Whereas a synergetic cycling of the sulphur compounds is expected to happen between the two groups, each member within each group will have to fill a certain niche or enter competition when very similar metabolisms are present.

Sulphur oxidizing bacteria

As mentioned, *Sulfurihydrogenibium* dominates the hot spring with almost one fifth of the total reads assigned to this genus. Their

plasticity, acting as anaerobes or microaerobes and as chemolitoautotrophs or facultatively heterotrophs, probably contributes to their success in the Muiño da Veiga hot spring as it makes them adaptable to small disturbances in the environmental conditions. Hot springs where *Sulfurihydrogenibium* dominates over other microorganisms have been reported, although our results show a more diverse community than other hot springs (33,34), which may be attributed to a process of adaptation to changing conditions in the hot spring. Also, Muiño da Veiga rather neutral pH and temperatures (pH 5 and 68°C) not in the hyperthermophilic range can be considered more favourable to more varied communities than acidic hot springs or with a very high temperature (30). Nevertheless, the Aquificae group (to which abundant genera *Sulfurihydrogenibium*, *Aquifex*, *Hydrogenobacter*, *Thermocrinis*, *Hydrogenobaculum* and *Hydrogenivirga* belong to) have been highlighted as the main carbon fixators in terrestrial hot springs, with the energy metabolism driven by either hydrogen or reduced sulphur compounds oxidation (34).

The microaerophile and strictly chemolithoautotrophic genus *Aquifex* uses hydrogen, thiosulfate and elemental sulphur as electron donors and oxygen and nitrate as electron acceptors (35). They fixate all necessary carbon from environmental CO_2 using the reverse (reductive) Tricarboxylic Acid Cycle (TCA) and perform gluconeogenesis through the Embden–Meyerhof–Parnas pathway (36). As they have an anaerobic anabolism while displaying an aerobic catabolism, they are deemed primitive. Their microaerobic growth conditions are explained by this duality in their metabolism (37). Very similarly, *Hydrogenobacter* are microaerophilic and obligately chemolithotrophic. They have a respiratory metabolism using O_2 as the electron acceptor and molecular hydrogen as the electron donor, but reduced sulphur compounds can also be used (38,39). Carbon dioxide is fixed through the TCA cycle (40), and ammonium and nitrate salts serve as nitrogen sources (41). As with *Aquifex*, this genus has an ambivalent metabolism of anaerobic

anabolism and aerobic catabolism supporting their position in the early-branching Aquificales order (42). Their very similar metabolism may suggest competence for the same metabolic niche within the community.

Thermocrinis is a genus of microaerophilic thermophiles with chemolithoautotrophic or chemoorganoheterotrophic growth (43). For chemolithoautotrophic growth all species can use molecular hydrogen, thiosulfate and elemental sulphur as sole electron donors and oxygen acting as electron acceptor (44,45). It has also been discovered that some species can use arsenite and monothioarsenate as electron donors producing arsenate, adding metabolic diversity to the microbial community (46). Chemoorganoheterotrophic growth is more varied, where simple or complex organic compounds can serve as carbon sources for some species, while others lack such capability (47). The genus has often been an abundant representative of microbial communities in terrestrial hot spring ecosystems (44).

Hydrogenobaculum is a genus with respiratory metabolism, more commonly associated with acidic hot springs (48). They can use molecular hydrogen and reduced sulphur compounds as electron donors and molecular oxygen as electron acceptor, whereas carbon dioxide can be the sole carbon source, fixed following the reductive tricarboxylic acid cycle (49). Arsenite is known to be oxidized to arsenate in the absence of H_2S in some species, but is not used as energy source (50,51), likely because sulphide can inhibit microbial As(III) oxidases (52). Dominance of this genera in certain hot springs has been explained by their ability to adapt to changes in temperature and oxygen concentration (28).

Thiomonas comprises a genus of acidophilic obligate aerobes, with some species moderately thermophilic. They are facultative chemolithoautotrophs unable to denitrify nor oxidize ferrous iron. They can use a range of complex organic compounds and sugars as carbon sources, or carbon dioxide through the Calvin-Benson-

Bassham cycle, using reduced sulphur compounds, like thiosulfate, or elemental sulphur as electron donors (53). As heterotrophs, they can grow on acetate or pyruvate, but molecular hydrogen has been reported to be an electron donor in some strains too (54). It has been reported that members of this genus are important As(III)-oxidising microorganisms in arsenic-contaminated ecosystems, removing this more toxic arsenic species coupled to their energy metabolism, with key metabolic genes like those involved in CO_2 fixation regulated differently between strains (55).

Hydrogenivirga are anaerobic to microaerobic thermophilic microorganisms with a strict NaCl requirement for growth. They are strictly chemolithoautotrophic, with the ability to use molecular hydrogen or elemental sulphur as electron donors, and oxygen or nitrate as electron acceptors (56,57).

Sulphur reducing bacteria

Both *Thermodesulfovibrio* (58) and *Geobacter* are strict anaerobes and both can undergo the reduction of sulphur compounds and Fe(III), but nitrate is also a possible electron acceptor for species in both genera (58,59). Coupling of acetate oxidation to the reduction of either Fe(III) or elemental sulphur creates a link in iron and sulphur geochemistry in anaerobic conditions (60). It is worth mentioning that *Geobacter* is considered a mesophilic genus. While the presence of non-thermophilic genera in the hot spring is surprising at first, studies of the genome structure had revealed a number of heat-shock proteins in some non-thermophilic strains, that could explain their presence in habitats with a higher temperature than their optima (61). Interestingly, there are species within the genus *Thermodesulfovibrio* that use hydrogenotrophic methanogens as the electron-accepting system (62), which may be of relevance because of the presence of methanogens in the rare genera portion of the community, as discussed further below.

On the other hand, the genus *Thermus* is considered one of the most metabolically varied within thermophilic microorganisms inhabiting hot springs, and thus is less specialized than the other two sulphur reducing bacteria. Although some species are obligate aerobes, analysis of the pangenome reveals anaerobic respiration capability, with species capable of polysulfide, ferric iron and nitrate reduction, in decreasing frequency order, and the role of the genus in the associated metabolic pathways seems to be more relevant than thought (63,64). The oxidation of pyruvate, lactate and acetate as electron donors coupled to Fe(III) reduction is documented (64). Metal-reducing species of *Thermus* are thought to be widely distributed due to the broad range of electron acceptors available through their metabolism (64).

Non sulphur metabolism-related microorganisms within the dominant genera of the Muiño da Veiga hot spring

Although most of the most abundant genera in the hot spring could be linked to various degrees to the sulphur cycle pathways, there were also several of them that did not have described species involved in this metabolism. As a common trait for these microorganisms not involved in the sulphur cycle, we found all of them specialized in the use of various simple to complex carbohydrates, and in most instances with some degree of involvement in the nitrogen cycle. Their success as prominent members of the hot spring, probably comes from this divergence from the apparent main metabolic pathways involved in the microbial community, as these species probably are filling niches that complement the use of nutrients and energy sources not exploited otherwise.

Acidovorax (1.34%) is a reclassified genus from *Pseudomonas* with at least 15 different species (65), with environmental species (from soil and freshwater) and phytopathogenic species. In water, they are often associated with blooming populations of cyanobacteria. Although first described as mesophilic species, several studies

across the world have now linked some *Acidovorax* species to moderate (66,67) to high temperature hot springs (68,69), and in particular to acid-sulphate springs or iron-rich springs. The presence of temperature-resistant proteins has been reported and explains their ability to grow in thermophilic habitats (69). They are aerobic and chemoorganotrophic, using oxygen as the terminal electron acceptor, or in some species using the heterotrophic denitrification of nitrate metabolic pathway (65,70). Their metabolism of carbohydrates is expected to complement other species enzymatic pathways (66), for example with the use of benzoate and xylene degradation.

The genus *Clostridium* (1.07%) is composed of a group of heterogenous obligate anaerobic species, very widely distributed. They are known for their genetic plasticity which enables species in the genus to use a wide range of nutrients and to colonize various habitats, including hot springs. The identification of *Clostridium* pangenome core genes points out that they can use glucose and glycerol as the sole carbon source, and that ammonium is the preferred source of inorganic nitrogen (72). Several thermophilic species have been described, with several biotechnologically interesting metabolic capabilities like the ability to degrade cellulose and the production of bioethanol by fermentation of cellulose and hemi-cellulose degradation-derived compounds (73), which also hints to the metabolic niche they fulfil in the community.

Caldicellulosiruptor (0.78%) are thermophilic strict anaerobes. Their metabolism consists in the fermentation of monosaccharides, disaccharides and polysaccharides as carbon sources, including cellulose as the sole carbon source, which makes them attractive from a bioprospecting stand point (74). All species are capable of glycolysis through the Embden-Meyerhof-Parnas pathway, fermentation of xylose following the non-oxidative pentose phosphate pathway and reduction of pyruvate with an incomplete TCA cycle (75).

Nitrogen metabolism

As mentioned when discussing their respective metabolisms, several predominant genera are also involved in the nitrogen cycling of the hot spring. *Sulfurihydrogenibium*, *Aquifex*, *Hidrogenivirga*, *Thermodesulfovibrio*, *Geobacter*, and some *Thermus*, *Acidovorax* and *Meiothermus* species are all reported to use nitrate as an electron acceptor and *Hydrogenobacter* and *Clostridium* grow on ammonium or nitrate salts as nitrogen sources. KEGG mapping allowed to reconstruct main metabolic reactions in the nitrogen cycle, with all the major enzymes present for all pathways except anaerobic ammonium oxidation (anammox) which was completely absent. The pathways present included dissimilatory and assimilatory nitrate reduction, denitrification, nitrogen fixation and complete nitrification.

The most abundant genus of the hot spring, *Sulfurydrogenibium*, is known to be able to perform complete denitrification of nitrate to molecular nitrogen (31). *Acidovorax* is known to also perform denitrification (65,70) and main genes involved in this process like *nirK*, *norB* and *noZ* could all be linked to *Acidovorax* in our metagenomic data by cross referencing KO and RefSeq databases, an association supported by previously reported data (76). Denitrification of nitrate to form nitrite and molecular nitrogen is also described in *Aquifex* (35) and *Hydrogenivirga* (57).

Hydrogenobacter species are able to perform nitrate reduction (77) and we found the *nirA* gene when we cross referenced the functional and taxonomical databases.

Regarding nitrogen fixation, the *nifH* gene for nitrogenase was found in the KO database alignment to the metagenome, and species from the genera *Thermodesulfovibrio*, *Thermocrinis*, *Hydrogenobacter* and *Caldicellulosiruptor* are all reported to possess the *nifH* gene and are putative nitrogen fixing microorganisms (78,79). Cross reference between the KO database and the RefSeq

database confirmed all four genera harbouring the *nifH* gene in the Muiño da Veiga metagenome.

Lastly, the oxidation of ammonia to nitrate through the nitrification pathway (through *amoABC*, *hao*, and *nxrAB*/*narGH* genes) was found to be present (Figure 6B). A single hit against the KO database for the *amoC* gene that cross referenced with the RefSeq database as a *Nitrosospira* genus bacterium hit, a rare member of the microbial community (0.13%). Moreover, the archaeal *amoA*-harbouring genus *Nitrosopumilus* (80) was also present in the metagenome, though as a very rare member of the ecosystem (0.01%), opposite to other reports where ammonia oxidizing archaea had been described as playing a dominant role in the nitrogen cycle in a big number of terrestrial hot springs (81,82). The *hao* gene was present in the metagenome, with hits only from the Deltaproteobacteria class, corresponding to relatively low abundant genera in the hot spring *Anaeromyxobacter* (0.35%) and *Desulfobacterium* (0.06%). Both *narG* and *narH* (but not *nxrA* nor *nxrB*) were present as revealed by alignment to the KO database, and by cross referencing with the RefSeq database main genera from the hot spring *Hydrogenobacter*, *Meiothermus*, *Geobacter* and *Acidovorax* were revealed to harbour it. As only the last step of nitrification was represented by dominant genera of the ecosystem, probably the relevance of nitrite conversion to nitrate is more significant in the hot spring than complete nitrification from ammonia, which may be used by other microorganisms as a nitrogen source rather than an energy source in the nitrogen cycle.

Hydrogen metabolism

Hydrogen plays an important role in some hot springs communities as it links several metabolic pathways, especially as a key intermediate in anaerobic conditions (83). For example, its generation by nitrogen-fixing bacteria or by fermentative metabolism, as well as its role as electron donor and consumption by sulphate reducing bacteria and methanogens in hot springs is

well documented (83,84). However, its importance in the metabolism is dependent of several factors (availability compared to other electron donors, presence of coupling electron acceptors and how thermodynamically favourable the coupling is compared to others) (85,86). Reportedly acidic hot springs are generally more prone to have a hydrogen-favoured metabolism compared to alkaline ones where sulphur compounds may be preferred (85). A considerable number of dehydrogenases were detected when the reads were aligned to the KO functional database. However, dehydrogenases constitute an heterogenous group of enzymes that are hard to classify and make functional predictions based on sequence alone (85). Nevertheless, its noteworthy that several of the dominant genera of the hot spring contain potential hydrogen-using species, as molecular hydrogen can be used as an electron donor by *Sulfurihydrogenibium*, the dominant genus of the hot spring. The same is true for all other sulphur-oxidizing genera prevalent in the hot spring: *Aquifex*, *Hydrogenobacter*, *Thermocrinis*, *Hydrogenobaculum*, some *Thiomonas* species and *Hydrogenivirga*.

Carbon metabolism

KEGG mapping using the KO database revealed the dissimilatory metabolism of carbon, including a complete TCA cycle and almost complete Gluconeogenesis and Pentose Phosphate pathways in the metagenome. For the assimilatory pathways, a complete reverse TCA cycle and an almost complete reverse Pentose Phosphate cycle (Calvin cycle) were also found. These results made apparent that main pathways for the use of carbon were all present in the microbial community. Given the results from the main genera survey with the RefSeq database, it seems likely that the reverse TCA cycle for CO_2 fixation is a main driving force in the metabolic network of the Muiño da Veiga ecosystem, (as mentioned before Aquificae may constitute the main carbon fixators in terrestrial hot springs (34)). Nevertheless, the range of substrates that can be used

by the community inferred from the metabolic potential of the dominant genera in the metagenome is remarkable, as described in the genera descriptions above. With simple to complex compounds as possible substrates, including (as an example) complex carbohydrates like cellulose and xylan, the potential for this habitat as a source of interesting metabolites from a biotechnological standpoint is to be highlighted.

Rare species metabolism

The so-called rare biosphere is often defined by an arbitrary numerical threshold of sequence counts (often around 0.1% to 0.01% relative abundance in a data set) (87). Although classically ignored in favour of the dominant genera in a microbial community, it is now widely accepted that the rare members of the community can play uneven key roles in metabolic pathway networks like element cycling and bioremediation (88) and that they constitute a genetic and functional diversity bank that supports ecological resilience (87,88). Moreover, the composition of the rare fraction of a microbial community is expected to be highly variable in the event of changes in the ecosystem environmental conditions (88), or in other words, some species might be only conditionally or temporary rare and alterations in the ecosystem would promote such species to more prominent ecological roles (29,88). Two particular metabolic processes were found by KEGG mapping using the KO database that could only be attributed to genera that could be regarded as rare, namely photosynthesis and methanogenesis.

Photosynthesis in extreme temperature conditions

Almost complete Photosyntesis (00195) and Carbon Fixation in Photosyntetic microorganisms (00710) KEGG mapping pathways were present in the Muiño da Veiga metagenome, indicating the possibility of this autotrophic metabolism playing a role in the microbial community. Indeed, photosynthesis is possible at the temperature of the hot spring (68°C) as the upper temperature limit

is considered 73°C in non-acidic habitats (89). Some species of the Cyanobacteria phyla and a limited number of Chloroflexi ones have been observed to play the role of primary production through the photosynthetic pathway in hot springs, especially in alkaline conditions, where temperature plays an important role in species diversity (90). We found that the major genera responsible of photosynthesis in hot springs were present in the metagenome, although their abundance was somewhat low compared to other much prominent members of the microbial community, which was also in agreement to the KEGG mapping often returning singular hits for specific enzymes involved in the photosynthetic pathway. For Cyanobacteria, we found both *Synechoccus* (0.23% of total reads) and *Thermosynechococcus* (0.06%) present in the sample, as well as photosyntetic Cloroflexi (90) *Roseiflexus* (0.56%), *Chloroflexus* (0.31%) and *Oscillochloris* (0.05%). The reads belonging to these major genera of photosynthetic microorganisms related to hot springs represented 1.21% of the total sequences. Moreover, the subterranean nature of these waters limits their exposition to sunlight (Figure 6A) although the surface of the hot springs could sustain the growth of microorganisms that do use these metabolic pathways. Taking it all together, photosynthesis probably plays a very minor role compared to other autotrophic metabolic routes that are also present in the metagenome and related to more abundant genera.

Archaea and roles in methane metabolism

Several different Archaea genera were found in the metagenome, representing a small fraction of the total reads of the metagenome (1.96%). 57.15% of the total Archaea assigned reads were aligned with 9 predominant genera, many of which were found to be strict anaerobes and methanogens, mainly belonging to the *Euryarchaeota* phylum, with the exception of the genus *Metallosphaera* from the *Crenarchaeota* phylum. Indeed, the Methane Metabolism pathway was mostly complete by KEGG

mapping of the metagenomic reads, including all four of the methanogenesis modules, although the aceticlastic fermentation of acetate (acetate decarboxylation) returned more hits than the others, and the catabolism of methylamines was the most incomplete.

Nevertheless, the most abundant Archaea genus was *Methanosarcina* (0.20%), considered one of the most diverse methanogenic archaea regarding their metabolism. These anaerobic thermophiles can use all known methanogenic substrates using the four different catabolic pathways (CO_2 reduction with H_2, methyl reduction with H_2, aceticlastic fermentation of acetate or methylotrophic catabolism of methanol, methylated amines and dimethylsulfide), some species can grow without methane production using carbon monoxide, and most species can fix molecular nitrogen (91). Other thermophilic anaerobic methanogens ranking as most abundant Archaea included *Methanocaldococcus* (0.14%) and *Methanococcus* (0.11%) (92,93), *Methanosaeta* (*Methanotrix*) (0.10%) that follows the acetoclastic (acetate decarboxylation) pathway for metanogenesis (94,95) and *Methanothermobacter* (0.8%), that is capable of reduction of CO_2 to methane, using H_2 (and some species formate as well) as electron donor. (96) Notably, we found the thermophilic, anaerobic and methanogenic *Archeoglobus* (97) (0.16%) the second most abundant Archaea genus. This genus includes species that grow as chemolithoautotrophs in the presence of H_2, CO_2 and thiosulfate, and as chemoorganotrophs with several complex organic compounds, using sulfate, sulfite and thiosulfate as electron acceptors.

Other non-methanogenic genera within the most abundant archaea included *Pyrococcus* (0.13%) and *Thermococcus* (0.11%), hyperthermophilic anaerobes that grow as chemoorganoheterotrophs capable of fermentation of complex organic substrates and reduction of sulphur species (98,99). Lastly,

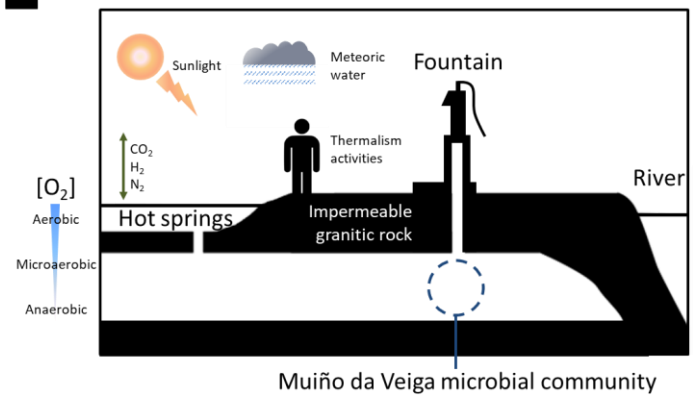
Methallosphaera (0.09%) was the only aerobic archaea found within the 9 dominant genera, that contains facultatively chemolithoautotrophic species, capable of growing on sulfidic ores or elemental sulphur, or chemoorganotrophically using complex organic substrates but not sugars (100).

With the exception of the methanogenic species, the fact that the metabolism characteristics of Archaea species are similar to those of the predominant bacteria genera make them compete for similar resources in the hot spring, serving as a putative explanation for the lesser occurrence of these other species in the community.

Conclusions

Taking into account all putative genetic and functional diversity found in the metagenome, a complex network of potential metabolic interactions within the microbial community can be stablished, where most elements cycles were found to be present. This diversity probably is the result of the optimization between using all possible nutrients and filling all the available niches minimizing metabolic overlaps, not unlike what has been described for other hot springs (101). Other factors that should be contributing are the water temperature not being high enough to be a limiting diversity factor (102) and a circumneutral pH favouring it too (103). A synergetic effect has been described between metabolic diversity and increase of biomass in the hot spring ecosystem observed when temperatures are below the hyperthermophilic range. The presence of varied primary producers with autotrophic pathways expands the range or available nutrients and with the increase of biomass microbial mats become thicker and in turn expand the available niches (104). Such niche-facilitating microbial mats has been described to be formed by the most abundant genus of the community, *Sulfurihydrogenibium* (105,106). A global network of metabolic interactions proposed for the Muiño da Veiga hot spring community is given in Figure 6.

A



B

Metabolism	KEGG pathway	EC number	Gene	Acid	Aqui	Cald	Clos	Geob	Hyvg	Hydr	Hybl	Meio	Sulf	Thcr	Thde	Ther	Thio	AM
Sulfate metabolism	Assimilatory sulfate reduction, sulfate => H2S	2.7.7.4	*met3*		59					302	8	23		178	1221	334		7666
			CysN	13		22		4										645
			CysD	32		6	11	36									51	1166
		2.7.1.25	*CysC*				15	3				12			10			646
			CysNC			12	12	1									84	715
		1.8.4.8	*CysH*					3				4				116		1218
		1.8.1.2	*CysJ*	2														199
			CysI	25														1140
		1.8.7.1	*sir*									18						126
	Dissimilatory sulfate reduction, sulfate => H2S	2.7.7.4	*met3*		59					302	8	23		178	1221	334		7666
			CysN	13		22		4										645
			CysD	32		6	11	36									51	1166
		1.8.99.2	*AprA*			16									3490			8045
			AprB												592			1685
		1.8.99.5	*DsrA*												2058			2807
			DsrB												1931			2766
	Cysteine biosynthesis, serine => cysteine	2.3.1.30	*CysE*	56		226	66	15				16					89	2361
		2.5.1.47	*CysK*	14		136	51	212				57				285		3525
			CysM	142	1692			114		533	309		2351	514	672		43	8000
		2.5.1.65	*CysO*															
Nitrogen metabolism	Nitrogen fixation, nitrogen => ammonia	1.18.6.1	*nifD*			26	1	8		173				1527	81			4078
			nifK			8		10							249			2375
			nifH			9		81		610				360	36			2213
	Assimilatory nitrate reduction, nitrate => ammonia	1.7.7.2	*narB*															28
		1.7.1.1	*NIA1*															365
		1.7.1.3	*NIA2*															3
		1.7.7.1	*nirA*							957			54	429		302		3049
		1.7.1.4	*nirB*	104	1421					1379	25			389				4757
			nirD	5														150
	Dissimilatory nitrate reduction, nitrate => ammonia	1.7.5.1	*narG*	137				443		3903		815					4	11450
			narH	22				140		992		325					10	5045
			narI	27				1				145						1134
			narZ															84
		1.9.6.1	*napA*					1					3124					13440
		1.7.1.4	*nirB*	104	1421					1379	25			389				4757
			nirD	5														150
		1.7.2.2	*nrfA*					759				16			1049			2341
	Denitrification, nitrate => nitrogen	1.7.5.1	*narG*	137				443		3903		815					4	11450
			narH	22				140		992		325					10	5045
			narI	27				1				145						1134
			narZ															84
		1.9.6.1	*napA*					1					3124					13440
		1.7.2.1	*nirK*	51							624							2494
		1.7.2.5	*norB*	9				6		2073			4207					9606
		1.7.2.4	*nosZ*	7						2970								6970
	Nitrification, ammonia => nitrite	1.14.99.39	*amoA*															
			amoB															
			amoC															1
		1.7.2.6	*hao*															204

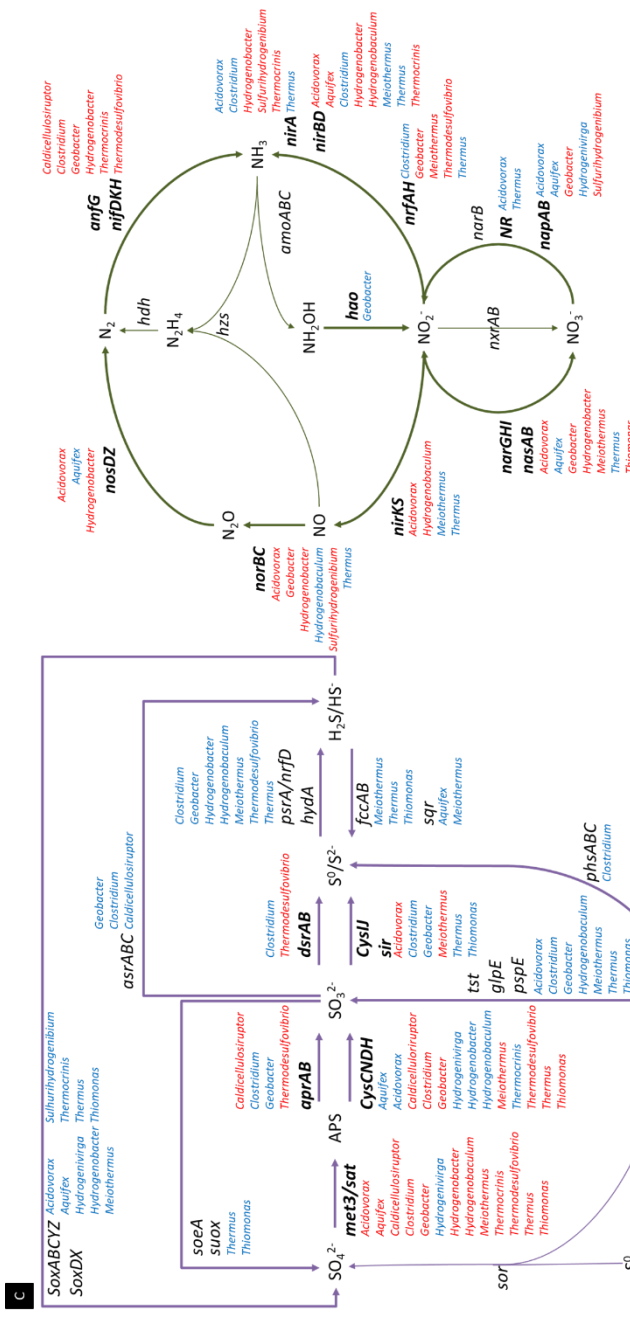

c
SoxABCYZ
SoxDX
Acidovorax
Aquifex
Hydrogenivirga
Hydrogenobacter
Meiothermus
Sulfurihydrogenibium
Thermocrinis
Thermus
Thiomonas
soeA
suox
Thermus
Thiomonas
Caldicellulosiruptor
Clostridium
Geobacter
Thermodesulfovibrio
aprAB
met3/sat
Acidovorax
Aquifex
Caldicellulosiruptor
Clostridium
Geobacter
Hydrogenivirga
Hydrogenobacter
Hydrogenobaculum
Meiothermus
Thermocrinis
Thermodesulfovibrio
Thermus
Thiomonas
CysCNDH
Aquifex
Acidovorax
Caldicellulosiruptor
Clostridium
Geobacter
Hydrogenivirga
Hydrogenobacter
Hydrogenobaculum
Meiothermus
Thermocrinis
Thermodesulfovibrio
Thermus
Thiomonas
tst
glpE
pspE
Acidovorax
Clostridium
Geobacter
Hydrogenobaculum
Meiothermus
Thermus
Thiomonas
CysIJ
sir
Acidovorax
Clostridium
Geobacter
Meiothermus
Thermus
Thiomonas
fccAB
Meiothermus
Thermus
Thiomonas
sqr
Aquifex
Meiothermus
asrABC
Caldicellulosiruptor
Geobacter
Clostridium
Clostridium
Geobacter
Hydrogenobacter
Hydrogenobaculum
Meiothermus
Thermodesulfovibrio
Thermus
dsrAB
Clostridium
Thermodesulfovibrio
psrA/nrfD
hydA
phsABC
Clostridium
doxDA
Clostridium
ttrABC
Acidovorax
Clostridium
Geobacter
Hydrogenivirga
Hydrogenobaculum
Thermodesulfovibrio
Thermus
Thiomonas
sor
APS
SO4 2-
SO3 2-
S0/S2-
H2S/HS-
S0
S2O3 2-
S4O6 2-
anfG
nifDKH
Caldicellulosiruptor
Clostridium
Geobacter
Hydrogenobacter
Thermocrinis
Thermodesulfovibrio
nosDZ
Acidovorax
Aquifex
Hydrogenobacter
norBC
Acidovorax
Geobacter
Hydrogenobacter
Hydrogenobaculum
Sulfurihydrogenibium
Thermus
nirKS
Acidovorax
Hydrogenobaculum
Meiothermus
Thermus
narGHI
nasAB
Acidovorax
Aquifex
Geobacter
Hydrogenobacter
Meiothermus
Thermus
Thiomonas
nxrAB
narB
NR
napAB
Acidovorax
Thermus
Aquifex
Geobacter
Hydrogenivirga
Sulfurihydrogenibium
nrfAH
Clostridium
Geobacter
Meiothermus
Thermodesulfovibrio
Thermus
nirA
nirBD
Acidovorax
Aquifex
Clostridium
Hydrogenobacter
Hydrogenobaculum
Meiothermus
Thermus
Thermocrinis
Acidovorax
Clostridium
Hydrogenobacter
Sulfurihydrogenibium
Thermocrinis
Thermus
amoABC
hao
Geobacter
hdh
hzs
N2
N2O
NO
NO2-
NO3-
N2H4
NH2OH
NH3

Figure 6. (A) Schematic representation of Muiño da Veiga hot springs and sampling site (fountain). The impermeable granitic rock prevents water from meteoric origin (rain, Miño river) to mix with the subterranean hot water. Sunlight exposition is limited to the surface of the hot spring pools, where the exchange of gaseous elements takes place. Oxygen concentration decreases with water depth. As the pools are open for thermalism activities, some degree of cross contamination may take place too. (B) Occurrence of genes in the metagenome from Muiño da Veiga involved in selected metabolic processes in main members of the microbial community as found by cross referencing KO and RefSeq databases. In green, genes that returned zero reads. In yellow to red gradient, number of reads (in ascending order) that have a hit both KO and RefSeq databases. *Acid*: *Acidovorax*; *Aqui*: *Aquifex*; *Cald*: *Caldicellulosiruptor*; *Clos*: *Clostridium*; *Geo*: *Geobacter*; *Hyvg*: *Hydrogenivirga*; *Hydr*: *Hydrogenobacter*; *Hybl*: *Hydrogenobaculum*; *Meio*: *Meiothermus*; *Sulf*: *Sulfurihydrogenibium*; *Thcr*: *Thermocrinis*; *Thde*: *Thermodesulfovibrio*; *Ther*: *Thermus*; *Thio*: *Thiomonas*; AM: all reads across the metagenome. (C) Sulfur and nitrogen cycles in the Muiño da Veiga metagenome. Names in bold represent genes that were present in the cross reference of the KO and RefSeq databases (our data) with the metagenome for the main genus of the community. Thick lines represent reactions that were supported by either the cross reference of KO and RefSeq databases with the metagenome or by cross reference of the Uniprot database for the Enzymatic function (EC number) and the Organism name (genera) of the main genus of the community (published data). Genus names in red were present in the cross reference between the KO and the RefSeq databases and genus names in blue were present in the cross refence of the Uniprot database for the Enzymatic function (EC number) and the Organism name (genera).

Acknowledgments

Part of the bioinformatic analysis in this chapter was performed by Escuder-Rodríguez JJ as part of his visit as a fellow student in Professor Robert Edwards Laboratory at San Diego State University, with help and supervision from Professor Robert Edwards and Dr. Genivaldo Gueiros. Additional training in bioinformatic tools usage was also provided by other members of the group. Access to the

supercomputing servers from the Department of Computer Sciencies in San Diego State University was also granted.

Bioinformatic analysis previous to the MG-RAST upload were performed using Finisterrae II servers provided by the Supercomputing Center of Galicia (CESGA).

Additional Founding

Escuder-Rodríguez JJ received a predoctoral fellowship "Axudas para estadías predoutorais INDITEX-UDC 2016" with budgement identification 0400470001 422D 48700; "Convenio de colaboración entre a UDC e Inditex para a internacionalización dos estudos de doutoramento".

Escuder-Rodríguez JJ received a predoctoral fellowship from "Agrupación CICA-INIBIC: ayudas para estancias pre y postdoctorales 2016" with budgement identification 6100240002 541A 6490211.

References

1. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev [Internet]. 1995 Mar;59(1):143–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7535888

2. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol [Internet]. 1985;39:321–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3904603

3. Torsvik V, Sørheim R, Goksøyr J. Total bacterial diversity in soil and sediment communities—A review. J Ind Microbiol Biotechnol [Internet]. 1996 Sep;17(3–4):170–8. Available from: https://academic.oup.com/jimb/article/17/3-4/170-178/5988979

4. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol [Internet]. 2012 Apr;21(8):1794–805. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2012.05538.x

5. Urbieta MS, Donati ER, Chan K-G, Shahar S, Sin LL, Goh KM. Thermophiles in the genomic era: Biodiversity, science, and applications. Biotechnol Adv [Internet]. 2015 Nov;33(6):633–47. Available from: http://dx.doi.org/10.1016/j.biotechadv.2015.04.007

6. Galloway-Peña J, Hanson B. Tools for Analysis of the Microbiome. Dig Dis Sci [Internet]. 2020 Mar 31;65(3):674–85. Available from: http://link.springer.com/10.1007/s10620-020-06091-y

7. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol [Internet]. 1998 Oct;5(10):R245–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1074552198901089

8. Escuder-Rodríguez J-J, DeCastro M-E, Rodríguez-Belmonte E, Becerra M, María-Isabel G-S. Microbiomes of Extreme Environments [Internet]. 1st ed. Yadav AN, Rastegari AA, Yadav N, editors. Microbiomes of Extreme Environments: Biodiversity and Biotechnological Applications. First edition. | Boca Raton : CRC Press, 2021.: CRC Press; 2021. 19 p. Available from: https://www.taylorfrancis.com/books/9781000734911

9. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics [Internet]. 2011 Mar 15;27(6):863–4. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr026

10. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics

[Internet]. 2014 Mar 1;30(5):614–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt593

11. Silva GGZ, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics [Internet]. 2016 Feb 1;32(3):354–61. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv584

12. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods [Internet]. 2015 Jan 17;12(1):59–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25402007

13. Bethesta MD, National Library of Medicine US. National Center for Biotechnology Information (NCBI) [Internet] [Internet]. [cited 2016 Sep 28]. Available from: https://www.ncbi.nlm.nih.gov/

14. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics [Internet]. 2008 Dec 19;9(1):386. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-386

15. Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, D'Souza M, et al. A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. PLoS Comput Biol. 2012;8(6).

16. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.

17. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics [Internet]. 2014 Jun 12;15(1):182.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/24925680

18. Aronesty E. Comparison of Sequencing Utility Programs. Open Bioinforma J [Internet]. 2013 Jan 31;7(1):1–8. Available from: https://openbioinformaticsjournal.com/VOLUME/7/PAGE/1/

19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Apr 4;9(4):357–9. Available from: http://www.nature.com/articles/nmeth.1923

20. Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28(24):3211–7.

21. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics [Internet]. 2012 Dec;28(23):3150–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts565

22. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Res [Internet]. 2002 Mar 20;12(4):656–64. Available from: http://www.genome.org/cgi/doi/10.1101/gr.229202

23. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res [Internet]. 2010 Nov;38(20):e191–e191. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq747

24. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D733–45. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189

25. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M.

KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D457-62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26476454

26. Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang H-Y, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res [Internet]. 2005;33(17):5691–702. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16214803

27. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics [Internet]. 2011 Dec 30;12(1):385. Available from: http://www.biomedcentral.com/1471-2105/12/385

28. Lin K-H, Liao B-Y, Chang H-W, Huang S-W, Chang T-Y, Yang C-Y, et al. Metabolic characteristics of dominant microbes and key rare species from an acidic hot spring in Taiwan revealed by metagenomics. BMC Genomics [Internet]. 2015 Dec 3;16(1):1029. Available from: http://dx.doi.org/10.1186/s12864-015-2230-9

29. Wang Y, Hatt JK, Tsementzi D, Rodriguez-R LM, Ruiz-Pérez CA, Weigand MR, et al. Quantifying the Importance of the Rare Biosphere for Microbial Community Response to Organic Pollutants in a Freshwater Ecosystem. Stams AJM, editor. Appl Environ Microbiol [Internet]. 2017 Apr 15;83(8):1–19. Available from: https://aem.asm.org/lookup/doi/10.1128/AEM.03321-16

30. Sahm K, John P, Nacke H, Wemheuer B, Grote R, Daniel R, et al. High abundance of heterotrophic prokaryotes in hydrothermal springs of the Azores as revealed by a network of 16S rRNA gene-based methods. Extremophiles [Internet]. 2013 Jul 26;17(4):649–62. Available from: http://link.springer.com/10.1007/s00792-013-0548-2

31. Takai K, Kobayashi H, Nealson KH, Horikoshi K. Sulfurihydrogenibium subterraneum gen. nov., sp. nov., from

a subsurface hot aquifer. Int J Syst Evol Microbiol [Internet]. 2003 May 1;53(3):823–7. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02506-0

32. Nakagawa S, Shtaih Z, Banta A, Beveridge TJ, Sako Y, Reysenbach A-L. Sulfurihydrogenibium yellowstonense sp. nov., an extremely thermophilic, facultatively heterotrophic, sulfur-oxidizing bacterium from Yellowstone National Park, and emended descriptions of the genus Sulfurihydrogenibium, Sulfurihydrogenibium subterraneum . Int J Syst Evol Microbiol [Internet]. 2005 Nov 1;55(6):2263–8. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.63708-0

33. Takacs-Vesbach C, Inskeep WP, Jay ZJ, Herrgard MJ, Rusch DB, Tringe SG, et al. Metagenome Sequence Analysis of Filamentous Microbial Communities Obtained from Geochemically Distinct Geothermal Channels Reveals Specialization of Three Aquificales Lineages. Front Microbiol [Internet]. 2013;4(MAY):1–25. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2013.00084/abstract

34. Dodsworth JA, Hungate BA, Hedlund BP. Ammonia oxidation, denitrification and dissimilatory nitrate reduction to ammonium in two US Great Basin hot springs with abundant ammonia-oxidizing archaea. Environ Microbiol [Internet]. 2011 Aug;13(8):2371–86. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2011.02508.x

35. Huber R, Wilharm T, Huber D, Trincone A, Burggraf S, König H, et al. Aquifex pyrophilus gen. nov. sp. nov., Represents a Novel Group of Marine Hyperthermophilic Hydrogen-Oxidizing Bacteria. Syst Appl Microbiol. 1992;15(3):340–51.

36. Deckert G, Warren P V., Gaasterland T, Young WG, Lenox AL, Graham DE, et al. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature. 1998;392(6674):353–8.

37. Beh M, Strauss G, Huber R, Stetter KO, Fuchs G. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium Aquifex pyrophilus and in the archaebacterium Thermoproteus neutrophilus. Arch Microbiol. 1993;160(4):306–11.

38. Bonjour F, Aragno M. Growth of thermophilic, obligatorily chemolithoautotrophic hydrogen-oxidizing bacteria related to Hydrogenobacter with thiosulfate and elemental sulfur as electron and energy source. FEMS Microbiol Lett. 1986;35:11–5.

39. Alfredsson GA, Ingason A, Kristjansson JK. Growth of thermophilic obligately autotrophic hydrogen-oxidizing bacteria on thiosulfate. Lett Appl Microbiol [Internet]. 1986 Aug;2(2):21–3. Available from: http://doi.wiley.com/10.1111/j.1472-765X.1986.tb01507.x

40. Shiba H, Kawasumi T, Igarashi Y, Kodama T, Minoda Y. The CO2 assimilation via the reductive tricarboxylic acid cycle in an obligately autotrophic, aerobic hydrogen-oxidizing bacterium, Hydrogenobacter thermophilus. Arch Microbiol [Internet]. 1985 Apr;141(3):198–203. Available from: http://link.springer.com/10.1007/BF00408058

41. Kawasumi T, Igarashi Y, Kodama T, Minoda Y. Hydrogenobacter thermophilus gen. nov., sp. nov., an Extremely Thermophilic, Aerobic, Hydrogen-Oxidizing Bacterium. Int J Syst Bacteriol [Internet]. 1984 Jan 1;34(1):5–10. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-34-1-5

42. Pitulle C, Yang Y, Marchiani M, Moore ERB, Siefert JL, Aragn M, et al. Phylogenetic Position of the Genus Hydrogenobacter. 2007;(22):620–6.

43. Huber R, Eder W, Heldwein S, Wanner G, Huber H, Rachel R, et al. Thermocrinis ruber gen. nov., sp. nov., a Pink-Filament-Forming Hyperthermophilic Bacterium Isolated from Yellowstone National Park. Appl Environ Microbiol [Internet].

1998 Oct 1;64(10):3576–83. Available from: https://aem.asm.org/content/64/10/3576

44. Dodsworth JA, Ong JC, Williams AJ, Dohnalkova AC, Hedlund BP. Thermocrinis jamiesonii sp. nov., a thiosulfate-oxidizing, autotropic thermophile isolated from a geothermal spring. Int J Syst Evol Microbiol [Internet]. 2015 Dec 1;65(Pt_12):4769–75. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.000647

45. Eder W, Huber R. New isolates and physiological properties of the Aquificales and description of Thermocrinis albus sp. nov. Extremophiles [Internet]. 2002 Aug 27;6(4):309–18. Available from: http://link.springer.com/10.1007/s00792-001-0259-y

46. Härtig C, Lohmayer R, Kolb S, Horn MA, Inskeep WP, Planer-Friedrich B. Chemolithotrophic growth of the aerobic hyperthermophilic bacterium Thermocrinis ruber OC 14/7/2 on monothioarsenate and arsenite. FEMS Microbiol Ecol [Internet]. 2014 Dec;90(3):747–60. Available from: https://academic.oup.com/femsec/article-lookup/doi/10.1111/1574-6941.12431

47. Caldwell SL, Liu Y, Ferrera I, Beveridge T, Reysenbach A-L. Thermocrinis minervae sp. nov., a hydrogen- and sulfur-oxidizing, thermophilic member of the Aquificales from a Costa Rican terrestrial hot spring. Int J Syst Evol Microbiol [Internet]. 2010 Feb 1;60(2):338–43. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.010496-0

48. Massello FL, Chan CS, Chan K-G, Goh KM, Donati E, Urbieta MS. Meta-Analysis of Microbial Communities in Hot Springs: Recurrent Taxa and Complex Shaping Factors beyond pH and Temperature. Microorganisms [Internet]. 2020 Jun 16;8(6):906. Available from: https://www.mdpi.com/2076-2607/8/6/906

49. Stohr R, Waberski A, Völker H, Tindall BJ, Thomm M.

Hydrogenothermus marinus gen. nov., sp. nov., a novel thermophilic hydrogen-oxidizing bacterium, recognition of Calderobacterium hydrogenophilum as a member of the genus Hydrogenobacter and proposal of the reclassification of Hydrogenobacter acidophilus a. Int J Syst Evol Microbiol [Internet]. 2001 Sep 1;51(5):1853–62. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-51-5-1853

50. Romano C, D'Imperio S, Woyke T, Mavromatis K, Lasken R, Shock EL, et al. Comparative genomic analysis of phylogenetically closely related Hydrogenobaculum sp. isolates from yellowstone national park. Appl Environ Microbiol. 2013;79(9):2932–43.

51. Hedlund BP, Reysenbach A-L, Huang L, Ong JC, Liu Z, Dodsworth JA, et al. Isolation of diverse members of the Aquificales from geothermal springs in Tengchong, China. Front Microbiol [Internet]. 2015 Feb 27;6(FEB):1–8. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00157/abstract

52. Donahoe-Christiansen J, D'Imperio S, Jackson CR, Inskeep WP, McDermott TR. Arsenite-Oxidizing Hydrogenobaculum Strain Isolated from an Acid-Sulfate-Chloride Geothermal Spring in Yellowstone National Park. Appl Environ Microbiol [Internet]. 2004 Mar;70(3):1865–8. Available from: https://aem.asm.org/content/70/3/1865

53. Moreira D, Amils R. Phylogeny of Thiobacillus cuprinus and Other Mixotrophic Thiobacilli: Proposal for Thiomonas gen. nov. Int J Syst Bacteriol [Internet]. 1997 Apr 1;47(2):522–8. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-47-2-522

54. Vésteinsdóttir H, Reynisdóttir DB, Örlygsson J. Thiomonas islandica sp. nov., a moderately thermophilic, hydrogen- and sulfur-oxidizing betaproteobacterium isolated from a hot spring. Int J Syst Evol Microbiol [Internet]. 2011 Jan

1;61(1):132–7. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.015511-0

55. Bryan CG, Marchal M, Battaglia-Brunet F, Kugler V, Lemaitre-Guillier C, Lièvremont D, et al. Carbon and arsenic metabolism in Thiomonas strains: differences revealed diverse adaptation processes. BMC Microbiol [Internet]. 2009;9(1):127. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-9-127

56. Nakagawa S, Nakamura S, Inagaki F, Takai K, Shirai N, Sako Y. Hydrogenivirga caldilitoris gen. nov., sp. nov., a novel extremely thermophilic, hydrogen- and sulfur-oxidizing bacterium from a coastal hydrothermal field. Int J Syst Evol Microbiol [Internet]. 2004 Nov 1;54(6):2079–84. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.03031-0

57. Nunoura T, Miyazaki M, Suzuki Y, Takai K, Horikoshi K. Hydrogenivirga okinawensis sp. nov., a thermophilic sulfur-oxidizing chemolithoautotroph isolated from a deep-sea hydrothermal field, Southern Okinawa Trough. Int J Syst Evol Microbiol [Internet]. 2008 Mar 1;58(3):676–81. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.64615-0

58. Henry EA, Devereux R, Maki JS, Gilmour CC, Woese CR, Mandelco L, et al. Characterization of a new thermophilic sulfate-reducing bacterium. Arch Microbiol [Internet]. 1994 Jan;161(1):62–9. Available from: http://link.springer.com/10.1007/BF00248894

59. Lovley DR, Giovannoni SJ, White DC, Champine JE, Phillips EJP, Gorby YA, et al. Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. Arch Microbiol [Internet]. 1993

Apr;159(4):336–44. Available from: http://link.springer.com/10.1007/BF00290916

60. Caccavo F, Lonergan DJ, Lovley DR, Davis M, Stolz JF, McInerney MJ. Geobacter sulfurreducens sp. nov., a hydrogen- and acetate-oxidizing dissimilatory metal-reducing microorganism. Appl Environ Microbiol [Internet]. 1994;60(10):3752–9. Available from: https://aem.asm.org/content/60/10/3752

61. Fortney NW, He S, Converse BJ, Beard BL, Johnson CM, Boyd ES, et al. Microbial Fe(III) oxide reduction potential in Chocolate Pots hot spring, Yellowstone National Park. Geobiology [Internet]. 2016 May;14(3):255–75. Available from: http://doi.wiley.com/10.1111/gbi.12173

62. Sekiguchi Y, Muramatsu M, Imachi H, Narihiro T, Ohashi A, Harada H, et al. Thermodesulfovibrio aggregans sp. nov. and Thermodesulfovibrio thiophilus sp. nov., anaerobic,thermophilic, sulfate-reducing bacteria isolated from thermophilic methanogenic sludge, and emended description of the genus Thermodesulfovibrio. Int J Syst Evol Microbiol. 2008;58(11):2541–8.

63. Zhou E-M, Adegboruwa AL, Mefferd CC, Bhute SS, Murugapiran SK, Dodsworth JA, et al. Diverse respiratory capacity among Thermus strains from US Great Basin hot springs. Extremophiles [Internet]. 2020 Jan 18;24(1):71–80. Available from: https://doi.org/10.1007/s00792-019-01131-6

64. Balkwill DL, Kieft TL, Tsukuda T, Kostandarithes HM, Onstott TC, Macnaughton S, et al. Identification of iron-reducing Thermus strains as Thermus scotoductus. Extremophiles. 2004;8(1):37–44.

65. Willems A, Falsen E, Pot B, Jantzen E, Hoste B, Vandamme P, et al. Acidovorax, a New Genus for Pseudomonas facilis, Pseudomonas delafieldii, E. Falsen (EF) Group 13, EF Group 16, and Several Clinical Isolates, with the Species Acidovorax facilis comb. nov., Acidovorax delafieldii comb. nov., and Acidovorax temperans sp. . Int J Syst Bacteriol [Internet].

1990 Oct 1;40(4):384–98. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-40-4-384

66. Saxena R, Dhakan DB, Mittal P, Waiker P, Chowdhury A, Ghatak A, et al. Metagenomic Analysis of Hot Springs in Central India Reveals Hydrocarbon Degrading Thermophiles and Pathways Essential for Survival in Extreme Environments. Front Microbiol [Internet]. 2017 Jan 5 [cited 2018 Feb 26];7:2123. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2016.02123/full

67. Padilla-Del Valle R, Morales-Vale LR, Ríos-Velázquez C. Unraveling the microbial and functional diversity of Coamo thermal spring in Puerto Rico using metagenomic library generation and shotgun sequencing. Genomics Data [Internet]. 2017 Mar;11:98–101. Available from: http://dx.doi.org/10.1016/j.gdata.2016.12.010

68. Kozubal MA, Macur RE, Jay ZJ, Beam JP, Malfatti SA, Tringe SG, et al. Microbial Iron Cycling in Acidic Geothermal Springs of Yellowstone National Park: Integrating Molecular Surveys, Geochemical Processes, and Isolation of Novel Fe-Active Microorganisms. Front Microbiol [Internet]. 2012;3(MAR):1–16. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2012.00109/abstract

69. Selvarajan R, Sibanda T, Tekere M. Thermophilic bacterial communities inhabiting the microbial mats of “indifferent” and chalybeate (iron-rich) thermal springs: Diversity and biotechnological analysis. Microbiologyopen [Internet]. 2018 Apr;7(2):e00560. Available from: http://doi.wiley.com/10.1002/mbo3.560

70. Heylen K, Lebbe L, De Vos P. Acidovorax caeni sp. nov., a denitrifying species with genetically diverse isolates from activated sludge. Int J Syst Evol Microbiol [Internet]. 2008 Jan 1;58(1):73–7. Available from: https://www.microbiologyresearch.org/content/journal/ijse

m/10.1099/ijs.0.65387-0

71. Raposo P, Viver T, Albuquerque L, Froufe H, Barroso C, Egas C, et al. Transfer of Meiothermus chliarophilus (Tenreiro et al.1995) Nobre et al. 1996, Meiothermus roseus Ming et al. 2016, Meiothermus terrae Yu et al. 2014 and Meiothermus timidus Pires et al. 2005, to Calidithermus gen. nov., as Calidithermus chliarophilus com. Int J Syst Evol Microbiol [Internet]. 2019 Apr 1;69(4):1060–9. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003270

72. Udaondo Z, Duque E, Ramos J-L. The pangenome of the genus Clostridium. Environ Microbiol [Internet]. 2017 Jul;19(7):2588–603. Available from: http://doi.wiley.com/10.1111/1462-2920.13732

73. Liu L, Jiao J-Y, Fang B-Z, Lv A-P, Ming Y-Z, Li M-M, et al. Isolation of Clostridium from Yunnan-Tibet hot springs and description of Clostridium thermarum sp. nov. with lignocellulosic ethanol production. Syst Appl Microbiol [Internet]. 2020 Sep;43(5):126104. Available from: https://doi.org/10.1016/j.syapm.2020.126104

74. Rainey FA, Donnison AM, Janssen PH, Saul D, Rodrigo A, Bergquist PL. Description of Caldicellulosiruptor saccharolyticus. FEMS Microbiol Lett. 1994;120:263–6.

75. Blumer-Schuette SE, Giannone RJ, Zurawski J V., Ozdemir I, Ma Q, Yin Y, et al. Caldicellulosiruptor Core and Pangenomes Reveal Determinants for Noncellulosomal Thermophilic Deconstruction of Plant Biomass. J Bacteriol [Internet]. 2012 Aug 1;194(15):4015–28. Available from: https://jb.asm.org/content/194/15/4015

76. Bellini MI, Gutiérrez L, Tarlera S, Scavino AF. Isolation and functional analysis of denitrifiers in an aquifer with high potential for denitrification. Syst Appl Microbiol [Internet]. 2013;36(7):505–16. Available from: http://dx.doi.org/10.1016/j.syapm.2013.07.001

77. Zeytun A, Sikorski J, Nolan M, Lapidus A, Lucas S, Han J, et al. Complete genome sequence of Hydrogenobacter thermophilus type strain (TK-6T). Stand Genomic Sci [Internet]. 2011 Apr 29;4(2):131–43. Available from: http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.1463589

78. Nishihara A, Thiel V, Matsuura K, McGlynn SE, Haruta S. Phylogenetic Diversity of Nitrogenase Reductase Genes and Possible Nitrogen-Fixing Bacteria in Thermophilic Chemosynthetic Microbial Communities in Nakabusa Hot Springs. Microbes Environ [Internet]. 2018;33(4):357–65. Available from: https://www.jstage.jst.go.jp/article/jsme2/33/4/33_ME18030/_article

79. Hamilton TL, Koonce E, Howells A, Havig JR, Jewell T, de la Torre JR, et al. Competition for ammonia influences the structure of chemotrophic communities in geothermal springs. Appl Environ Microbiol. 2014;80(2):653–61.

80. Pester M, Rattei T, Flechl S, Gröngröft A, Richter A, Overmann J, et al. AmoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. Environ Microbiol. 2012;14(2):525–39.

81. Zhang CL, Ye Q, Huang Z, Li WJ, Chen J, Song Z, et al. Global occurrence of archaeal amoA genes in terrestrial hot springs. Appl Environ Microbiol. 2008;74(20):6417–26.

82. Reigstad LJ, Richter A, Daims H, Urich T, Schwark L, Schleper C. Nitrification in terrestrial hot springs of Iceland and Kamchatka. FEMS Microbiol Ecol. 2008;64(2):167–74.

83. Revsbech NP, Trampe E, Lichtenberg M, Ward DM, Kühl M. In Situ Hydrogen Dynamics in a Hot Spring Microbial Mat during a Diel Cycle. Spormann AM, editor. Appl Environ Microbiol [Internet]. 2016 Jul 15;82(14):4209–17. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.00710-16

84. Otaki H, Everroad RC, Matsuura K, Haruta S. Production and Consumption of Hydrogen in Hot Spring Microbial Mats Dominated by a Filamentous Anoxygenic Photosynthetic Bacterium. Microbes Environ [Internet]. 2012;27(3):293–9. Available from: https://www.jstage.jst.go.jp/article/jsme2/27/3/27_ME11348/_article

85. Lindsay MR, Colman DR, Amenabar MJ, Fristad KE, Fecteau KM, Debes R V., et al. Probing the geological source and biological fate of hydrogen in Yellowstone hot springs. Environ Microbiol. 2019;21(10):3816–30.

86. Fenchel T, King GM, Blackburn TH. Bacterial Metabolism. In: Bacterial Biogeochemistry [Internet]. Elsevier; 2012. p. 1–34. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780124158368000013

87. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol [Internet]. 2015 Apr 2;13(4):217–29. Available from: http://dx.doi.org/10.1038/nrmicro3400

88. Yang S, Winkel M, Wagner D, Liebner S. Community structure of rare methanogenic archaea: insight from a single functional group. FEMS Microbiol Ecol [Internet]. 2017 Nov 1;93(11):1–10. Available from: https://academic.oup.com/femsec/article/doi/10.1093/femsec/fix126/4331631

89. Boyd ES, Leavitt WD, Geesey GG. CO2 Uptake and Fixation by a Thermoacidophilic Microbial Community Attached to Precipitated Sulfur in a Geothermal Spring. Appl Environ Microbiol [Internet]. 2009 Jul 1;75(13):4289–96. Available from: https://aem.asm.org/content/75/13/4289

90. Bennett AC, Murugapiran SK, Hamilton TL. Temperature impacts community structure and function of phototrophic Chloroflexi and Cyanobacteria in two alkaline hot springs in Yellowstone National Park. Environ Microbiol Rep.

2020;12(5):503–13.

91. Maeder DL, Anderson I, Brettin TS, Bruce DC, Gilna P, Han CS, et al. The Methanosarcina barkeri genome: Comparative analysis with Methanosarcina acetivorans and Methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes. J Bacteriol. 2006;188(22):7922–31.

92. Huber H, Thomm M, König H, Thies G, Stetter KO. Methanococcus thermolithotrophicus, a novel thermophilic lithotrophic methanogen. Arch Microbiol. 1982;132(1):47–50.

93. Jones WJ, Leigh JA, Mayer F, Woese CR, Wolfe RS. Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch Microbiol. 1983;136(4):254–61.

94. Smith KS, Ingram-Smith C. Methanosaeta, the forgotten methanogen? Trends Microbiol. 2007;15(4):150–5.

95. Liu C, Sun D, Zhao Z, Dang Y, Holmes DE. Methanothrix enhances biogas upgrading in microbial electrolysis cell via direct electron transfer. Bioresour Technol [Internet]. 2019;291(July):121877. Available from: https://doi.org/10.1016/j.biortech.2019.121877

96. Wasserfallen A, Nölling J, Pfister P, Reeve J, De Macario EC. Phylogenetic analysis of 18 thermophilic Methanobacterium isolates supports the proposals to create a new genus, Methanothermobacter gen. nov., and to reclassify several isolates in three species, Methanothermobacter thermautotrophicus comb. nov., Methano. Int J Syst Evol Microbiol. 2000;50(1):43–53.

97. Stetter KO. Archaeoglobus fulgidus gen. nov., sp. nov.: a New Taxon of Extremely Thermophilic Archaebacteria. Syst Appl Microbiol [Internet]. 1988 Mar;10(2):172–3. Available from: http://dx.doi.org/10.1016/S0723-2020(88)80032-8

98. Callac N, Oger P, Lesongeur F, Rattray JE, Vannier P, Michoud

G, et al. Pyrococcus kukulkanii sp. nov., a hyperthermophilic, piezophilic archaeon isolated from a deep-sea hydrothermal vent. Int J Syst Evol Microbiol. 2016;66(8):3142–9.

99. Dalmasso C, Oger P, Selva G, Courtine D, L'Haridon S, Garlaschelli A, et al. Thermococcus piezophilus sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman Rise. Syst Appl Microbiol [Internet]. 2016;39(7):440–4. Available from: http://dx.doi.org/10.1016/j.syapm.2016.08.003

100. Huber G, Spinnler C, Gambacorta A, Stetter KO. Metallosphaera sedula gen, and sp. nov. Represents a New Genus of Aerobic, Metal-Mobilizing, Thermoacidophilic Archaebacteria. Syst Appl Microbiol [Internet]. 1989;12(1):38–47. Available from: http://dx.doi.org/10.1016/S0723-2020(89)80038-4

101. Colman DR, Lindsay MR, Boyd ES. Mixing of meteoric and geothermal fluids supports hyperdiverse chemosynthetic hydrothermal communities. Nat Commun [Internet]. 2019 Dec 8;10(1):681. Available from: http://dx.doi.org/10.1038/s41467-019-08499-1

102. Anda D, Makk J, Krett G, Jurecska L, Márialigeti K, Mádl-Szőnyi J, et al. Thermophilic prokaryotic communities inhabiting the biofilm and well water of a thermal karst system located in Budapest (Hungary). Extremophiles [Internet]. 2015 Jul;19(4):787–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25952671

103. Castelán-Sánchez HG, Meza-Rodríguez PM, Carrillo E, Ríos-Vázquez DI, Liñan-Torres A, Batista-García RA, et al. The Microbial Composition in Circumneutral Thermal Springs from Chignahuapan, Puebla, Mexico Reveals the Presence of Particular Sulfur-Oxidizing Bacterial and Viral Communities. Microorganisms [Internet]. 2020 Oct 29;8(11):1677. Available from: https://www.mdpi.com/2076-2607/8/11/1677

104. Everroad RC, Otaki H, Matsuura K, Haruta S. Diversification of Bacterial Community Composition along a Temperature Gradient at a Thermal Spring. Microbes Environ [Internet]. 2012;27(4):374–81. Available from: https://www.jstage.jst.go.jp/article/jsme2/27/4/27_ME11350/_article

105. Hamamura N, Meneghin J, Reysenbach A-L. Comparative community gene expression analysis of Aquificales-dominated geothermal springs. Environ Microbiol [Internet]. 2013 Apr;15(4):1226–37. Available from: http://doi.wiley.com/10.1111/1462-2920.12061

106. Kubo K, Knittel K, Amann R, Fukui M, Matsuura K. Sulfur-metabolizing bacterial populations in microbial mats of the Nakabusa hot spring, Japan. Syst Appl Microbiol [Internet]. 2011 Jun;34(4):293–302. Available from: http://dx.doi.org/10.1016/j.syapm.2010.12.002

CHAPTER 4

Sequence analysis, cloning, purification and characterization of a novel thermostable endoglucanase and sequence analysis and cloning of a predicted β-xylosidase found by sequence-based metagenomics

Abstract

In the present work we analysed the metagenome of a hot spring located in the northwest of Spain (next to the city of Ourense in the Galicia region) in search for novel thermostable enzymes involved in plant biomass degradation. Sequence analysis revealed two promising novel predicted proteins, a β-xylosidase (Xyl7992) and an endoglucanase (Cel776). Predictions on protein structure and conserved amino acid sequences were conducted, and expression in a heterologous system with *Escherichia coli* as the host. Xyl7992 was not expressed as a soluble protein in the system tested, probably due to inclusion bodies formation. Cel776 was correctly expressed and purified, and its biochemical characterization revealed a biotechnologically promising enzyme. Cel776 exhibited high thermostability maintaining its activity for 24 hours at 60°C, with an optimal temperature of 80°C and optimal pH of 5. Moreover, its substrate specificity allowed the degradation of both cellulose and xylan. It was more active in the presence of metal ions including calcium and magnesium and was not severely affected by detergents other than SDS.

Introduction

Sequence-based metagenomics

The development of high-throughput Next Generation Sequencing (NGS) technologies, and more specifically of shotgun metagenomics, has allowed to directly sequence the large number of genomes present in environmental samples, using multiple templates in

parallel without targeting specific genes (1,2). The method relies on annotated data (reference genomes and gene and protein databases), as gene search and functional annotation using bioinformatic tools is based on alignment and homology to deposited sequences (3,4). Shotgun metagenomics has a number of advantages over other methods for bioprospecting enzymes in metagenomes, as it can retrieve more divergent sequences than that of consensus genes obtained through PCR-based targeted sequencing (2), favours the detection of less abundant genes in complex microbiomes (2) and provides a scalable method that bypasses limitations in the detection of novel gene products due to the heterologous host expression systems and the need for high-throughput screening methods that typically hinder functional metagenomics (3). Moreover, the sequencing costs have progressively and significantly lowered, making the technology more accessible to all research levels (1,2), and an ever-growing repertory of bioinformatic tools and databases (4) is constantly being updated providing researchers the means to analyse the vast amount of data contained in metagenomes. Nevertheless, sequence based metagenomic studies also face their own challenges, including quality and length of the reads generated, cost and labour of the methodology, and amplification bias and other artifacts like chimeric sequences and secondary structures (1). It is also important to consider that the success of the method heavily relies on the quality of the database annotation and is limited to find somewhat similar sequences in known protein families (3). Lastly, functional characterization of predicted gene products is still necessary in order to confirm the results of the *in silico* analysis (5).

Of all the available NGS platforms developed, the platform from Illumina (originally Solexa) is by far the most widespread (1), but other strategies are also available. Illumina requires the construction of a library and uses a PCR amplification step before the sequencing. In short, a library is constructed with the metagenomic DNA and adapters are ligated to these library fragments. The samples are loaded in the flow cell of the sequencer machine (each flow cell

divided in eight different lanes), in which the interior surface of the flow cell is covered with covalently bound sequences complementary to said adapters. The next steps consist in heating and cooling so that hybridization between the two takes place. Then, incubation with amplification reactants including the polymerase allows to generate millions of clusters of library fragments. For sequencing, the clusters are supplied with polymerase and nucleotides labelled with fluorescence and modified to stop amplification in that flow cycle. After measurement of fluorescence to identify the nucleotide, the flow cell is chemically treated to remove the fluorescent group and unblock the amplification for the next flow cycle. Computation of the sequence on each cluster allows the generation of raw sequencing data with a quality score for each base, in Illumina 1.3+ FASTQ format (6).

After sequencing, a wide variety of bioinformatic tools allows for a correct interpretation of the data. In shotgun metagenomics, when the objective is to bioprospect for novel gene products, the short reads can be linked together into a bigger sequence (contig). Due to the high computational demand of methods based on overlapping reads, many de novo assemblers use instead a de Bruijn graph approach (4). Predicted Open Reading Frames (ORF) within the reconstructed contigs can then be submitted and aligned to known sequences deposited in annotated databases, and thus functions of ecological or biotechnological interest can be identified. As mentioned, a functional characterization is needed at the end of the process to confirm the predicted activities.

Plant-polymer degrading enzymes: endoglucanases and β-xylosidases

Cellulose is the most abundant polymer on Earth and the main component of plant biomass. It normally presents itself in fibres associated to other biopolymers, namely hemicelluloses and lignin, in a complex structural matrix in plant cell walls (7). This complex matrix limits the extent and rate of utilization of plant biomass,

usually requiring harsh pre-treatments and the action of multiple enzymes to perform the full breakdown of the structure (8). Although multiple enzymatic activities are responsible of the conversion of cellulose into simpler molecules, collectively known as cellulases, β-1-4-endoglucanases (EC 3.2.1.4) are especially important as they act on the cellulose chain cleaving internal glycosidic bonds and releasing oligosaccharides of different lengths. These are further hydrolysed by other cellulolytic enzymes like non-reducing end cellobiohydrolases (EC 3.2.1.91), reducing-end cellobiohydrolases (EC 3.2.1.176) and cellodextrinases (EC 3.2.1.74), and finally converted in glucose by β-glucosidases (EC 3.2.1.21) reducing the product inhibition of all the enzymes mentioned before (9). Endoglucanases have been classified along with other enzymes based on sequence similarity in the Carbohydrate Active Enzymes (CAZy) Database (10) (http://www.cazy.org/) in 12 Glycosyl Hydrolase families: GH5, GH6, GH7, GH8, GH9, GH12, GH44, GH45, GH48, GH51, GH74, and GH124. Lists of thermophilic cellulases found by metagenomics (5) and characterized thermophilic cellulases (9) are available, and novel cellulases found following this strategy continue to be discovered (11,12) showing the interest for these biocatalysts. Moreover, the discovery of multifunctional enzymes that can act on more than one biomass polymer (13–15), and the characterization of microbial consortiums that produce multiple lignocellulolytic enzymes (16) are also recent research focuses.

Endoglucanases have multiple biotechnological applications, and are especially important in the valorisation of agro-industrial by-products (8) and in the production of biofuels combined with β-glucosidases to produce glucose later fermented into (bio)ethanol (17), as these two applications are directly targeting environmental challenges. In the textile industry there are several processes like biostoning (to give a wash-down look on cotton clothes) and biopolishing (softening and brightening of cotton surfaces) that remove cellulose fibres and replace more harsh treatments. Similarly, detergent formulations can include endoglucanases, also brightening and softening cotton

fabrics (18). They are also employed in the food and brewing industries, improving digestibility of food and decreasing viscosity, and increasing fermentable compounds for the elaboration of alcoholic drinks. These properties have also been exploited in the animal feed industry, enhancing digestibility and nutrient bioavailability. The pulp and paper industry has many uses for endoglucanases including biopulping, treatment of pulp wastes, deinking and removal of pollutants from paper. Lastly, other reported uses include waste management, improvement of soils for agriculture and extraction of bioactive compounds, pigments and oils from plants (9).

After cellulose, hemicelluloses (polymers of pentoses, hexoses and sugar acids) are the second most abundant polymers on Earth and their main component is xylan (19,20) in the case of hardwoods, whereas glucomannans are main components in softwood hemicelluloses (21). Xylan is composed of a backbone of β-(1,4)-D-xylose subunits with a variety of side chains depending on its origin. The composition of the side chains can be used to classify xylan into different families (22,23), including arabinoxylans (α-L-arabinofuranosyl subunits); glucuronoxylans (α-D-glucuronic acid and its 4-O-methyl ether derivative); glucuronoarabinoxylans (α-L-arabinose and α-D-glucuronic acid); and galactoglucuronoarabinoxylans (β-D-galactopyranosyl residues on complex oligosaccharide side chains), but side chains may also include other units like acetic, ferulic and *p*-coumaric acids (21,24). This structural complexity explains the need for multiple enzymes acting in synergy for its complete breakdown (21,25). Endoxylanases (EC 3.2.1.8) act on the main chain polysaccharide releasing xylo-oligosaccharides that are then further hydrolyzed from the non-reducing ends to xylose by the action of β-xylosidases (EC 3.2.1.37) (25). As xylo-oligosaccharides are strong inhibitors of hemicellulose-acting enzymes, β-xylosidases play a crucial role in the whole xylan-degrading process, but are in turn inhibited by D-xylose, L-arabinose and other monosaccharides (24). β-xylosidases have been identified

in eleven Glucosyl Hydrolase (GH) families in the CaZy database (GH1, 3, 5, 30, 39, 43, 51, 53, 54, 116 and 120) but GH1, GH54 and GH116 proteins annotated as β-xylosidases have their activity on natural substrates yet to be observed (24). Several thermophilic β-xylosidases have been reported, with optimum temperatures ranging from 50 to 90°C (25). Moreover, a number of β-xylosidases have been obtained through metagenomic strategies, mostly from the microbiota found in the rumen of ruminants like cows (26), goats (27) and yaks (28–30), but also from environmental samples like deep-sea sediments (31) and thermophilic habitats like compost (32) and hot springs (33).

β-xylosidases have found a variety of applications, for example in the food industry where they are often combined with other xylan-acting enzymes, including juice clarification as one of the maceration enzymes used; reduction of viscosity in wine clarification; solubilization of arabinoxylans to reduce beer viscosity and turbidity; enhancement of wheat flour for bakery improving the dough handling, as well as its volume and half-life; and extraction of starch, plant oils and coffee (25,34). More uses include the processing of cellulose pulp prior to the whitening process in the pulp and paper industry to reduce the use of chemical compounds, and improved digestibility of animal feeds to reduce their costs (34). Its main product, xylose, can be used to prepare Maillard flavours, and is a precursor in the synthesis of products like xylitol and succinic acid for the food industry (25,34). Xylose is reportedly one of the most widely used sugars for the production of bioethanol (ranging from 5 to 20% of the sugars employed) (34). In xylan saccharification, β-xylosidases are responsible for most of the glycosidic bond cleavages (24) and enhance the whole process by degrading xylooligosaccharides that inhibit other enzymes involved (24). The antitumoral diterpenoid drug paclitaxel (taxol) is isolated from plant species of the genus *Taxus* but in a complex and low yielding process. More abundant analogues can be used after treatment with β-xylosidases that remove xylose for the semisynthesis of this valuable pharmaceutical

product (35). Xylooligosaccharydes can be used for their prebiotic properties in the nutraceutical industry. Although β-xylosidase free preparations are desired for the breakdown of xylan, the reverse hydrolysis of xylose using these enzymes at high substrate concentrations have been deemed a promising process as an alternative for the production of xylooligosaccharydes (23,34).

As mentioned, many of these applications are benefited from the use of combinations of various enzymes (enzyme cocktails) as the plant biomass is composed of a complex matrix of cellulose, hemicelluloses and lignin (34). Industrial processes like biofuel production, food processing, treatments in the pulp and paper industry and production of nutraceuticals have been explored in this context, among others (21,36).

In the present work, we conducted a metagenomic sequence-based screening for genes that were predicted to encode putative β-1-4-endoglucanases and β-xylosidases with the premise that the metagenome from a hot spring was expected to include thermophilic variants of these biocatalysts. A purification and biochemical characterization followed this screening process to confirm the *in silico* analysis with functional assays.

Materials and Methods

Metagenomic DNA from a hot spring in the Galicia region of Spain was obtained as described in Chapter 1 of this work (Muiño da Veiga hot spring). Sequencing was performed by an external service provided by Health in Code (A Coruña, Spain) as described in Chapter 3 of this work, using an Illumina Hi-Seq platform with a 2x100pb sequence length. Quality filtering and read merging of the raw FASTQ files were performed as outlined in the Materials and Methods section in Chapter 3.

Assembly of short reads into contigs using a de-novo assembler and decontamination

Quality-filtered merged reads in FASTA format and singletons (sequences that could not be merged from the sequencing pair) were all used as input for an assembly using a de-novo de Brujin graph based algorithm IDBA-UD (37) on default settings. In short, the assembler uses de Brujin graphs in which the vertices are unique length-k substrings (k-mers) and edges connect them only if both k-mers are present consecutively in a read. The resulting paths in the graph represent longer sequences than the original reads called contigs. The process is iterative, starting from a smaller k and ending in a large value, using the contigs generated in each iteration as the new reads input. A process of contig error detection and removal is also involved, in which IDBA-UD introduces a local assembly with paired-end information using a variable threshold, to solve problems of highly uneven sequence depths between reads, as well as commonly addressed problems such as the gap problem (missing k-mers due to regions with low sequencing depth using large k values) and the branching problem (introduction of branches in the graph due to repeat regions and erroneous reads using small k values).

The resulting contigs were decontaminated from sequences aligning to the human genome (contaminating sequences) using the Deconseq algorithm (38) with the NCBI GRCh38 data (RefSeq ID 884148) used as the database. Sequences aligning to the human genome were then removed from the analysis.

Sequence upload to MG-RAST web service and identification of putative endoglucanase genes

The FASTA format file containing the contigs sequence data was uploaded to the MG-RAST (39) web service. The pipeline options for the upload were as follows: dereplication “yes”, screening “none”, length filtering “no”, ambiguous base filtering yes”, maximum ambiguous basepairs “5”. A brief description of MG-RAST automated pipeline for data upload can be found in the Materials and Methods section of Chapter 3 of this work.

An in-depth analysis was conducted to retrieve contig sequences containing genes of interest. The analysis tool from MG-RAST was used with the following parameters: e-value “5”, %-ident “60”, length “15”, min. abundance “1”, method “representative hit”. The database for gene product annotation employed was SEED Subsystems (40). Annotated sequences at the function level were search with the terms “Endoglucanase (EC 3.2.1.4)” and “Beta-xylosidase (EC 3.2.1.37)”. The resulting contigs containing putative genes coding for the selected activities were analyzed using ExPASy Translate Tool (41) to identify ORFs. The selected ORFs where then analyzed using NCBI BLASTp algorithm (42) on the default settings (protein-protein BLAST with BLOSUM62 matrix, cost of gap existence = 11 and of gap extension = 1) with the Non-redundant protein sequences database (nr). Selected putative proteins were structurally modelled using the SWISS-MODEL (43) web tool with the best matching template. Protein models were visualized with the PyMol software (Schrödinger LCC, USA) (44). Protein parameters were predicted with the PROTPARAM algorithm from ExPASy (41). Multiple sequence alignments used EMBL-EBI Clustal Omega (45) and sequences obtained from the UniProt Knowledgebase database (46). Prediction of signal peptide features was performed using the SignalP 0.5 web service (47).

Synthesis and cloning of a DNA fragment containing a putative β-xylosidase gene

The following sequence was ordered for gene synthesis:

>Xyl7992_gateway_attB

GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**atg**gcaaatttcgggaggacgca
gatgcggatgggatgggcgatgaaggtgggggtggcagcggcggtgctggggctgccatcgg
cagcaagcgcgggcggctccgcggcccggttcgactatcttgcctacgaatcgctcgagcctgc
ccagccgctgggcgagggcacttaccgcaatccggtgatcccgggctttcaccccgatccttca
atcgtgcgcgtgggcaaggatttctatgccgtcacctccaccttcagctggtttccaggcctgccg
atcttccacagcactgatctcgtcaactggcgggtgattggcaatgccattgatcggccgggcca
gctcgatttcagcgggctgggcaccaatcgcgggctctttgccccggcgatcacccatcatcagg
gcaggttctggatcgtcaacacctgcatcgattgcggcggcaatttcgtcatcaccgccgacgat
ccggccgggccgtggagcgatcccaaatggctcgatttcggcgggatcgacccgtcgctgttctt
cgatgcggatggcagcgcctggatcgtctataatgatgcgcccccggttccgccgcgctatgag
ggccaccgcgccctgtggctccagcgcttcgatcccgaggccatgcaggtgctgcccgagcgca
ccctgctggtcgatggcggggtcgatcccaaggccaacccgatctgggccgaagggccgcata
tctacaaggtgggcgggtggtattatctgctcgcggccgagggcggcacggcagaccagcattc
gcagacgatctaccgcgcgcgcaagatcaccgggccctatgtgccggggccgttcaacccgat
cctcacccagcgcgatctgccgcccaaccgtcccgaccgcatcgaggcgaccggccatgccgat
attgtccagctcgatgatggcacgtggtgggggctgttccttgcgacccggccctttgccgggca
atcgacgctgatggggcgcgagaccttcttgctgccgctgcgctggaaggatggctggccgcgc
tttctcgatcgcggcgagccggtgccgctggtgctggagcgcccgcgcctgccgcagtccgcgc
cttcggcatgggacagctggcgcgaggagttcgatggtccgctcggcccggaatggatcggcct
gcgcacccccgggccgcgtgcgctgatcgcaacggatgacggggcgctgcgggtcgcggcgg
gcagcgccgcggcgggcagcctgggcaagccggccttcatcggccggcgcttgcgccaccacc
gcgccgcctttgtcacccggttcgccttcgcgcccgagcggcgggggggattttgccgggttgctc
gccttcatggacgagcaccactttctcgccgccggcaaggagggcgcgcgcctggtggtgcgtc
tgcgcaaggccgcgggcgagaggcgcaatgcggcaagcgagaaccccgcgtcctctgcgatg
tccatccacgagatcgagcaggcctatgccaagaccgagatcgccatcgcgattgccgaaggc
gttttgcccgacgaccccgatttaccgctcatcgcccggtcggtgagggcctggtgg**tga**GACC
CAGCTTTCTTGTACAAAGTGGTCCCC

The adapter sequences (*att*B sites) for the Gateway recombination vector system are in capital letters, whereas the start and stop codons are in bold.

The synthesis was ordered to ThermoScientific gene synthesis service. The gene was provided in the pMA vector (ampicillin resistance, ColE1 origin of replication, T7 promoter). Adapter sequences (*att*B sites) were added to both sequence ends to allow subcloning in the Gateway system (Invitrogen).

For subcloning the Gateway Technology with Clonase II (Invitrogen) kit was used, following the manufacturer protocol. For the BP recombination reaction, the cloned β-xylosidase insert and pMA vector were linearized using the *Sca*I restriction enzyme (Roche, Switzerland) which produces one cut on the vector and none in the insert sequence. The restriction reaction was carried out in SureCut Buffer H (Roche, Switzerland) for 1 hour at 37°C. The linearized DNA was precipitated using 0.1 volumes of 3M sodium acetate and then 2.5 volumes of 100% ethanol. After centrifugation at maximum speed the pelleted DNA was washed twice with 70% ethanol. The DNA was then resuspended in TE Buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA). The *att*P sites containing vector pDNOR221 was used in the BP reaction as donor vector to generate an entry vector (*att*L site containing vector generated by the recombination of *att*B and *att*P sites). An e equimolar amount of linearized DNA and donor vector was employed, and the reaction was carried out using TE Buffer pH 8 supplied in the kit and the Gateway BP Clonase II enzyme mix with an incubation at 25°C for 1 hours. The reaction was stopped adding Proteinase K solution to the reaction mix and incubating at 37°C for 10 minutes. This reaction mix was used to transform chemically competent *E. coli* One Shot™ OmniMAX™ 2 $T1^R$ (F´ {*pro*AB *lac*I^q *lac*ZΔM15 *Tn*10(Tet^R) Δ(*ccd*AB)} *mcr*A Δ(*mrr hsd*RMS-*mcr*BC) Φ 80(*lac*Z)ΔM15 Δ(*lac*ZYA-*arg*F)U169 *end*A1 *rec*A1 *sup*E44 *thi*-1 *gyr*A96 *rel*A1 *ton*A *pan*D) (Invitrogen, USA). A vial of chemically competent cells (50µL) was thawed on ice and 1µL of BP recombination reaction was added and mixed. The mixture was

incubated for 30 minutes on ice, heat-shocked at 42°C for 30s and cooled on ice for 2 minutes. For recovery, cells were incubated with shaking at 37°C for 1 hour with 500µL LB media. Excess media was removed by centrifugation and 100µL were plated on LB media containing kanamycin for positive selection of transformant recombinant clones with an overnight period for growth. Correct recombination was tested by sequencing both ends of the DNA fragment using the M13 sequencing primer pair, with the sequencing service provided by the Molecular Biology Unit from the Research Support Services of Universidade da Coruña (Spain). Plasmid DNA was obtained by performing a miniprep using the NZYMiniprep kit (NZYTECH) on a 10mL culture from a single colony growing in the transformation plate (up to 4 colonies were tested). The plasmid harbouring the insert (flanked by *att*L sites) was used for the next step of the protocol of LR recombination as supercoiled DNA. The destination vector was pDEST527 (which contains *att*R sites) and was also used as supercoiled DNA. The reaction was carried out in TE Buffer pH 8.0 with the Gateway LR Clonase II enzyme mix, with an incubation at 25°C for 1 hours, and it was stopped by adding 1µL of Proteinase K and incubating for 10 minutes at 37°C. Transformation was carried out using 1µL of the LR reaction and a vial (50µL) of chemically competent *E. coli* One Shot™ OmniMAX™ 2 $T1^R$ following the same protocol for the thermal shock, recovery and plating as with the BP reaction. The LB media used for positive selection contained ampicillin instead of kanamycin. Correct insertion was verified plating on LB media containing both ampicillin and chloramphenicol where negative selection occurred as chloramphenicol resistance is lost if a LR recombination takes place, and by sequencing both ends using the T7 sequencing primer pair. DNA for sequencing was obtained from a fresh 10mL culture from a single colony from the transformation plate using the NZYMiniprep kit (NZYTech, Portugal). The resulting expression vector pDEST521 contained the insert with the gene for the putative β-xylosidase flanked by *att*B sites and with a His(x6) tag to enable easier purification and an inducible T7 promoter for controlled gene expression. An aliquot of the DNA used for

sequencing was also used to transform chemically competent *E. coli* T7 Express (New England Biolabs, USA) (*fhu*A2 *lac*Z::T7 *gene*1 [lon] *omp*T *gal sul*A11 R(*mcr*-73::*miniTn*10--*Tet*^{S})2 [dcm] R(*zgb*-210::*Tn*10--*Tet*^{S}) *end*A1 Δ(*mcr*C-*mrr*)114::IS10). Transformation was carried out as described before changing the heat-shock step to 10 seconds at 42°C and the incubation on ice to 5 minutes as recommended by the manufacturer. LB medium for positive selection contained ampicillin.

Synthesis and cloning of a DNA fragment containing a putative endoglucanase gene

The following sequence was ordered for gene synthesis:

>Cel776_gateway_attB

GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**atg**ctaacaagttgcagagaag
aaaaaactgtcaaaaaaagcgcatttgaatacaacaatctaatcggtcatggtgtgaatatgg
gaaatgcacttgaagcacctgtagaaggttcttggggcgtattcatcgatgatgagtatttcaag
ataataaaacaacggggatttaattcggtgagaatacctataagatggtctgcacatatctcag
atgattacacaatagataaaaatttcatggaaagagtcgttcatgtggttgataaagcgttaga
aaatggtttagtaacaattataaacacacaccatttcgaagaactttaccaagatccagacaaa
tatggtcctgtgcttgtaaagatatgggaacagatagcagaaaaattcaaaaactatccagaca
cgctctttttttgaaatatacaatgagccagcccaacaattaactcctgacaaatggaataaaata
taccctgaagcgttaaaagtgatacgaaaaatcaatcccaaaagaatagttataatcgatgtac
cgaattgggcacattactctgctgtgaaggatctcaaattaccagacgatgagaatataatagt
atccttccattactatgaaccattcaatttcactcaccaaggtgcagaatgggtcaccccacaac
taccagtaggagtaaaatgggaaggtaaagactgggaagtggagcagatcaaaaaccacttc
aaatacgttagtgattgggcgaagaaaaacaatgtaccgatattccttggcgaattcggagctt
actcaaaagcggatatggattcaagagtaaaatggacggaaacagtaagaaaaaccgccga
agaatttggattttcactcgcatactgggaattttgcgcaggattcggtatctacgatagaacatc
cggaacgtgggtagaaccactggcaacagcagcgcttgggaag**tag**GACCCAGCTTTCT
TGTACAAAGTGGTCCCC

The adapter sequences (*att*B sites) for the Gateway recombination vector system are in capital letters, whereas the start and stop codons are in bold.

The synthesis was ordered to ThermoScientific gene synthesis service. The gene was provided as a DNA fragment readily available for cloning. Adapter sequences (*att*B sites) were added to both sequence ends to allow subcloning in the Gateway system (Invitrogen). The Gateway Technology with Clonase II (Invitrogen) kit was used for subcloning as described for the putative β-xylosidase Xyl7992 gene in the previous section, following the same steps and using the same materials.

Purification of a putative β-xylosidase and activity test

A fresh culture *of E. coli* T7 Express harbouring the pDEST527 vector and the Xyl7992 gene was started from an overnight inoculum started from a single colony, with all incubation temperatures at 37°C. When OD_{600} reached between 0.8 to 1.0 units the culture was induced with 0.4mM (final concentration) IPTG for 2 hours at 37°C. Release of intracellular content was achieved using the NZYLysis kit (NZYTech, Portugal) following the manufacturer protocol. Alternatively, lysis was performed using a VCX130 Vibra-Cell sonicator (SONICS & MATERIALS INC., USA) with a setting of 70% amplitude, 3 minutes active time and pulses of 3 seconds ON and 7 seconds OFF.

A method for quantification of β-xylosidase activity was adopted using pNP-xylopyranoside (pNPX) as the substrate (48–50). 10µL of enzyme extract were incubated with 10µL of 20mM pNPX, 180µL of 0.5mM Sodium Phosphate buffer pH 6.0 at the reaction temperature for 30 minutes. Absorbance was read in at wavelength 405nm.

Further testing was performed using varying concentrations of IPTG (0.05, 0.1 and 1mM) for 8 hours, with the sonication method used for obtaining the crude extract. SDS-PAGE was performed loading both the crude extracts and the precipitated pellets (51).

Purification of a novel endoglucanase

Fresh cultures were stablished using an overnight inoculum at 37°C from a single colony of *E. coli* T7 Express harbouring the pDEST527 vector and the cloned endoglucanase gene Cel776. The cells were harvested using a refrigerated centrifuge (4°C) at maximum speed and resuspended in resuspension buffer 100mM Sodium Acetate pH 5, 100mM NaCl and 1mM DTT. Cell lysates were obtained using the sonicator with a setting of 70% amplitude, 3 minutes active time and pulses of 3 seconds ON and 7seconds OFF. Cell debris was precipitated in a refrigerated (4°C) centrifuge at maximum speed, and the recovered supernatant was used as a crude enzyme extract. This extract was further purified using the following filtered and sonicator-degasified buffers: wash buffer 100mM sodium acetate pH 5, 500mM NaCl and 25mM imidazol; elution buffer 100mM sodium acetate pH 5, 100mM NaCl and 300mM imidazol. The extract was loaded in a niquel-sepharose resin-filled column HisTrap 5mL (Cytiva, USA) and washed with 5 volumes (25mL) of wash buffer. A manual run with a peristaltic pump P-1 (Cytiva, USA) was conducted using a flow rate setting of 3.0 and pressure setting at 0.5. The elution buffer was diluted to obtain 10% elution buffer (30mM imidazol) and 50% elution buffer (150mM imidazol). Increasingly concentrations of elution buffer were loaded in the column to obtain the different elution fractions. Protein concentration determination using the Bradford method (52), enzymatic activity tests (described in the next section) and SDS-PAGE were all conducted on the fractions to monitor the purification process. A western-blot was also performed using anti-his-tag antibodies to detect the purified protein. The developed polyacrylamide gel was cut and placed in contact with the western-blot membrane, and introduced in a sandwich like fashion between two filter papers. For the bands transfer to the membrane, a cooled transfer buffer was employed with the following composition: Tris 25mM pH 8.3, glycine 192mM, 20% (v/v) methanol. The settings for the transfer were 300mA for 60 minutes. After the transfer was completed, the membrane was incubated with a

blocking solution consisting in Tris buffered saline (TBS), 0.1% (v/v) Tween 20 and 5% (w/v) BSA for 2 hours in cold. After blocking, the membrane was incubated with a TBS solution containing 0.1% (v/v) Tween 20, 5% (w/v) BSA and HRP-conjugated 6XHis, His-Tag Monoclonal antibody (Proteintech, USA) in cold. To develop the chemiluminescence signal the components from Pierce ECL Plus Western Blotting Substrate kit (Fisher Thermo Scientific, USA) were employed, following the manufacturer protocol. Pictures were taken using a Chemidoc MP Imaging System (Bio-rad, USA), with a chemiluminescence setting for "high sensitivity" and exposition time "120 seconds" (several photos were taken in-between), and a colorimetric photo was merged to include the protein ladder. The purified protein was concentrated to its final volume for activity assays using a Pierce Concentrator column with 10K Molecular Weight cut-off (Fisher Thermo Scientific, USA).

Biochemical characterization of a novel endoglucanase

Endoglucanase activity was tested using the reducing sugars method with dinitrosalicylic acid (DNS) using soluble carboxymethylcellulose (CMC) as the substrate (53). 5µL of the purified enzyme were mixed with 45µL of reaction buffer and 50µL of 1% CMC. Reaction time was 30 minutes, and the reaction was stopped by adding 150µL DNS and incubating for 5 minutes at 100°C. Then, 1mL ice-cold milliQ water were added and the tubes were placed on ice until the absorbance was read at wavelength 540nm. All conditions were assayed in triplicates and blanks were also assayed where the volume of purified enzyme was replaced with the same volume of milliQ water. An enzymatic unit is defined as the amount of enzyme that releases a micromol of glucose per minute in the reaction conditions. A glucose standard was prepared to perform a linear regression of the relationship between absorbance and glucose concentration.

The optimal temperature was evaluated at pH 5 (reaction buffer was 100mM sodium acetate pH 5 at 60°C) in the range between 30 and 90°C with a 10°C interval. The optimal pH was evaluated at 60°C in a

range between 4 and 9 with an interval of 1. Reaction buffers for the pH intervals were 100mM Sodium Acetate buffer for the range 4-6, 100mM Sodium Phosphate buffer for the range 6-8 and 100mM Tris HCl buffer for the range 8-10. Thermostability was assayed at optimal pH and temperature after incubation of the enzyme at 60°C, 70°C, 80°C and 90°C for a set amount of time. Alternative substrates were assayed at 1% concentration (w/v), insoluble microcrystalline cellulose AVICEL (Sigma-Aldrich), beechwood xylan (Megazyme, Ireland), filter paper (Scharlab, Spain), cotton (Corman Spa, Italy) and starch (Sigma-Aldrich, USA). The following additives and detergents were tested to evaluate the effect on the enzymatic activity (final concentrations 5mM): NaCl, KCl, $CaCl_2$, $ZnSO_4$, $CuSO_4$, $MgCl_2$, $MnCl_2$, EDTA, SDS, Triton X-100, Tween 20, Tween 80, and CHAPS.

Results

Assembly, decontamination and data upload to MG-RAST server

From the original 170,738,281 reads with a length of 100bp from the Illumina HiSeq sequencing used for assembly, the IDBA-UD algorithm successfully assembled 371,171 contigs, with a n50 value of 950 (the shortest contig length when 50% of the total sequences are considered). The longest contig sequence was 191,032bp and the mean contig length was 777bp. The total length of all contigs assembled was 288,544,399bp. Decontamination using Deconseq and the NCBI GRCh38 human genome database resulted in the removal of 3,183 sequences. The remaining 367,988 contigs had an average length of 781bp. In the MG-RAST upload, 2,670 sequences were removed as they were detected as artificial duplicates in the quality control pipeline. The final sequence count was 365,318 contigs with an average length of 780bp.

Identification of putative endoglucanase and β-xylosidase genes and sequence analysis

The search for endoglucanases in the contig sequences using the MG-RAST server and the Subsystems database resulted in 19 contigs containing putative endoglucanases. Of those, three contigs contained predicted complete ORF (with a Met and STOP codon). These were named based on putative activity and numbered after the contig they were found in. Through BLASTp alignment to the NCBI nr database, it was found that Cel232 had a 99% identity with an endoglucanase from *Dictyoglomus thermophilum* and Cel652 had a 97% identity with an endoglucanase from *Fervidobacterium nodosum*, whereas Cel776 had a 73% identity with a glycoside hydrolase family 1 protein from *Fervidobacterium pennivorans*. Based on the lower identity of Cel776 to known proteins, it was selected for further characterization. The search for β-xylosidades in the contig sequences resulted in 29 contigs harbouring putative β-xylosidase genes according to the Subsystems database. 4 of those contigs contained a predicted complete ORF and were named and numbered as before. BLASTp alignment using the NCBI nr database revealed that Xyl468 had 99% identity with a xylan 1,4-β-xylosidase from *D. thermophilum*; Xyl7992 had 99% identity with a glycoside hydrolase family 43 protein from *Porphyrobacter cryptus*; Xyl26640 had 68% identity with a β-glucosidase from *Caldanaerobius fijiensis* DSM 17918 and Xyl59213 had 86% identity with a glycoside hydrolase from *Caldicellulosiruptor kronotskyensis*. As Xyl26640 and Xyl59213 did not feature conserved domains that supported the β-xylosidase annotation they were not favoured for further analysis. Xyl468 and Xyl7992 had high percentage of identity to proteins in the database but the 1,4-β-xylosidase from *D. thermophilum* had been studied elsewhere (49,54), and thus Xyl7992 was selected as a more interesting subject for biochemical characterization.

Xyl7992 protein product was predicted to be 538 amino acids long with a molecular weight of 59,015.96 Da and with a theorical

isoelectric point of 6.25. Protein modelling in SWISSPROT achieved the highest GMQE score (0.53) with a β-xylosidase from *Geobacillus thermoleovorans* IT-08 with a relatively low (39.03%) identity and a QMeanDisCo Global score of 0.66±0.05, which belongs to the subfamily 12 of GH43. The protein model is shown in Figure 1A. Clustal Omega alignment of conserved residues of characterized proteins in the family GH43 subfamily 12 retrieved from the CAZy database and the putative β-xylosidase Xyl7992 is shown in Figure 1B. A signal peptide feature was predicted for Xyl7992 with a likelihood of 0.4324 for a Sec/SPI, 0.5602 for a Tat/SPI and a probability for a cleavage site of 0.6874 between amino acids 32 and 33 if the sequence was considered of Gram-negative bacteria origin. These three values changed to 0.665, 0.3237 and 0.4866 respectively if the sequence was considered of Gram-positive bacteria origin.

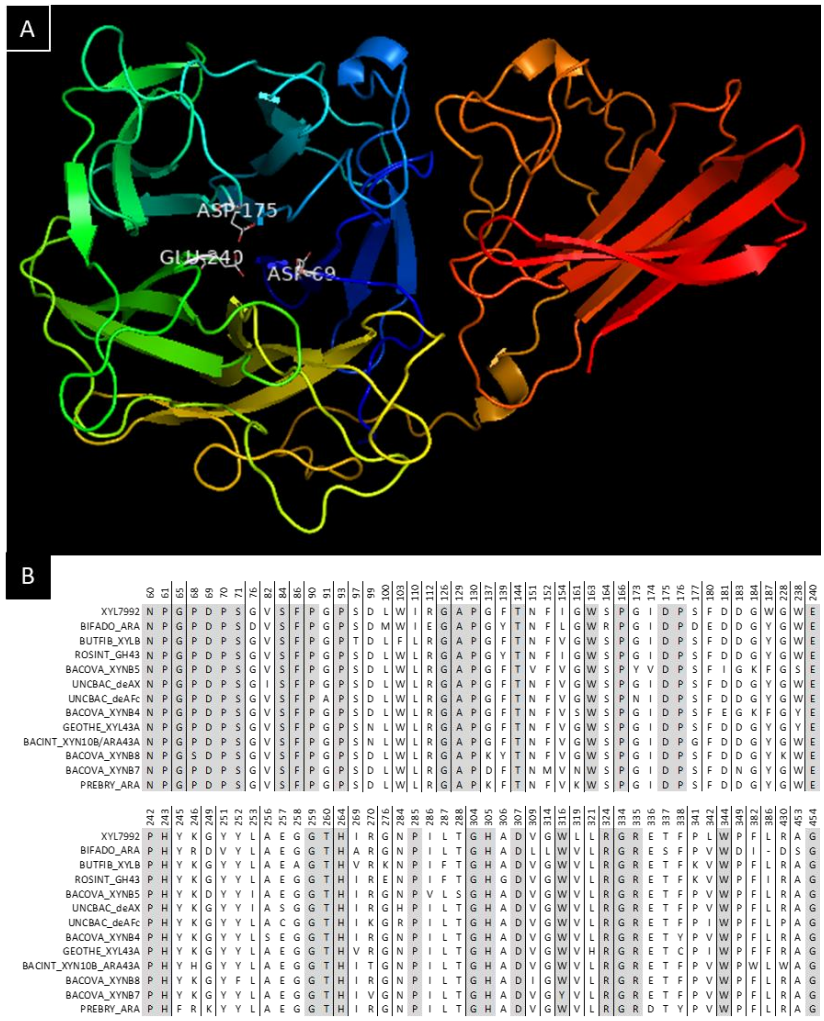
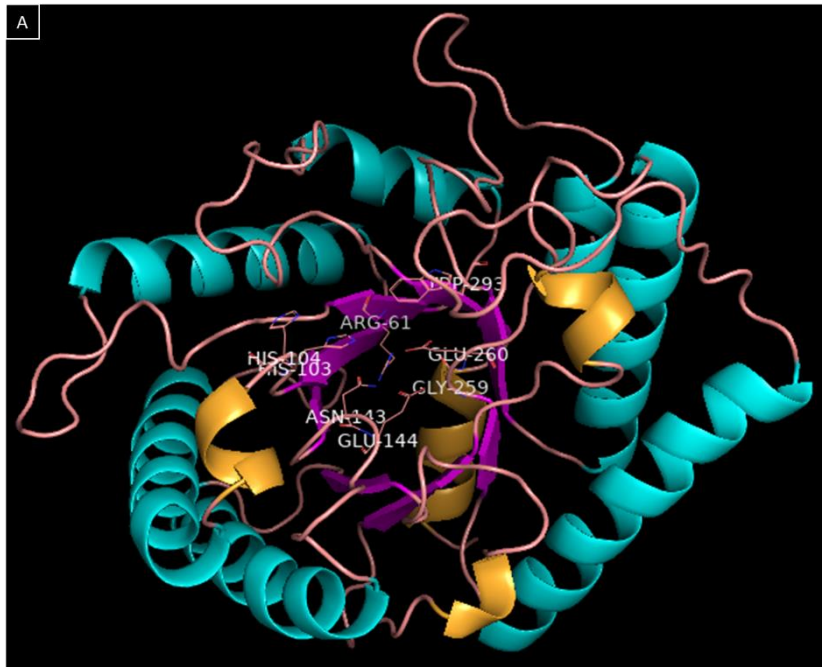


Figure 1. (A) Protein model for the gene product of Xyl7992 using the β-xylosidase from *Geobacillus thermoleovorans* IT-08 (SWISSMODEL Template Library ID: 5z5d.1.A) template in SWISSMODEL, visualized with the Pymol software. Colors follow a rainbow pattern from the initial amino acids in blue (N-terminus) to the last amino acids in red (C-terminus). Residues Asp69, Asp175 and Glu240 are highlighted. The modelled structure has a five-bladed β-propeller catalytic domain towards the N-terminus and a β-sandwich domain towards the C-terminus. (B) Alignment of conserved residues of characterized proteins in family GH43 subfamily 12 and Xyl7992. Numbers correspond to the position of the residue in the Xyl7992 sequence.

Conserved residues across all selected sequences are highlighted in gray. BIFADO_ARA: *Bifidobacterium adolescentis* α-L-arabinofuranosidase (GeneBank: BAF39204.1); BUTFIB_XYLB: *Butyrivibrio fibrisolvens* xylosidase/arabinosidase (AAA63610.1); ROSINT_GH43: *Roseburia intestinalis* L1-82 glycosyl hydrolase family 43 (EEU99940.1); BACOVA_XYNB5: *Bacteroides ovatus* β-xylosidase (ALJ47682.1); UNCBAC_deAX: uncultured bacterium α-L-arabinofuranosidase/β-xylosidase (ACF39706.1); UNBAC_deAFc: uncultured bacterium arabinofuranosidase (ABB92159.1); BACOVA_XYNB4: *Bacteroides ovatus* β-xylosidase (ALJ47681.1); GEOTHE_XYL43A: *Geobacillus thermoleovorans* β-xylosidase (ABC75004.1); BACINT_XYN10B_ARA43A: *Bacteroides intestinalis* DSM 17393 glycosyl hydrolase family 10 (EDV05059.1); BACOVA_XYNB8: *Bacteroides ovatus* β-xylosidase (ALJ48326.1); BACOVA_XINB7: *Bacteroides ovatus* β-xylosidase (ALJ48322.1); PREBRY_ARA: *Prevotella bryantii* B14 arabinoxylan arabinofuranohydrolase (ADI70680.1).

Cel776 gene product was predicted to be 320 amino acids long, with a theorical molecular weight of 37,232.31 Da and an isoelectric point of 5.67. BLASTp against the NR protein database resulted in the best hit for a glycoside hydrolase family 5 protein from *Fervidobacterium* sp. 2310opik-2 (ID: WP_164541660.1) with a score of 531 bits and 76.69% identity. SWISSMODEL modelling had the best hit with the template for the crystal structure of FnCel5A from *F. nodosum* Rt17-B1 with a sequence identity of 79.68% and a GMQE of 0.92, whereas the QMeanDisCo Global score was 0.90±0.05. The template is annotated in the CAZy database as a GH5 family protein belonging to subfamily 25. The protein model is given in Figure 2A visualized with the PyMol software and a Clustal Omega alignment performed with all ten characterized proteins in this GH subfamily deposited in the CAZy database is also given in Figure 2B. No signal peptide was predicted for the amino acid sequence of Cel776.


A
ARG-61
HIS-104
HIS-103
GLU-260
GLY-259
ASN-143
GLU-144

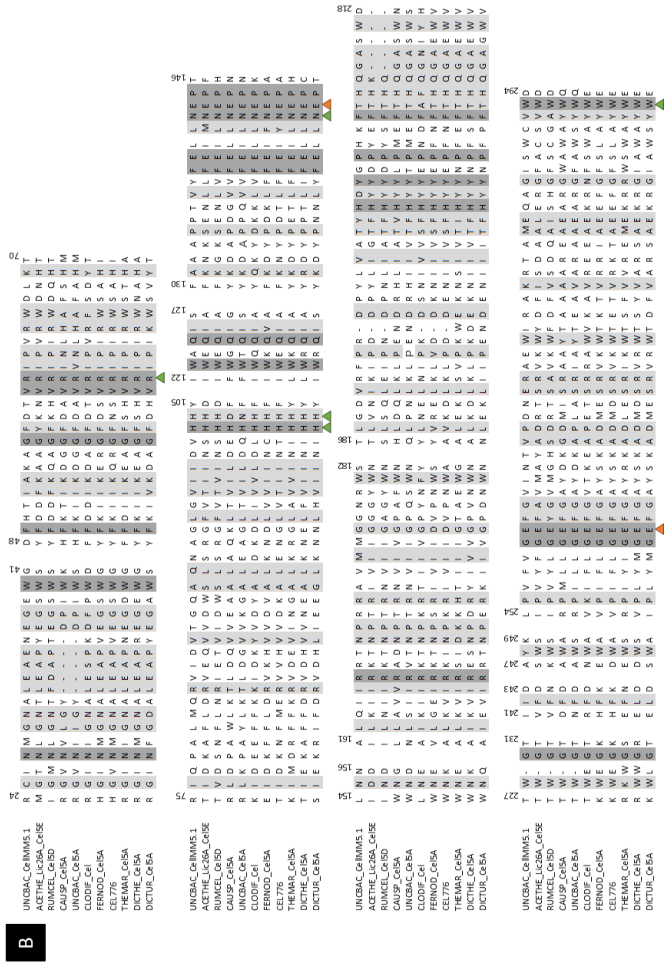


Figure 2. (A) Protein model of the gene product of Cel776 using the endoglucanase Cel5A from *F. nodosum* Rt17-B1 as template (SWISSMODEL ID: 3nco.1.A). The structure modelled consists in a $(\beta/\alpha)_8$-TIM barrel consisting in the 8 β-sheets (purple) and the 8 surrounding α-helices (cyan). 4 additional α-helices are represented in orange. Main conserved amino acids from the GH5 family are highlighted. The model was constructed in SWISSMODEL and visualized using the PyMol Software. (B) Alignment of Cel776 to sequences in the CAZy database annotated as GH5 subfamily 25 proteins. Numbers correspond to the residue below in the Cel776 sequence. Conserved residues (appearing in the same position in more than 70% of the sequences) are highlighted in gray. Strictly conserved residues (appearing in

the same position in all the sequences) are coloured in a darker shade of gray. Green triangles indicate conserved residues of GH5 that have been linked to the catalytic activity of the enzyme. Orange triangles highlight the conserved catalytic residue pair that act as proton donor and nucleophile. UNCBAC_CellMM5.1: Uncultured bacterium endoglucanase (GeneBank: AID57617.1); ACETHE_Lic26A_Cel5E: *Acetibrivio thermocellus* bifunctional endo-β-1,4-glucanase / β-1,3:1,4-glucanase (AAA23225.1); RUMCEL_Cel5D: *Ruminiclostridium cellulolyticum* H10 endo-β-1,4-glucanase (ACL75216.1); CAUSP_Cel5A: *Caulobacter* sp. K31 glycoside hydrolase family 5 (ABZ70413.1); UNCBAC_Cel5A: uncultured bacterium endo-β-1,4-glucanase (ABE60714.1); CLODIF_Cel: *Clostridioides difficile* DSM 28668 endoglucanase (AXU72614.1); FERNOD_Cel5A: *Fervidobacterium nodosum* Rt17-B1 endo-β-1,4-glucanase (ABS61403.1); THEMAR_Cel5A: *Thermotoga maritima* MSB8 bifunctional endoglucanase/β-mannanase (AAD36816.1); DICTHE_Cel5A: *Dictyoglomus thermophilum* H-6-12 endoglucanase (ACI18520.1); DICTUR_Cel5A: *Dictyoglomus turgidum* DSM 6724 endoglucanase (ACK41955.1).

Cloning of DNA fragments containing putative novel hydrolases

Correct BP and LR recombination reactions were verified by sequencing using the M13 and T7 sequencing primer pairs, respectively. Both Xyl7992 and Cel776 sequences were correctly cloned in the pDNOR221 and pDEST527 vectors in the corresponding step. Additionally, for the LR recombination, no growth on LB plates supplemented with chloramphenicol and ampicillin was observed for each of the clones harbouring the pDEST527 vectors with an insert DNA.

Purification of a putative β-xylosidase and activity tests

Experiments to detect β-xylosidase activity in the crude extract failed to find significant results that proved a correct expression of the heterologous gene in the expression system used. Experiments performed using both the crude extract and the pelleted fractions (Figure 3) revealed the an insoluble protein precipitate was formed during the induction experiments (51). As such, further

characterization was deemed unachievable for the putative β-xylosidase gene product.

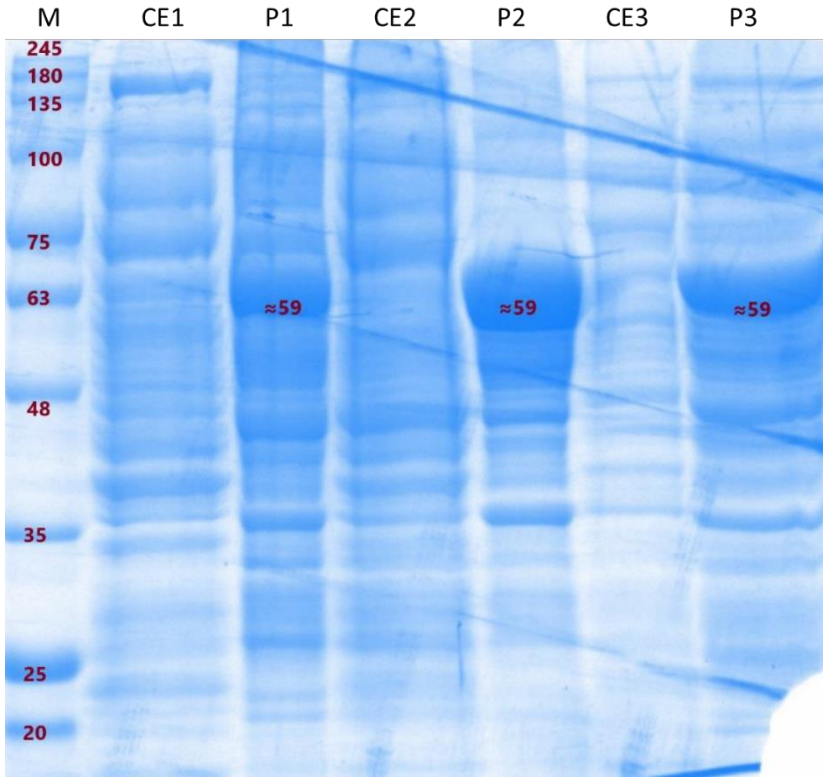


Figure 3. SDS-PAGE of the expression experiment for Xyl7992. Numbers indicate molecular weight in KDa. M: NZYColour protein marker II (NZYTech, Portugal); CE1: Crude extract with 0.05 mM IPTG induction; P1: Pelleted fraction with 0.05 mM IPTG induction; CE2: Crude extract with 0.1 mM IPTG induction; P2: Pelleted fraction with 0.1 mM IPTG induction; CE3: Crude extract with 1 mM IPTG induction; P3: Pelleted fraction with 1 mM IPTG induction.

Purification of a novel endoglucanase

As shown in Figure 4, the Cel776 gene product was purified by affinity chromatography taking advantage of the 6xHIS tag in the C-terminus

of the protein and then concentrated. The protein was recovered in the elution fraction containing 10% elution buffer.

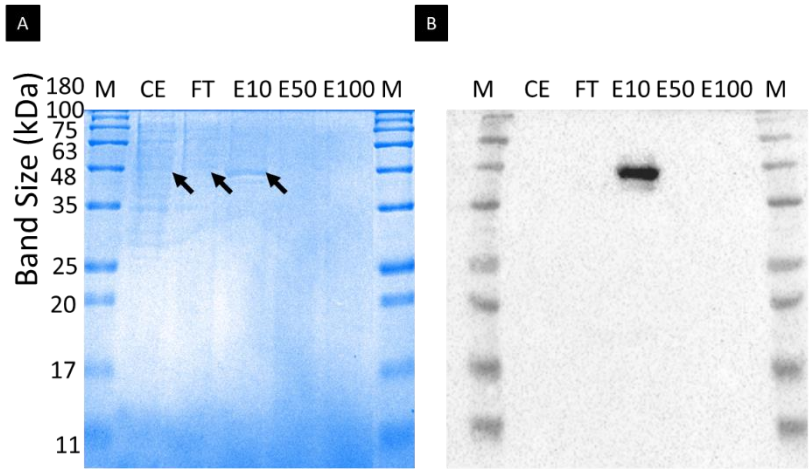


Figure 4. (A) SDS-PAGE of the purification process for the Cel776 gene product. M: NZYColour protein marker II (NZYTech, Portugal) 2.5µL; CE: Crude Extract 9µL; FT: flow-through from the column load and wash steps 9µL; E10: Elution fraction at 10% elution buffer concentration 18µL; E50: Elution fraction at 50% elution buffer concentration 18µL; E100: Elution fraction at 100% elution buffer concentration. Black arrows mark the band corresponding to the Cel776 protein. (B) Western blot of the purification steps for the Cel776 gene product. Lanes are named as in (A).

Biochemical characterization of a novel endoglucanase

The optimal temperature for the endoglucanase Cel776 was 80°C (Figure 5A). The optimum pH for the enzyme at 80°C and 60°C was found to be 5. The enzyme activity was severely reduced with varying pH conditions, and more so at 80°C than at 60°C (Figure 5B). The enzyme was found to be thermostable, maintaining its activity for over 24 hours at 60°C and with a half-life of 6.38 hours at 70°C and 3.55 hours at 80°C (Figure 5C), whereas treatment at 90°C quickly inactivated the enzyme with a half-life of 6.11 minutes (Figure 5D). Some metal ions enhanced the enzymatic activity, particularly $CaCl_2$ and $MgCl_2$ (Figure 5E). Others did not have a strong effect on it, like

KCl, $MnCl_2$ and NaCl, nor did the addition of EDTA, whereas the addition of $ZnSO_4$ and $CuSO_4$ strongly inhibited the enzymatic activity. Regarding the effect of detergents, Tween 80 did not affect the enzymatic activity, Triton X-100 and Tween 20 had slights reduction effects and CHAPS reduced it by more than 20%. SDS addition almost rendered the enzyme inactive (Figure 5F). The enzyme displayed almost no activity towards cotton and very little activity towards filter paper and starch. It showed activity when the substrate was insoluble microcrystalline cellulose (AVICEL) and was able to degrade xylan as well (Figure 5G).

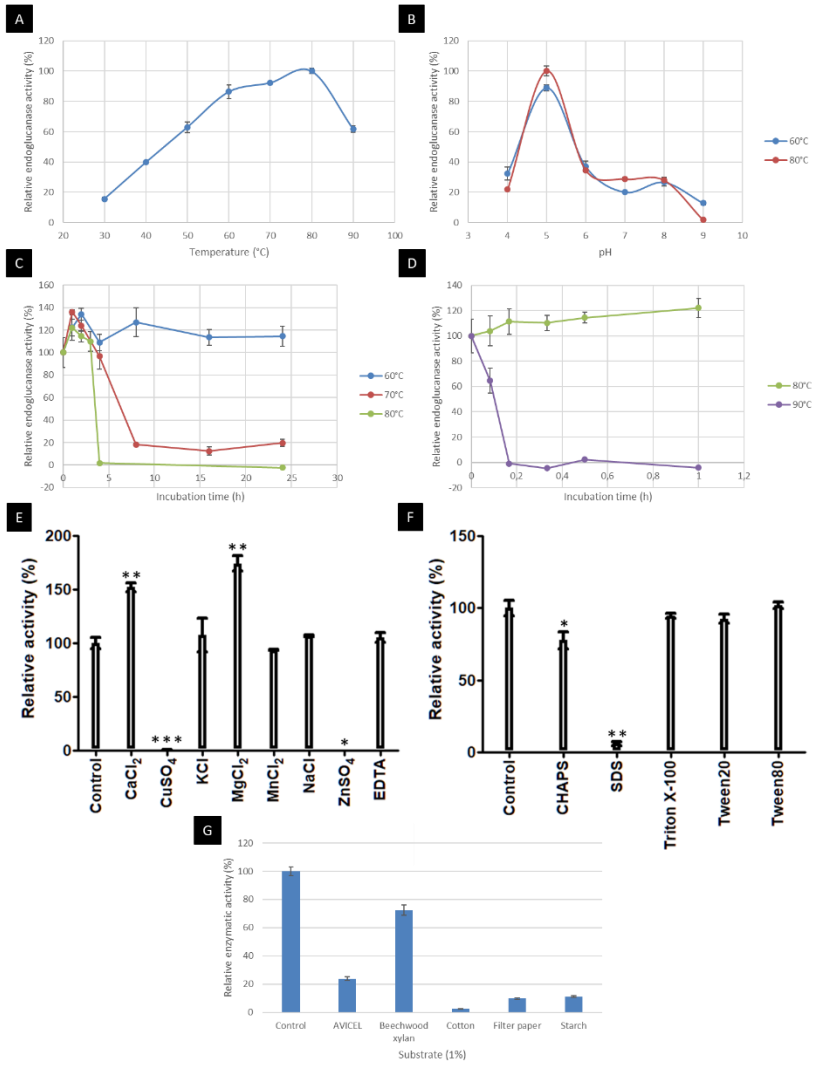


Figure 5. Biochemical characterization of Cel776. All measurements were the result of three independent experiments and an enzyme blank was subtracted to each condition tested. (A) Optimal temperature at pH 5. 100% activity is 0.6917 U/mL. (B) Optimal pH at temperatures 80°C and 60°C. 100% activity is 0.5378 U/mL. (C) Thermal stability at 60°C, 70°C, 80°C for long incubation times and (D) thermal stability at 80°C and 90°C for short incubation times. 100% activity 0.9086 U/mL. (E) Effect of additives at 5mM concentration on the enzymatic activity. The non-additive control was stablished as 100% activity with 0.6748 U/mL. An asterisk represents a p-

value lower than 0.05 in a two-sample t-student test assuming equal variances for the control and each additive condition. Two asterisks represent a p-value lower than 0.005. Three asterisks represent a p-value lower than 0.0005. (F) Effect of detergents at 5mM concentration on the enzymatic activity. A non-detergent control was used to set the 100% activity value with 0.6748 U/mL. An asterisk represents a p-value lower than 0.05 in a two-sample t-student test assuming equal variances for the control and each detergent condition. Two asterisks represent a p-value lower than 0.005. (G) Activity of Cel776 towards alternative substrates. Activity towards CMC at optimal temperature and pH was used to compare as a 100% control activity with 0.5377 U/mL.

Discussion

The role of thermozymes in lignocellulose degradation and their discovery process has been the subject of multiple reviews (2,5,8,21,55) as their biotechnological applications raises their interest as a promising environmentally friendly alternative to stablished processes along other advantages including operational cost reductions. Although endoglucanases (7,9,17,56) and xylanases (19,34,57,58) have multiple applications where only one of the activities is required, synergism of both enzymes and even other accessory enzymes is also described for particular processes (36).

In this study we identified a putative β-xylosidase belonging to the GH43 family. Enzymes categorized in GH family 43 of the CAZy database are mainly β-xylosidases and α-L-arabinofuranosidases and all of them have an inverting mechanism of catalysis (the anomeric carbon configuration of the D-xylose or L-arabinose is inverted) with a single transition state (59). This specificity towards multiple substrates is expected due to D-xylose and L-arabinose having a similar conformation near the glycosidic bond (60). The glycosidic bond is hydrolized with a single-displacement mechanism: a water molecule is deprotonated by the general base and performs a nucleophilic attack on the anomeric carbon together with the general acid, thus protonating the glycosidic oxygen and cleaving the bond (24). The general base is an Asp residue (Asp175 in Xyl7992), and the

general acid a Glu residue (Glu240 in Xyl7992) (Figure 1). The 3D structure of these enzymes includes a 5-fold β-propeller where the active site is located and in some of the subfamilies a non-catalytic β-sandwich, called X19 domain, at the C-terminus (59), which is indeed present in the protein Xyl7992 reported in this work (Figure 1A). This non-catalytic domain forms a pocket with a portion of the active site allowing access to small molecules like the substrate (61). Up to 35 residues were found to be strictly conserved when sequence alignments were performed on the twelve enzymes deposited and characterized in the CAZy database belonging to family GH43 subfamily 12 and Xyl7992 (Figure 1B). No divalent metal ion dependency has been observed for enzymes belonging to this subfamily (24,62). Most of the sequences in the CAZy database deposited in this subfamily belong to the Bacteria domain, with at least one enzyme characterized from a thermophilic organism (63). The protein sequence was highly similar to a predicted GH43 protein from *Porphyrobacter cryptus*, a moderately thermophilic organism first isolated in hot springs from Portugal (64). Despite efforts to produce and purify the gene product in the laboratory to biochemically validate the predictions made based on amino acid sequence, the protein proved to be recalcitrant to its correct expression in the *E. coli* host and pDEST527 vector system, and a soluble form was not obtained in expression experiments, with the most likely explanation being the formation of insoluble protein aggregates (inclusion bodies). A signal peptide feature was predicted in the putative β-xylosidase amino acid sequence, and failure to correctly recognize and cleave the peptide could further explain why the protein was not soluble in our expression experiments.

We also identified an enzyme in this study as a putative endoglucanase belonging to the GH5 subfamily 25, named Cel776. Proteins in the GH5 family have a retaining mechanism of catalysis (the anomeric carbon configuration is conserved in the enzymatic reaction), with two Glu residues acting as the nucleophile and the electron donor (Glu144 and Glu260 in Cel776). Their structure

consists in a $(\beta/\alpha)_8$-barrel, with some subfamilies featuring a modular domain architecture (65). Modelling of the predicted protein Cel776 fitted the $(\beta/\alpha)_8$-barrel structure (Figure 2A) using the endoglucanase Cel5A from *F. nodosum* Rt17-B1 as template, which also belongs to subfamily 25 of GH5. Indeed, proteins in subfamily 25 include thermophilic and multifunctional enzymes (65) like said endoglucanase from *F. nodosum* Rt17-B1 which can also use β-D-glucan and galactomannan as substrates (66); a endoglucanase/β-mannanase from *Thermotoga maritima* MSB8 (67), and a cellulase/xylanase from *Clostridium thermocellum* (68) (that appears as two entries in the database due to renaming of the species as *Hungateiclostridium thermocellum* (69) and *Acetibrivio thermocellus* (70)). Alignment to these and other characterized members of GH5 subfamily 25 with Cel776 confirmed conserved residues involved in the active site of the enzymes including the catalytic Glu pair (Figure 2B). Features like a carbohydrate binding module present in some members of the subfamily like the multifunctional endoglucanase from *C. thermocellum* (71) were not present in the Cel776, clearly evidenced in the Clustal Omega alignment where nearly 300 amino acids at both ends of this protein did not align with the shorter GH5_25 endoglucanases analysed. Key conserved residues in other GH5 family proteins were identified in Cel776 sequence, including the mentioned catalytic pair of Glu144 and Glu260. In the GH5 endoglucanase from *Thermotoga maritima* MSB8 hydrogen bonds between an Arg and His residues with these two catalytic Glu residues are described (72), that in the Cel776 protein would correspond to Arg61 with nucleophile Glu260 and His203 with proton donor Glu144. The conserved Asn143 in Cel776 also forms a stabilizing hydrogen bond with conserved Arg61 and stabilizes the transition state with a hydrogen bond with glucan substrates in the enzyme from *T. maritima*. Trp293 is described to allow glucose-binding in an hydrophobic context in the -1 subsite (72). These four conserved residues could be identified in the multiple sequence alignment (Figure 2B) and their spatial distribution in the active site is similar across the members of the GH5_25 family (68). Studies on

protein structure bound to the substrates xylobiose and cellobiose and site directed mutagenesis had been conducted on the endoglucanase/xylanase from *C. thermocellum* (68) and with substrates cellotetraose, cellobiose and mannotriose on the GH5 endoglucanase from *T. maritima* (73). Identified conserved residues important to catalytic activity in these two enzymes where also present in Cel776, including: Asn30, His103, His104, Asn143, Glu144, Tyr205, His212, Trp227 and Trp293. These are thought to be involved in substrate recognition mechanisms in enzymes of GH5 family (68,73).

Activity assays confirmed that Cel776 is an endoglucanase, with multifunctional activity including xylanase activity with birchwood xylan as the substrate. The biochemical characterization also revealed that it is a thermozyme, with an optimal temperature of 80°C that is maintained over a range of temperatures and capable of maintaining its activity for over 24 hours incubation periods at 60°C. Moreover, at its optimal temperature it showed a half-life of 3.55h. The optimal pH was 5 and the enzyme activity was severely affected by pH changes. There are many thermophilic endoglucanases that have been already characterized (9), and both the GH5 endoglucanase from *Acidothermus cellulolyticus* (74,75) and the endoglucanase Cel5A from *T. maritima* (72) most closely resemble the temperature and pH optimums observed for Cel776. Also, the behaviour of the enzyme towards variations on temperature and pH is similar to other GH5 family endoglucanases recently reported from metagenomic surveys (76,77). In addition, the enzymatic inhibition by zinc and copper that Cel776 exhibited is present in many endoglucanases, as well as a non-reduction effect from the addition of EDTA (76–78). An enhancing effect of $CaCl_2$ and $MgCl_2$ additives on the Cel776 was observed. Such an effect has been studied in several endoglucanases with calcium (79–81), where higher affinity for the substrate CMC is reported, whereas the effect of magnesium is reported in some enzymes, possibly because of a stabilization of the structure (78). Most detergents had moderate effects on the

enzymatic activity of Cel776, but SDS remarkably inhibited the activity, an effect observed for other endoglucanases too (77).

Conclusion

In this study an endoglucanase from a hot spring metagenome was identified by sequence-based analysis and was successfully purified and biochemically characterized. The enzyme was multifunctional, mainly acting on CMC as an endoglucanase but also able to degrade xylan as a xylanase. Its high temperature optimum and thermotolerance are desirable traits regarding its potential as a biotechnological catalyst in processes like biofuel production and saccharification of plant biomass.

References

1. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol [Internet]. 2012 Apr;21(8):1794–805. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2012.05538.x

2. Berini F, Casciello C, Marcone GL, Marinelli F. Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett [Internet]. 2017 Nov 15;364(21). Available from: https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnx211/4329276

3. Mardanov A V., Kadnikov V V., Ravin N V. Metagenomics: A Paradigm Shift in Microbiology. In: Metagenomics [Internet]. Elsevier; 2018. p. 1–13. Available from: https://doi.org/10.1016/B978-0-08-102268-9.00001-X

4. Galloway-Peña J, Hanson B. Tools for Analysis of the Microbiome. Dig Dis Sci [Internet]. 2020 Mar 31;65(3):674–85. Available from: http://link.springer.com/10.1007/s10620-020-06091-y

5. DeCastro M-E, Rodríguez-Belmonte E, González-Siso M-I.

Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol [Internet]. 2016 Sep 27;7(SEP):1–21. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01521/abstract

6. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res [Internet]. 2010 Apr;38(6):1767–71. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp1137

7. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS. Microbial Cellulose Utilization: Fundamentals and Biotechnology. Microbiol Mol Biol Rev [Internet]. 2002 Sep;66(3):506–77. Available from: https://journals.asm.org/doi/10.1128/MMBR.66.3.506-577.2002

8. Barcelos MCS, Ramos CL, Kuddus M, Rodriguez-Couto S, Srivastava N, Ramteke PW, et al. Enzymatic potential for the valorization of agro-industrial by-products. Biotechnol Lett [Internet]. 2020 Oct 9;42(10):1799–827. Available from: https://doi.org/10.1007/s10529-020-02957-3

9. Escuder-Rodríguez J-J, DeCastro M-E, Cerdán M-E, Rodríguez-Belmonte E, Becerra M, González-Siso M-I. Cellulases from Thermophiles Found by Metagenomics. Microorganisms [Internet]. 2018 Jul 10;6(3):66. Available from: http://www.mdpi.com/2076-2607/6/3/66

10. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res [Internet]. 2014 Jan;42(D1):D490–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24270786

11. Zhao X, Liu L, Deng Z, Liu S, Yun J, Xiao X, et al. Screening, cloning, enzymatic properties of a novel thermostable cellulase enzyme, and its potential application on water

hyacinth utilization. Int Microbiol [Internet]. 2021 Mar 8;(Okada 1976). Available from: http://link.springer.com/10.1007/s10123-021-00170-4

12. Pabbathi NPP, Velidandi A, Gandam PK, Koringa P, Parcha SR, Baadhe RR. Novel buffalo rumen metagenome derived acidic cellulase Cel-3.1 cloning, characterization, and its application in saccharifying rice straw and corncob biomass. Int J Biol Macromol [Internet]. 2021 Feb;170:239–50. Available from: https://doi.org/10.1016/j.ijbiomac.2020.12.041

13. Ariaeenejad S, Kavousi K, Mamaghani ASA, Motahar SFS, Nedaei H, Salekdeh GH. In-silico discovery of bifunctional enzymes with enhanced lignocellulose hydrolysis from microbiota big data. Int J Biol Macromol [Internet]. 2021;177:211–20. Available from: https://doi.org/10.1016/j.ijbiomac.2021.02.014

14. Chai S, Zhang X, Jia Z, Xu X, Zhang Y, Wang S, et al. Identification and characterization of a novel bifunctional cellulase/hemicellulase from a soil metagenomic library. Appl Microbiol Biotechnol [Internet]. 2020 Sep 15;104(17):7563–72. Available from: http://link.springer.com/10.1007/s00253-020-10766-x

15. Patel M, Patel HM, Dave S. Determination of bioethanol production potential from lignocellulosic biomass using novel Cel-5m isolated from cow rumen metagenome. Int J Biol Macromol [Internet]. 2020 Jun;153:1099–106. Available from: https://doi.org/10.1016/j.ijbiomac.2019.10.240

16. Gavande P V., Basak A, Sen S, Lepcha K, Murmu N, Rai V, et al. Functional characterization of thermotolerant microbial consortium for lignocellulolytic enzymes with central role of Firmicutes in rice straw depolymerization. Sci Rep [Internet]. 2021 Dec 4;11(1):3032. Available from: https://doi.org/10.1038/s41598-021-82163-x

17. Wilson DB. Cellulases and biofuels. Curr Opin Biotechnol [Internet]. 2009 Jun;20(3):295–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S095816690900

0676

18. Wohlgemuth R, Littlechild J, Monti D, Schnorr K, van Rossum T, Siebers B, et al. Discovering novel hydrolases from hot environments. Biotechnol Adv [Internet]. 2018 Dec;36(8):2077–100. Available from: https://doi.org/10.1016/j.biotechadv.2018.09.004

19. Basit A, Liu J, Rahim K, Jiang W, Lou H. Thermophilic xylanases: from bench to bottle. Crit Rev Biotechnol [Internet]. 2018 Oct 3;38(7):989–1002. Available from: https://doi.org/10.1080/07388551.2018.1425662

20. Tomazini A, Higasi P, Manzine LR, Stott M, Sparling R, Levin DB, et al. A novel thermostable GH5 β-xylosidase from Thermogemmatispora sp. T81. N Biotechnol [Internet]. 2019 Nov;53(July):57–64. Available from: https://doi.org/10.1016/j.nbt.2019.07.002

21. Saha BC. Hemicellulose bioconversion. J Ind Microbiol Biotechnol [Internet]. 2003 May 1;30(5):279–91. Available from: https://academic.oup.com/jimb/article/30/5/279-291/5992143

22. Alokika, Singh B. Production, characteristics, and biotechnological applications of microbial xylanases. Appl Microbiol Biotechnol [Internet]. 2019 Nov 22;103(21–22):8763–84. Available from: http://link.springer.com/10.1007/s00253-019-10108-6

23. Santibáñez L, Henríquez C, Corro-Tejeda R, Bernal S, Armijo B, Salazar O. Xylooligosaccharides from lignocellulosic biomass: A comprehensive review. Carbohydr Polym [Internet]. 2021 Jan;251(September 2020):117118. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0144861720312911

24. Rohman A, Dijkstra BW, Puspaningsih NNT. β-Xylosidases: Structural Diversity, Catalytic Mechanism, and Inhibition by Monosaccharides. Int J Mol Sci [Internet]. 2019 Nov

6;20(22):5524. Available from: https://www.mdpi.com/1422-0067/20/22/5524

25. Liu Y, Huang L, Zheng D, Xu Z, Li Y, Shao S, et al. Biochemical characterization of a novel GH43 family β-xylosidase from Bacillus pumilus. Food Chem [Internet]. 2019 Oct;295(May):653–61. Available from: https://doi.org/10.1016/j.foodchem.2019.05.163

26. Jordan DB, Braker JD, Wagschal K, Stoller JR, Lee CC. Isolation and divalent-metal activation of a β-xylosidase, RUM630-BX. Enzyme Microb Technol [Internet]. 2016 Jan;82:158–63. Available from: http://dx.doi.org/10.1016/j.enzmictec.2015.10.001

27. Souto B de M, de Araújo ACB, Hamann PRV, Bastos A de R, Cunha I de S, Peixoto J, et al. Functional screening of a Caatinga goat (Capra hircus) rumen metagenomic library reveals a novel GH3 β-xylosidase. Lee S-W, editor. PLoS One [Internet]. 2021 Jan 15;16(1):e0245118. Available from: https://dx.plos.org/10.1371/journal.pone.0245118

28. Zhou J, Bao L, Chang L, Zhou Y, Lu H. Biochemical and kinetic characterization of GH43 β- <scp>d</scp> -xylosidase/α- <scp>l</scp> -arabinofuranosidase and GH30 α- <scp>l</scp> -arabinofuranosidase/β- <scp>d</scp> -xylosidase from rumen metagenome. J Ind Microbiol Biotechnol [Internet]. 2012 Jan 1;39(1):143–52. Available from: https://academic.oup.com/jimb/article/39/1/143/5994629

29. Zhou J, Bao L, Chang L, Liu Z, You C, Lu H. Beta-xylosidase activity of a GH3 glucosidase/xylosidase from yak rumen metagenome promotes the enzymatic degradation of hemicellulosic xylans. Lett Appl Microbiol [Internet]. 2012 Feb;54(2):79–87. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1472-765X.2011.03175.x

30. Bao L, Huang Q, Chang L, Sun Q, Zhou J, Lu H. Cloning and Characterization of Two β-Glucosidase/Xylosidase Enzymes

from Yak Rumen Metagenome. Appl Biochem Biotechnol [Internet]. 2012 Jan 22;166(1):72–86. Available from: http://link.springer.com/10.1007/s12010-011-9405-x

31. Klippel B, Sahm K, Basner A, Wiebusch S, John P, Lorenz U, et al. Carbohydrate-active enzymes identified by metagenomic analysis of deep-sea sediment bacteria. Extremophiles [Internet]. 2014 Sep 10;18(5):853–63. Available from: http://link.springer.com/10.1007/s00792-014-0676-3

32. Matsuzawa T, Kaneko S, Yaoi K. Screening, identification, and characterization of a GH43 family β-xylosidase/α-arabinofuranosidase from a compost microbial metagenome. Appl Microbiol Biotechnol [Internet]. 2015 Nov 14;99(21):8943–54. Available from: http://link.springer.com/10.1007/s00253-015-6647-5

33. Sato M, Suda M, Okuma J, Kato T, Hirose Y, Nishimura A, et al. Isolation of highly thermostable β-xylosidases from a hot spring soil microbial community using a metagenomic approach. DNA Res [Internet]. 2017 Dec 1;24(6):649–56. Available from: http://academic.oup.com/dnaresearch/article/24/6/649/4077064

34. Bosetto A, Justo PI, Zanardi B, Venzon SS, Graciano L, dos Santos EL, et al. Research Progress Concerning Fungal and Bacterial β-Xylosidases. Appl Biochem Biotechnol [Internet]. 2016 Feb 4;178(4):766–95. Available from: http://link.springer.com/10.1007/s12010-015-1908-4

35. Li Q, Jiang Y, Tong X, Pei J, Xiao W, Wang Z, et al. Cloning and characterization of the β-xylosidase from Dictyoglomus turgidum for high efficient biotransformation of 10-deacetyl-7-xylosltaxol. Bioorg Chem [Internet]. 2020 Jan;94(October 2019):103357. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0045206819309071

36. Bajaj P, Mahajan R. Cellulase and xylanase synergism in industrial biotechnology. Appl Microbiol Biotechnol

[Internet]. 2019 Nov 18;103(21–22):8711–24. Available from: http://link.springer.com/10.1007/s00253-019-10146-0

37. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics [Internet]. 2012 Jun 1;28(11):1420–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22495754

38. Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. Rodriguez-Valera F, editor. PLoS One [Internet]. 2011 Mar 9;6(3):e17288. Available from: https://dx.plos.org/10.1371/journal.pone.0017288

39. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics [Internet]. 2008 Dec 19;9(1):386. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-386

40. Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang H-Y, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res [Internet]. 2005;33(17):5691–702. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16214803

41. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook [Internet]. Totowa, NJ: Humana Press; 2005. p. 571–608. Available from: http://link.springer.com/10.1385/1592598900

42. NCBI_Resource_Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D7–19. Available from: https://academic.oup.com/nar/article-

lookup/doi/10.1093/nar/gkv1290

43. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–303.

44. SchrödingerLCC. The PyMOL Molecular Graphics System, Version 1.2r3pre [Internet]. Available from: https://pymol.org/2/#page-top

45. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol [Internet]. 2011 Jan 11;7(1):539. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21988835

46. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D506–15. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

47. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol [Internet]. 2019 Apr 18;37(4):420–3. Available from: http://www.nature.com/articles/s41587-019-0036-z

48. Shi H, Li X, Gu H, Zhang Y, Huang Y, Wang L, et al. Biochemical properties of a novel thermostable and highly xylose-tolerant β-xylosidase/α-arabinosidase from Thermotoga thermarum. Biotechnol Biofuels [Internet]. 2013;6(1):27. Available from: Biotechnology for Biofuels

49. Li Q, Wu T, Qi Z, Zhao L, Pei J, Tang F. Characterization of a novel thermostable and xylose-tolerant GH 39 β-xylosidase from Dictyoglomus thermophilum. BMC Biotechnol [Internet]. 2018 Dec 21;18(1):29. Available from: https://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-018-0440-3

50. Margolles-Clark E, Tenkanen M, Nakari-Setälä T, Penttilä M.

Cloning of genes encoding alpha-L-arabinofuranosidase and beta-xylosidase from Trichoderma reesei by expression in Saccharomyces cerevisiae. Appl Environ Microbiol [Internet]. 1996 Oct;62(10):3840–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8837440

51. Palleiro-Mosquera P. Approach to the cloning of a beta-xylosidase obtained by sequence-based metagenomics. Master Degree Final Project. Univerdade da Coruña; 2021.

52. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem [Internet]. 1976 May 7;72:248–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/942051

53. Silveira MHL, Aguiar RS, Siika-aho M, Ramos LP. Assessment of the enzymatic hydrolysis profile of cellulosic substrates based on reducing sugar release. Bioresour Technol [Internet]. 2014;151:392–6. Available from: http://dx.doi.org/10.1016/j.biortech.2013.09.135

54. Bretagne D, Pâris A, de Vaumas R, Lafite P, Daniellou R. Crystal structure of Dictyoglomus thermophilum β-d-xylosidase DtXyl unravels the structural determinants for efficient notoginsenoside R1 hydrolysis. Biochimie [Internet]. 2021 Feb;181:34–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0300908420303059

55. Chettri D, Verma AK, Sarkar L, Verma AK. Role of extremophiles and their extremozymes in biorefinery process of lignocellulose degradation. Extremophiles [Internet]. 2021 May 25;25(3):203–19. Available from: https://doi.org/10.1007/s00792-021-01225-0

56. Sahoo K, Sahoo RK, Gaur M, Subudhi E. Cellulolytic thermophilic microorganisms in white biotechnology: a review. Folia Microbiol (Praha) [Internet]. 2020 Feb 17;65(1):25–43. Available from: http://link.springer.com/10.1007/s12223-019-00710-6

57. Kumar V, Dangi AK, Shukla P. Engineering Thermostable Microbial Xylanases Toward its Industrial Applications. Mol Biotechnol [Internet]. 2018 Mar 29;60(3):226–35. Available from: https://doi.org/10.1007/s12033-018-0059-6

58. Chadha BS, Kaur B, Basotra N, Tsang A, Pandey A. Thermostable xylanases from thermophilic fungi and bacteria: Current perspective. Bioresour Technol [Internet]. 2019 Apr;277(November 2018):195–203. Available from: https://doi.org/10.1016/j.biortech.2019.01.044

59. Jordan DB, Wagschal K, Grigorescu AA, Braker JD. Highly active β-xylosidases of glycoside hydrolase family 43 operating on natural and artificial substrates. Appl Microbiol Biotechnol [Internet]. 2013 May 10;97(10):4415–28. Available from: http://link.springer.com/10.1007/s00253-012-4475-4

60. Mewis K, Lenfant N, Lombard V, Henrissat B. Dividing the Large Glycoside Hydrolase Family 43 into Subfamilies: a Motivation for Detailed Enzyme Characterization. Nojiri H, editor. Appl Environ Microbiol [Internet]. 2016 Mar 15;82(6):1686–92. Available from: https://journals.asm.org/doi/10.1128/AEM.03453-15

61. Brunzelle JS, Jordan DB, McCaslin DR, Olczak A, Wawrzak Z. Structure of the two-subsite β-d-xylosidase from Selenomonas ruminantium in complex with 1,3-bis[tris(hydroxymethyl)methylamino]propane. Arch Biochem Biophys [Internet]. 2008 Jun;474(1):157–66. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0003986108001240

62. Jordan DB, Stoller JR, Lee CC, Chan VJ, Wagschal K. Biochemical Characterization of a GH43 β-Xylosidase from Bacteroides ovatus. Appl Biochem Biotechnol [Internet]. 2017 May 16;182(1):250–60. Available from: http://link.springer.com/10.1007/s12010-016-2324-0

63. Rohman A, van Oosterwijk N, Puspaningsih NNT, Dijkstra BW. Structural basis of product inhibition by arabinose and xylose

of the thermostable GH43 β-1,4-xylosidase from Geobacillus thermoleovorans IT-08. Permyakov EA, editor. PLoS One [Internet]. 2018 Apr 26;13(4):e0196358. Available from: https://dx.plos.org/10.1371/journal.pone.0196358

64. Rainey FA, Silva J, Nobre MF, Silva MT, da Costa MS. Porphyrobacter cryptus sp. nov., a novel slightly thermophilic, aerobic, bacteriochlorophyll a-containing species. Int J Syst Evol Microbiol [Internet]. 2003 Jan 1;53(1):35–41. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02308-0

65. Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol Biol [Internet]. 2012 Dec 20;12(1):186. Available from: BMC Evolutionary Biology

66. Zheng B, Yang W, Zhao X, Wang Y, Lou Z, Rao Z, et al. Crystal Structure of Hyperthermophilic Endo-β-1,4-glucanase. J Biol Chem [Internet]. 2012 Mar;287(11):8336–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820609796

67. Chhabra SR, Shockley KR, Ward DE, Kelly RM. Regulation of Endo-Acting Glycosyl Hydrolases in the Hyperthermophilic Bacterium Thermotoga maritima Grown on Glucan- and Mannan-Based Polysaccharides. Appl Environ Microbiol [Internet]. 2002 Feb;68(2):545–54. Available from: https://journals.asm.org/doi/10.1128/AEM.68.2.545-554.2002

68. Yuan S-F, Wu T-H, Lee H-L, Hsieh H-Y, Lin W-L, Yang B, et al. Biochemical Characterization and Structural Analysis of a Bifunctional Cellulase/Xylanase from Clostridium thermocellum. J Biol Chem [Internet]. 2015 Feb;290(9):5739–48. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925819468452

69. Zhang X, Tu B, Dai L, Lawson PA, Zheng Z, Liu L-Y, et al. Petroclostridium xylanilyticum gen. nov., sp. nov., a xylan-degrading bacterium isolated from an oilfield, and reclassification of clostridial cluster III members into four novel genera in a new Hungateiclostridiaceae fam. nov. Int J Syst Evol Microbiol [Internet]. 2018 Oct 1;68(10):3197–211. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31526446

70. Tindall BJ. The names hungateiclostridium zhang et al. 2018, hungateiclostridium thermocellum (viljoen et al. 1926) zhang et al. 2018, hungateiclostridium cellulolyticum (patel et al. 1980) zhang et al. 2018, hungateiclostridium aldrichii (yang et al. 1990) zhang et . Int J Syst Evol Microbiol. 2019;69(12):3927–32.

71. Carvalho AL, Goyal A, Prates JAM, Bolam DN, Gilbert HJ, Pires VMR, et al. The Family 11 Carbohydrate-binding Module of Clostridium thermocellum Lic26A-Cel5E Accommodates β-1,4- and β-1,3–1,4-Mixed Linked Glucans at a Single Binding Site. J Biol Chem [Internet]. 2004 Aug;279(33):34785–93. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820773151

72. Pereira JH, Chen Z, McAndrew RP, Sapra R, Chhabra SR, Sale KL, et al. Biochemical characterization and crystal structure of endoglucanase Cel5A from the hyperthermophilic Thermotoga maritima. J Struct Biol [Internet]. 2010 Dec;172(3):372–9. Available from: http://dx.doi.org/10.1016/j.jsb.2010.06.018

73. Wu T-H, Huang C-H, Ko T-P, Lai H-L, Ma Y, Chen C-C, et al. Diverse substrate recognition mechanism revealed by Thermotoga maritima Cel5A structures in complex with cellotetraose, cellobiose and mannotriose. Biochim Biophys Acta - Proteins Proteomics [Internet]. 2011 Dec;1814(12):1832–40. Available from: http://dx.doi.org/10.1016/j.bbapap.2011.07.020

74. Himmel ME, Adney William S, Tucker MP, Grohmann K.

Thermostable purified endoglucanases from thermophilic bacterium Acidothermus cellulolyticus [Internet]. WO 93/15186, 1993. Available from: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO1993015186&tab=PCTBIBLIO

75. Lindenmuth BE, McDonald KA. Production and characterization of Acidothermus cellulolyticus endoglucanase in Pichia pastoris. Protein Expr Purif [Internet]. 2011 Jun;77(2):153–8. Available from: http://dx.doi.org/10.1016/j.pep.2011.01.006

76. Suleiman M, Schröder C, Klippel B, Schäfers C, Krüger A, Antranikian G. Extremely thermoactive archaeal endoglucanase from a shallow marine hydrothermal vent from Vulcano Island. Appl Microbiol Biotechnol [Internet]. 2019 Feb 13;103(3):1267–74. Available from: http://link.springer.com/10.1007/s00253-018-9542-z

77. Joshi N, Kaushal G, Singh SP. Biochemical characterization of a novel thermo-halo-tolerant GH5 endoglucanase from a thermal spring metagenome. Biotechnol Bioeng [Internet]. 2021 Apr 21;118(4):1531–44. Available from: https://onlinelibrary.wiley.com/doi/10.1002/bit.27668

78. Shi R, Li Z, Ye Q, Xu J, Liu Y. Heterologous expression and characterization of a novel thermo-halotolerant endoglucanase Cel5H from Dictyoglomus thermophilum. Bioresour Technol [Internet]. 2013 Aug;142:338–44. Available from: http://dx.doi.org/10.1016/j.biortech.2013.05.037

79. Gaur R, Tiwari S. Isolation, production, purification and characterization of an organic-solvent-thermostable alkalophilic cellulase from Bacillus vallismortis RG-07. BMC Biotechnol [Internet]. 2015 Dec 19;15(1):19. Available from: https://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-015-0129-9

80. Wang H-J, Hsiao Y-Y, Chen Y-P, Ma T-Y, Tseng C-P. Polarity Alteration of a Calcium Site Induces a Hydrophobic

Interaction Network and Enhances Cel9A Endoglucanase Thermostability. Kelly RM, editor. Appl Environ Microbiol [Internet]. 2016 Mar 15;82(6):1662–74. Available from: https://journals.asm.org/doi/10.1128/AEM.03326-15

81. Chauvaux S, Beguin P, Aubert JP, Bhat KM, Gow LA, Wood TM, et al. Calcium-binding affinity and calcium-enhanced activity of Clostridium thermocellum endoglucanase D. Biochem J [Internet]. 1990 Jan 1;265(1):261–5. Available from: https://portlandpress.com/biochemj/article/265/1/261/24767/Calciumbinding-affinity-and-calciumenhanced

CONCLUSIONS

1. Two metagenomic libraries were constructed, a 27,898 clone library from As Burgas and a 4,399 clone library from Muiño da Veiga, using the bacteria host *E. coli* EPI300 $T1^R$ and the arabinose inducible vector pCT3FK, and with backup copies for long-term storage and functional screenings.

2. The following number of clones with enzymatic activities of interest were identified in the metagenomic libraries: 2 endoglucanases, 5 β-glucosidases, 2 β-xylosidases, 4 proteases and 1 enzyme with lipolytic activity from As Burgas, and 4 endoglucanases, 1 β-glucosidase, 6 proteases and 8 enzymes with lipolytic activity from Muiño da Veiga.

3. The lipolytic activity enzyme LipB12_A11 was sequenced, cloned, expressed and purified. It was identified by sequence analysis as a protein from hydrolytic family SNGH/GDSL, with conserved domains and amino acids. Based on its preference for short chain substrates (C8 and C12) it was classified as an esterase, with an optimum temperature of 60°C, optimum pH of 9 and high thermostability.

4. The Muiño da Veiga metagenome was sequenced, and the presence of biotechnologically interesting enzymes was predicted. The microbial community was described, with the most abundant genus being *Sulfurihydrogenibium*. The most abundant genera were mainly bacteria related to the sulfur cycle, but nitrogen and hydrogen pathways were also present. The ability to use multiple carbon sources was detected, including the biotechnologically relevant ability to use lignocellulosic biomass. Insight into the rare biosphere composition revealed the possibility of photosynthetic and methanogenic microorganisms contributing to the metabolic variety of the hot spring.

5. The metagenome was assembled to reconstruct longer sequences, called contigs. Several open reading frames (ORF) codifying for putative endoglucanases and β-xylosidases were identified. Putative β-xylosidase Xyl7992 belonged to glycosyl hydrolase (GH) family 43 subfamily 12 and homology modelling showed a 5-sheet beta propeller structure as the catalytic domain and a beta sandwich non-catalytic domain, both with conserved amino acid residues. On the other hand, putative endoglucanase Cel776 belonged to GH family 25 subfamily 5 and homology modelling predicted a $(\beta/\alpha)_8$ barrel tridimensional structure in which conserved amino acids were identified as well.

6. Xyl7992 was cloned in an expression system with *E. coli* T7 Express and vector pJET1.2, but inclusion body formation prevented the characterization of the enzyme. Cel776 was cloned, expressed and purified using the same system, and a biochemical characterization was conducted. The enzyme was most active towards carboxymethyl-celulose (CMC) substrate confirming that it was an endoglucanase, but it proved to be active against xylan as well. The enzyme was thermostable, with an optimal temperature and pH of 80°C and 5, respectively. An enhancer effect was observed on enzymatic activity with the addition of calcium and magnesium.

ANNEX I

Resumen en castellano

En la presente tesis doctoral se ha estudiado el metagenoma de dos fuentes termales localizadas en la ciudad de Ourense (Galicia, España): As Burgas y Muiño da Veiga. Los objetivos principales fueron la identificación y la caracterización bioquímica de enzimas termófilas y termotolerantes (termozimas) codificadas en dichos metagenomas y con posible utilidad biotecnológica, así como la secuenciación de uno de estos metagenomas para estudiar la composición de la comunidad microbiana y sus posibles interacciones metabólicas. Además, el estudio de la secuencia de ADN ha permitido realizar predicciones y búsquedas de nuevas termozimas, expandiendo el alcance de la técnica para la bioprospección de estos catalizadores con interés biotecnológico.

Objetivos

1. Construir metagenotecas a partir del ADN ambiental extraído de muestras de dos aguas termales de la región de Ourense: As Burgas y Muiño da Veiga.
2. Realizar cribados funcionales con las metagenotecas para identificar actividades enzimáticas de interés usando sustratos específicos.
3. Secuenciar, clonar, expresar y purificar una enzima con actividad lipolítica para realizar predicciones a partir de su secuencia y para su caracterización bioquímica.
4. Secuenciar el metagenoma de la muestra de Muiño da Veiga para realizar un análisis bioinformático que permita describir la comunidad microbiana y sus interacciones metabólicas.
5. Analizar el metagenoma secuenciado para identificar termozimas que puedan ser clonadas y expresadas en el laboratorio.
6. Clonar, expresar y purificar una β-xilosidasa y una endoglucanasa predichas por el análisis de secuencia para realizar una caracterización bioquímica y confirmar su potencial biotecnológico.

Capítulo 1: construcción de metagenotecas a partir de muestras de dos fuentes termales

En primer lugar, se realizó el muestreo de las aguas de la fuente termal de As Burgas, localizada en la ciudad de Ourense, y de la fuente termal de Muiño da Veiga, localizada en Ourense junto al río Miño. Esta región atrae turismo termal (1) gracias a las numerosas fuentes de aguas que brotan a elevadas temperaturas tras recorrer las fallas y fracturas de la roca granítica, que apenas presenta permeabilidad. El agua recorre este circuito acumulando minerales e intercambiando calor con la roca a gran profundidad (2), dando origen a numerosas fuentes termales de naturaleza no volcánica al final del recorrido (1,3). Además del turismo termal, estas aguas pueden suponer un novedoso recurso para la obtención de enzimas termófilas con interés biotecnológico.

Los volúmenes de agua recogidos fueron transportados al laboratorio y procesados lo antes posible. El agua fue filtrada usando filtros de 0.22µm para atrapar los microorganismos presentes en la muestra, y se siguió un protocolo para la lisis celular y la extracción del ADN metagenómico. La extracción del ADN metagenómico supone un punto crítico en los estudios de metagenómica, ya que intervienen una serie de factores que pueden comprometer el éxito de la investigación (4). La baja densidad de células en este tipo de ambientes limita la cantidad de biomasa disponible y por tanto la cantidad de ADN que se puede recuperar (5), y por ello un paso previo de filtrado facilita contar con suficiente masa celular en un volumen adecuado. Otro factor a tener en cuenta es la presencia de contaminantes que interfieren con las reacciones enzimáticas posteriores, que pueden ser eliminados durante el proceso de extracción realizando una electroforesis en gel de agarosa y recuperando el ADN que ha tenido una migración diferencial respecto a estos compuestos inhibidores. Además, el ADN recuperado debe tener un tamaño adecuado para el tipo de aplicación que deseamos, en este caso se requería un tamaño lo suficientemente grande para el sistema de clonado empleado (6). Por último, la cantidad de ADN recuperado debe ser suficiente como para

ser representativo de todas las especies presentes en la muestra, ya que las especies más abundantes estarán más representadas. Para la obtención del ADN metagenómico con la suficiente cantidad, pureza, tamaño y representación se utilizaron los componentes de "Metagenomic DNA isolation kit for 189áter" (Epicentre). Una vez recuperado este ADN metagenómico, y usando componentes de "CopyControl Fosmid Library Production Kit" (Epicentre), se repararon los extremos generando extremos romos y fosforilados para permitir el ligamiento a un vector de tipo fósmido (pCT3FK) que contiene secuencias adaptadoras para la introducción en cabezas vacías de fago λ (7). Tras introducir el ADN en estos fagos, usando células del hospedador *Escherichia coli* que expresan el receptor de maltosa (necesario para que el fago infecte las células), el material genético foráneo se introduce a su vez en el hospedador (transformación) y se construye una metagenoteca, es decir, una colección de clones en los que cada clon porta un fragmento de ADN metagenómico distinto. De cada muestra procesada fue posible construir una metagenoteca, siendo la de As Burgas de un tamaño de 27,898 clones y la de Muiño da Veiga de 4,399 clones. Los clones fueron colocados en placas de 96 pocillos y se realizaron duplicados de las metagenotecas para poder almacenarlas a largo plazo con glicerol. Los tamaños de las metagenotecas concuerdan con otros estudios en los que se clonan fragmentos largos de ADN (8), que permiten contener en un mismo inserto múltiples genes completos y sus elementos reguladores, e incluso operones en los que los genes se organizan con una relación funcional. Dado que se utilizó un vector de número de copia inducible para construir las metagenotecas, se realizaron experimentos para determinar las mejores condiciones para dicha inducción. Se encontró que una concentración final de arabinosa de 0.02% era óptima para lograr la inducción a alto número de copias. Los vectores inducibles aportan ventajas significativas en las metagenotecas, al conseguir mayor estabilidad génica durante el almacenamiento y propagación de la metagenoteca, manteniendo pocas copias por clon, pero permitiendo alcanzar mayores niveles de expresión de los genes insertados y mayor recuperación de ADN cuando se induce el vector a un alto número de copias.

Capítulo 2: Bioprospección de termozimas con interés biotecnológico y caracterización bioquímica de una nueva enzima lipolítica perteneciende a la familia de hidrolasas SGNH/GDSL

La siguiente etapa de la investigación consistió en la realización de experimentos de búsqueda funcional utilizando los clones de las metagenotecas y métodos de cribado basados en la degradación de sustratos específicos de las actividades enzimáticas de interés (5,9–16). Este método tiene una serie de obstáculos a sortear ya que requiere la correcta expresión de un producto génico en un sistema de expresión heterólogo (4,9,17). La distancia evolutiva de los organismos termófilos que habitan las aguas termales y los hospedadores habitualmente usados en los laboratorios muchas veces dificulta aprovechar todo el potencial de los metagenomas. La no detección en los cribados funcionales de determinadas actividades enzimáticas puede pues deberse a numerosos factores y no tiene por qué implicar que el producto génico no se encuentre codificado en el metagenoma (falso negativo). Estos factores incluyen el no reconocimiento de sistemas reguladores (incluyendo promotores y sitios de unión del ribosoma) y de procesamiento, plegamientos incorrectos, uso preferencial de codones y acumulación intracelular, entre otros. La elección de un método de detección adecuado también resulta crucial para el éxito de la técnica. En ambas metagenotecas, se evaluó la capacidad de los clones de utilizar sustratos que revelaran actividades enzimáticas con relevancia biotecnológica, a saber, exo y endoglucanasas (utilizando diferentes sustratos derivados de la celulosa) (18,19,28,20–27), xilanasas (29), β-xilosidasas (30), β-glucosidasas (28), proteasas (28), enzimas lipolíticas (31–34) y feruloil esterasas (31,35,36). No se consiguieron detectar exoglucanasas, xilanasas ni feruloil esterasas en ninguna de las dos metagenotecas pero sí endoglucanasas, β-xilosidasas, β-glucosidasas, proteasas y enzimas lipolíticas.

Se seleccionó una enzima con actividad lipolítica, LipB12_A11 para un estudio en profundidad. Se realizó un subclonado con la enzima de restricción *Eco*RV para transferir el inserto desde el fósmido pCT3FK hasta el plásmido pJET1.2 para facilitar el estudio. La secuencia completa se obtuvo mediante secuenciación y *primer walking* desde ambos extremos. La construcción génica se transfirió a una cepa de expresión de *E. coli* y se consiguió expresar la enzima utilizando IPTG para la inducción. Se siguieron los siguientes pasos para la purificación de la enzima: precipitación térmica diferencial, incubando a 60°C durante 20 minutos para desnaturalizar proteínas no termófilas como las presentes en el hospedador; precipitación con sulfato amónico, aumentando la concentración de sulfato amónico hasta el 50% que provocaba la precipitación de la enzima lipolítica; diálisis, para retirar el sulfato amónico del paso anterior; cromatografía de exclusión molecular, para descartar las proteínas con tamaño distinto al de la enzima lipolítica; concentración en columna de ultrafiltración por centrifugación, obteniendo así una enzima purificada y concentrada para realizar la caracterización bioquímica.

La enzima LipB12_A11 tiene preferencia por los sustratos para-nitrofenil-octanoato (C8) y para-nitrofenil-laurato (C12), por lo que se puede clasificar como una esterasa (37–39). La temperatura óptima de 60°C y el pH óptimo de 9, y su elevada termoestabilidad a 60°C la hacen atractiva para diferentes aplicaciones biotecnológicas. La actividad enzimática resultó muy favorecida por la presencia de determinados detergentes, en especial Tween80, y además mostró actividad sobre algunos sustratos subrogados. Por el contrario, la presencia de determinados iones metálicos provocó la reducción o completa inhibición de la actividad enzimática, así como ocurrió con el detergente Triton X-100. Además de la caracterización bioquímica, el análisis de la secuencia mediante alineamientos BLASTp (40) usando la base de datos no-redundante NR reveló que LipB12_A11 pertenece a la familia de hidrolasas SGNH/GDSL y se pudo proponer un modelo de su estructura basado en homología con otros miembros de dicha familia. El mejor resultado de alineamiento fue con la proteína de la familia de hidrolasas SGNH/GDSL procedente de

Bacillus sp. (código WP_028398363.1) con un e-valor de 6e-84, y un porcentaje de identidad de tan solo 58.73%. También en base al análisis de la secuencia, se identificaron mediante alineamientos múltiples usando la herramienta Clustal Omega (41) los aminoácidos conservados con otros miembros de la familia de hidrolasas SGNH/GDSL y se observaron bloques conservados descritos con anterioridad (42). La representación gráfica usando la herramienta Web Logo (43) facilitó la visualización del grado de conservación de las posiciones de los aminoácidos.

Capítulo 3: Descripción de la comunidad microbiana de la fuente termal de Muiño da Veiga mediante metagenómica basada en secuencia y análisis bioinformático

Tras seguir la estrategia basada en la caracterización funcional del metagenoma descrita en los capítulos anteriores, se optó por expandir el estudio de la muestra de la fuente termal de Muiño da Veiga adoptando una metodología de metagenómica basada en secuencia. El análisis del metagenoma con los datos de secuencia ofrece una alternativa que evita la necesidad del uso de hospedadores heterólogos con los consiguientes inconvenientes descritos con anterioridad (44), además de aportar información adicional con respecto a la composición de la comunidad microbiana presente en la muestra y al posible metabolismo que explica las interrelaciones entre los miembros de la misma.

El ADN metagenómico fue extraído de la muestra de aguas termales de Muiño da Veiga siguiendo el procedimiento descrito en el capítulo 1. Tras ello, fue secuenciado por un servicio externo proporcionado por Health In Code (A Coruña, España). Se utilizó la plataforma de secuenciación Illumina HiSeq generando lecturas de extremos emparejados con 100pb de longitud. Los dos archivos resultantes, en formato FASTQ, contienen la información de cada base y la calidad de secuenciación de cada base, y cada secuencia tiene su lectura emparejada en el otro archivo. Las secuencias se filtraron según criterios de calidad usando la secuencia de comandos PrinSeq (45). A continuación, las secuencias emparejadas se unieron para dar lugar a una única secuencia consenso, descartando aquellas que no podían ser emparejadas para evitar alterar los recuentos de secuencia (abundancia de cada lectura), utilizando el programa bioinformático PEAR (46).

Para evaluar el potencial del metagenoma en cuanto al contenido de secuencias codificantes para enzimas putativas con posible interés industrial se realizaron dos alineamientos con dos bases de datos distintas utilizando la secuencia de comandos DIAMOND (47). Se

pudo acceder a una base de datos basada en SEED subsystems (48) y a la base de datos no-redundante (NR) de NCBI (49) a fecha 28 de septiembre del 2016. Solamente los resultados con un identificador único y con el mayor e-valor fueron conservados para su análisis. Los resultados muestran que un 11.94% de los recuentos con la base de datos SEED subsystems son hidrolasas, y se pudieron identificar proteasas, peptidasas, esterasas, lipasas, amilasas, glucosidadas, galactosidadas, enzimas proteolíticas, xilanasas, pululanasas, glucanasas, despolimerasas y celulasas. La base de datos NR reveló un 2.97% de enzimas con potencial biotecnológico en el metagenoma de Muiño da Veiga, identificando proteasas y peptidasas, esterasas y lipasas, amilasas, celulasas, glucosidasas, galactosidasas, pululanasas, xilosidasas, xilanasas, lacasas, depolimerasas, caspasas y glucanasas.

Para estudiar la composición y el metabolismo de la comunidad microbiana de Muiño da Veiga, los datos de lecturas emparejadas se cargaron en el servidor de análisis metagenómico MG-RAST (50). Las bases de datos utilizadas para los análisis en este servidor fueron RefSeq (51) para el análisis taxonómico y KEGG Ortology (KO) (52) y SEED subsystems (53) para los análisis de función y el mapeado de rutas metabólicas usando KEGG. La curva de rarefacción obtenida con las secuencias depositadas en MG-RAST indica que la profundidad de secuenciación ha sido adecuada respecto a la diversidad de especies encontrada en el metagenoma. Usando la base de datos KO, el 60.7% de las lecturas asignadas a una categoría correspondían a “Metabolismo”, un 20.7% a “Procesamiento de Información Génica” y un 11.3% a “Procesamiento de Información Ambiental”. En el caso de la base de datos SEED Subsystems un 12.5% correspondían a “Subsistemas basados en agrupamientos”, 11.1% a “Metabolismo de proteínas”, un 10.1% a “Aminoácidos y derivados”, un 9.2% a “Carbohidratos”, un 6.9% a “Cofactores, Vitaminas, Grupos Prostéticos y Pigmentos”, un 5.9% clasificados como “Misceláneos” y un 5.7% como “Respiración”.

El género más abundante en el metagenoma de Muiño da Veiga es *Sulfurihydrogenibium* (54) (19.3% del total de lecturas) y muchos de los géneros que se encontraron como predominantes en el

metagenoma tenían relación con el metabolismo del azufre. La plasticidad de *Sulfurihydrogenibium* en cuanto a su requerimiento de oxígeno y su metabolismo quimiolitoautótrofo o heterótrofo facultativo, probablemente expliquen su éxito como género predominante de la fuente termal. La alta diversidad de especies encontrada en la fuente termal tiene relación con su pH en el rango neutro y las temperaturas elevadas sin entrar en el rango hipertermófilo, ya que la reducción del pH y la elevación de las temperaturas son desfavorables para la diversidad microbiana en estos ambientes (55). El grupo Aquificae ha sido descrito como el principal fijador de carbono en las aguas termales terrestres (56), con un metabolismo de energía basado en la oxidación de hidrógeno o compuestos reducidos del azufre.

Muchos de los géneros más abundantes son oxidadores de compuestos del azufre: además del descrito *Sulfurihydrogenibium*, los géneros *Aquifex* (57) e *Hidrogenobacter* (58,59) tienen características metabólicas similares, con un anabolismo anaeróbico y un catabolismo aeróbico, fijando CO_2 mediante el ciclo de los ácidos tricarboxílicos, indicando una probable relación de competencia entre especies de ambos géneros. *Thermocrinis* (60) representa otro género termófilo con un metabolismo adaptable que resulta común como miembro abundante de las comunidades de aguas termales (61). El género *Hydrogenobaculum* (62) está normalmente asociado a aguas termales con pH más ácido, y tiene la capacidad de adaptarse a cambios en la temperatura y la concentración de oxígeno (63). *Thiomonas* es otro género con especies acidófilas y crecimiento quimiolitoautótrofo facultativo (64). Por último, *Hydrogenivirga* incluye microorganismos anaerobios o microaerobios con un crecimiento quimiolitoautótrofo estricto (65,66). Por otro lado, numerosas bacterias reductoras del azufre cuyo metabolismo complementa a las anteriores son también géneros dominantes en el metagenoma de Muiño da Veiga. *Thermodesulfovibrio* y *Geobacter* (67,68) comparten nicho metabólico como anaerobios estrictos con capacidad de reducir compuestos del azufre y hierro (III). La capacidad de algunas especies de *Thermodesulfovibrio* para acoplar metanógenos hidrogenotróficos a su metabolismo puede tener

relevancia en la fuente termal debido a la presencia de estos microorganismos en la porción de géneros raros del metagenoma (69). El género *Thermus* es reconocido como uno de los más variados metabólicamente en el hábitat de las fuentes termales, y las especies capaces de reducir compuestos metálicos están ampliamente distribuidas gracias a su capacidad de utilizar un amplio rango de aceptores de electrones (70,71).

A pesar de que gran parte de los géneros dominantes tienen relación con el metabolismo del azufre, otros géneros también abundantes no tienen implicación en dicho metabolismo, y por tanto ocupan nichos para el uso de nutrientes y fuentes de energías alternativos. Se han relacionado especies del género *Acidovorax* (72) con fuentes termales sulfurosas ácidas o ricas en hierro (73–76). Su metabolismo capaz de degradar benzoato y xileno (73) probablemente contribuye a su éxito en esta fuente termal. *Meiothermus* puede metabolizar proteínas y péptidos y algunas especies pueden degradar almidón. Además, pueden usar carbohidratos y algunos compuestos orgánicos como fuentes de carbono y energía (77). El género de anaerobios obligados *Clostridium* debe su éxito de distribución a su plasticidad génica con una amplia capacidad para utilizar diferentes nutrientes (78), incluyendo especies que pueden degradar polímeros complejos (79). También *Caldicellulosiruptor* representa un género de anaerobios estrictos capaces de fermentar monosacáridos, disacáridos y polisacáridos (80).

Aunque el uso de formas del azufre es probablemente el principal motor del metabolismo de la fuente termal, muchos de los géneros dominantes mencionados también están implicados en el ciclo del nitrógeno. *Sulfurihydrogenibium*, *Aquifex*, *Hidrogenivirga*, *Thermodesulfovibrio*, *Geobacter*, algunas especies de *Thermus*, *Acidovorax* y especies de *Meiothermus* pueden usar nitrato como aceptor de electrones y tanto *Hydrogenobacter* como *Clostridum* pueden usar amonio o sales de nitrato como fuentes de nitrógeno. Salvo la ruta de oxidación anaeróbica del amonio (anammox), el resto de los procesos metabólicos relacionados con el ciclo del nitrógeno estaban presentes en el metagenoma. Estos incluyen la denitrificación de nitrato a nitrito a nitrógeno molecular por

Sulfurihydrogenibium (54), *Acidovorax* (72,81), *Aquifex* (57) e *Hydrogenivirga* (66), la reducción de nitrato por *Hydrogenobacter* (82), la fijación de nitrógeno por *Thermodesulfovibrio*, *Thermocrinis*, *Hydrogenobacter* y *Caldicellulosiruptor* (83,84) y la oxidación de amonio a nitrato por la vía de nitrificación, aunque en este caso en géneros raros de la comunidad como los géneros de bacterias *Nitrosospira, Anaeromyxobacter* y *Desulfobacterium* y el género de arqueas *Nitrosopumilus* (85), que en otras fuentes termales tiene un papel más prominente en el ciclo del nitrógeno. Solamente el último paso de la nitrificación estaba representado por géneros dominantes de la comunidad: *Hydrogenobacter*, *Meiothermus*, *Geobacter* y *Acidovorax*. Probablemente la conversión de nitrito a nitrato sea más significativa en esta fuente termal que la nitrificación completa desde el amonio, que puede ser usado por los microorganismos preferentemente como fuente de nitrógeno en lugar de fuente de energía.

El metabolismo del hidrógeno juega un papel clave en condiciones anaeróbicas, como intermediario que conecta varias rutas metabólicas, ya que es generado por bacterias fijadoras del nitrógeno y por el metabolismo fermentativo, y es usado como donador de electrones y consumido por bacterias reductoras del azufre y metanógenos (86,87). Su importancia viene dada por su disponibilidad y por como de termodinámicamente favorable resulta su emparejamiento con aceptores de electrones. Normalmente el equilibrio es favorable al uso de metabolismos basados en hidrógeno en fuentes termales ácidas, mientras que las alcalinas favorecen un metabolismo basado en compuestos del azufre (88,89). Varios géneros dominantes tienen la capacidad de usar hidrógeno incluyendo *Sulfurihydrogenibium*, *Aquifex*, *Hydrogenobacter*, *Thermocrinis*, *Hydrogenobaculum*, algunas especies de *Thiomonas* e *Hydrogenivirga*.

Con relación al metabolismo disimilatorio del carbono, la ruta de los ácidos tricarboxílicos, la gluconeogénesis y la vía de las pentosas-fosfato están presentes en el metagenoma. En las rutas asimilatorias se encuentra un ciclo reverso de los ácidos tricarboxílicos y el ciclo de Calvin. El hecho de que muchos géneros dominantes pertenecen al

grupo Aquificae, los principales fijadores de CO_2 por la vía del ciclo reverso de los ácidos tricarboxílicos, indica la importancia de esta ruta en particular en la fuente termal. La posibilidad de usar sustratos de diferente complejidad, y en especial de carbohidratos complejos como la celulosa y el xilano, revelan una posible fuente de metabolitos con interés biotecnológico.

Por último, la biosfera rara, tradicionalmente ignorada en favor de los géneros dominantes, merece atención ya que puede jugar papeles metabólicos claves, aparte de representar un depósito de diversidad genética y funcional que aporta resiliencia ecológica a la comunidad microbiana (90,91). Además, en ambientes cuyas condiciones ambientales pueden sufrir variaciones, algunas de estas especies raras pueden convertirse en especies predominantes (91,92). En el metagenoma se encontraron dos rutas metabólicas relacionadas con la fotosíntesis, cuyo límite de temperatura se considera 73°C en hábitats no ácidos (93), por encima de los 68°C que presenta la fuente termal de Muiño da Veiga. Sin embargo, la pertenencia a la biosfera rara y la presencia de otras rutas autotróficas en el metabolismo de la comunidad microbiana indican que el papel de la fotosíntesis seguramente sea minoritario en Muiño da Veiga. Por otro lado, se encontraron géneros del dominio Archaea en el metagenoma, también como parte de la biosfera rara, muchos de ellos anaerobios estrictos y metanógenos. Todas las rutas relacionadas con el metabolismo del metano estaban presentes al menos parcialmente, y el género de arqueas más abundante *Methanosarcina* es considerado uno de los metanógenos más diversos metabólicamente al contar con las cuatro rutas catabólicas (94). Además de otras arqueas metanógenas, se encontraron géneros cuyas características metabólicas eran similares a las de las bacterias predominantes, lo que probablemente explica su reducida abundancia en el metagenoma debido a la competencia por los mismos recursos.

Capítulo 4: Análisis de secuencia, clonado, purificación y caracterización de una nueva endoglucanasa termoestable y análisis de secuencia y clonado de una β-xilosidasa putativa encontradas por metagenómica basada en secuencia

Además de permitir estudiar la composición y las interacciones metabólicas de las comunidades microbianas, los datos de secuenciación masiva pueden ser utilizados para realizar predicciones y encontrar nuevas termozimas codificadas en los metagenomas (11,16,44,95). En este capítulo se utilizaron los datos de secuenciación del metagenoma de Muiño da Veiga para encontrar enzimas relacionadas con la degradación de material lignocelulósico, que encuentran utilidad en una gran variedad de industrias y que muchas veces suponen una alternativa más ecológica y barata a los procesos tradicionales. En concreto, la búsqueda se centró en encontrar posibles termozimas con actividad endoglucanasa y β-xilosidasa. Las endoglucanasas son enzimas que actúan sobre la celulosa, el principal polímero estructural de las paredes celulares vegetales (96), con aplicaciones que van desde los detergentes y los tratamientos de prendas de algodón a la producción de biocombustibles (97–100). Las β-xilosidasas son enzimas que actúan sobre xilooligosacáridos, productos de la degradación del xilano, un componente de la hemicelulosa que se entrecruza en la estructura de celulosa y la hace más resistente a la hidrólisis (101–104). Además de las aplicaciones directas de las β-xilosidasas (103,105), el producto de la reacción (xilosa) puede emplearse a su vez en variedad de procesos productivos (103,105). Además, la acción conjunta de ambas enzimas e incluso de otras enzimas accesorias también suscita interés por sus actividades sinérgicas frente a estos polímeros (104,106).

Las lecturas de secuenciación del capítulo anterior se utilizaron para reconstruir secuencias más largas o “contigs” usando un ensamblador *de novo* basado en gráficos de De Bruijn IDBA-UD (107). Para eliminar secuencias contaminantes procedentes del genoma

humano, se utilizó la secuencia de comandos Deconseq (108) con los datos de NCBI GRCh38 (RefSeq ID 884148) como base de datos de referencia. Los datos fueron cargados al servidor de análisis de metagenomas MG-RAST (50) para analizarlos frente a la base de datos SEED Subsystems (53). Se recuperaron las secuencias que en el nivel jerárquico "función" contenían el término "Endoglucanase (EC 3.2.1.4)" y "Beta-xylosidase (EC 3.2.1.37)". Las secuencias que contenían los genes que codificaban para estas proteínas putativas se analizaron usando la herramienta ExPASy Translate Tool (109) para encontrar los marcos de lectura abiertos, y con NCBI BlastP (40) usando los parámetros por defecto y la base de datos NR de proteínas. Se seleccionó una enzima putativa de cada clase: Cel776 como endoglucanasa y Xyl7992 como β-xilosidasa; se realizaron predicciones usando la herramienta ProtParam de ExPASy (109); se crearon modelos tridimensionales usando la herramienta SWISS-MODEL (110) y se visualizaron usando el programa PyMOL (Schrödinger LCC) (111). Además, se recuperaron secuencias de la base de datos Uniprot Knowledgebase (112) para realizar alineamientos múltiples de secuencia usando la herramienta de EMBL-EBI Clustal Omega (41).

Xyl7792 se clasificó como una β-xilosidasa putativa de la familia glicosil hidrolasa 43 subfamilia 12, con alta identidad (99%) con la proteína glicósido hidrolasa familia 43 de *Porphyrobacter cryptus*. El modelado tridimensional reveló una estructura de dominio catalítico en propulsor beta de cinco láminas, y un dominio no catalítico en sándwich beta. Se identificaron residuos conservados y en especial la triada catalítica Asp69, Asp175 y Glu240.

Cel776 se clasificó como una endoglucanasa putativa de la familia GH5 subfamilia 25, con un 73% de identidad con la proteína glicosil hidrolasa 1 de *Fervidobacterium pennivorans*. La estructura tridimensional reveló un barril $(\beta/\alpha)_8$-TIM con 8 láminas beta y 8 hélices alfa, y 4 hélices alfa adicionales. Los residuos conservados fueron identificados y en especial la pareja de donador de protones Glu144 y el nucleófilo Glu260.

Se encargó la síntesis de los genes seleccionados, usando los servicios de síntesis de ThermoScientific. A los genes se les añadieron secuencias adaptadoras a ambos extremos que permitían el clonado por medio del sistema Gateway (Invitrogen) basado en recombinaciones homólogas y vectores que incluyen regiones para la recombinación. Los genes sintetizados fueron clonados en primer lugar en el vector de entrada pDNOR221 y a continuación en el vector de destino pDEST527, que contiene una cola de histidinas para facilitar la purificación de las proteínas expresadas. Las células usadas para las transformaciones fueron *E. coli* OneShot™ OmniMAX™ 2 T1^{R} (Invitrogen) y *E. coli* T7 Express (New England Biolabs) para la expresión final. Cada paso de transformación fue comprobado por secuenciación (por el servicio proporcionado por la Unidad de Biología Molecular de los Servicios de Apoyo a la Investigación de la Universidade da Coruña) usando la pareja de cebadores adecuada para cada vector (M13 para el vector pDNOR221 y T7 para el vector pDEST527). Se indujo la expresión de las proteínas clonadas usando 0.4mM IPTG durante 2h a 37°C, y se procedió a la lisis celular usando un protocolo de sonicación usando un VCX130 Vibra-Cell sonicator (SONICS & MATERIALS INC., USA) y alternativamente un protocolo de lisis químico-enzimática con el preparado comercial NZYLysis (NZYTECH).

Los experimentos de expresión utilizando la construcción pDEST527 con el inserto Xyl7992 y un protocolo para cuantificar la actividad enzimática β-xilosidasa usando el sustrato pNP-*xylopyranoside* (113–115) no resultaron en una enzima soluble que pudiera ser purificada o caracterizada. La explicación más probable es la formación de cuerpos de inclusión comúnmente asociada a la expresión heteróloga de proteínas en *E. coli*.

Se consiguió expresar y purificar la enzima endoglucanasa Cel776 con una columna con resina niquel-sefarosa HisTrap 5mL (Cytiva) con un *buffer* de elución que fue añadido en concentraciones crecientes de imidazol. Los resultados de purificación se cuantificaron con el método de Bradford para determinar concentraciones de proteínas (116), y se visualizaron mediante SDS-PAGE y Western-blot. La proteína purificada se concentró usando una columna Pierce

Concentrator 10K Molecular Weight Cut-Off (THERMO SCIENTIFIC, UK). La actividad endoglucanasa se determinó usando el sustrato carboximetilcelulosa (CMC) y calculando los azúcares reductores liberados usando ácido dinitrosalicílico (DNS) (117).

La enzima Cel776 además de su actividad endoglucanasa presentaba actividad xilanasa, con una temperatura óptima de 80°C y un pH óptimo de 5. Se demostró que era una enzima termoestable manteniendo su actividad durante más de 24h a 60°C, y con una vida media de 3.55h a 80°C. La enzima era inhibida por zinc y cobre, y no se veía afectada por el quelante EDTA, de forma similar a otras endoglucanasas descritas (118–120). También se observó que la actividad enzimática se veía potenciada por calcio y magnesio, un efecto observado en otras endoglucanasas debido a la estabilización de la estructura de la proteína (120–123). Aunque la mayoría de detergentes eran tolerados por la enzima, el SDS inhibía la actividad endoglucanasa, un efecto también descrito en otras endoglucanasas (119).

Conclusiones

1. Se construyeron dos metagenotecas, una de 27,898 clones de la muestra de As Burgas y una de 4,399 clones de Muiño da Veiga, usando como hospedador la bacteria *E. coli* EPI300 $T1^R$ y el vector pCT3FK inducible por arabinosa, creando copias para su conservación a largo plazo y para realizar cribados funcionales.
2. Se identificaron mediante cribados funcionales clones que presentaban actividades enzimáticas de interés en ambas metagenotecas: 2 endoglucanasas, 5 β-glucosidasas, 2 β-xilosidasas, 4 proteasas y 1 enzima con actividad lipolítica en As Burgas, y 4 endoglucanasas, 1 β-glucosidasa, 6 proteasas y 8 enzimas con actividad lipolítica en Muiño da Veiga.
3. La enzima con actividad lipolítica LipB12_A11 fue secuenciada, clonada, expresada y purificada. Se identificó como una proteína hidrolítica de la familia SNGH/GDSL por el análisis de secuencia, con dominios y aminoácidos conservados. La preferencia por sustratos cortos (C8 y C12) permite clasificarla como una esterasa, con temperatura óptima de 60°C, pH óptimo de 9 y elevada termoestabilidad.
4. El metagenoma de Muiño da Veiga fue secuenciado y se pudo predecir la presencia de enzimas con interés biotecnológico. La comunidad microbiana fue descrita, siendo el género más abundante *Sulfurihydrogenibium*. Los géneros más abundantes correspondían a bacterias relacionadas con el ciclo del azufre, pero rutas para el ciclo del nitrógeno y del hidrógeno también estaban presentes. También se detectó la capacidad de utilizar numerosas fuentes de carbono, incluida la biotecnológicamente relevante capacidad para aprovechar material lignocelulósico. El análisis de la biosfera rara reveló la posibilidad de que organismos fotosintéticos y metanógenos contribuyeran a la variedad metabólica observada en la fuente termal.
5. El metagenoma fue ensamblado reconstruyendo secuencias más largas, o contigs. Se pudieron identificar varias pautas de

lectura abiertas (ORF) que codificaban para posibles endoglucanasas y β-xilosidasas. La β-xilosidasa putativa Xyl7992 pertenece a la familia glicosil hidrolasa GH43 subfamilia 12 y en base al modelado por homología presenta una estructura propulsor beta de 5 láminas como dominio catalítico y un sándwich de láminas beta no catalítico, ambas con residuos de aminoácidos conservados. Por otro lado, la endoglucanasa putativa Cel776 pertenece a la familia glicosil hidrolasa 25 subfamilia 5 y su estructura tridimensional en base al modelado por homología es de barril $(\beta/\alpha)_8$. También se encontraron aminoácidos conservados en la secuencia primaria de esta proteína.

6. Se clonó la enzima Xyl7992 en un sistema de expresión con el hospedador *E. coli* T7 Express y el vector pJET1.2, pero la formación de cuerpos de inclusión impidió que se pudiera caracterizar esta enzima. La enzima Cel776 fue clonada, expresada y purificada usando el mismo sistema, y se pudo realizar una caracterización bioquímica. La enzima mostraba su mayor actividad con el sustrato carboximetil-celulosa (CMC) confirmando que se trataba de una endoglucanasa, pero también demostró la capacidad de degradar xilano. La enzima es termoestable, con un óptimo de temperatura y pH de 80°C y 5 respectivamente. Se observó un efecto potenciador de la actividad enzimática con la adición de calcio y magnesio.

Referencias

1. González-Barreiro C, Cancho-Grande B, Araujo-Nespereira P, Cid-Fernández JA, Simal-Gándara J. Occurrence of soluble organic compounds in thermal waters by ion trap mass detection. Chemosphere [Internet]. 2009 Mar;75(1):34–47. Available from: http://dx.doi.org/10.1016/j.chemosphere.2008.11.067

2. López DL, Araujo PA, Outeiriño ID, Cid JA, Astray G. Geochemical signatures of the groundwaters from Ourense thermal springs, Galicia, Spain. Sustain Water Resour Manag [Internet]. 2019 Mar 5;5(1):103–16. Available from: http://dx.doi.org/10.1007/s40899-018-0239-3

3. Delgado-Outeiriño I, Araujo-Nespereira P, Cid-Fernández JA, Mejuto JC, Martínez-Carballo E, Simal-Gándara J. Behaviour of thermal waters through granite rocks based on residence time and inorganic pattern. J Hydrol [Internet]. 2009 Jul;373(3–4):329–36. Available from: http://dx.doi.org/10.1016/j.jhydrol.2009.04.028

4. Ekkers DM, Cretoiu MS, Kielak AM, van Elsas JD. The great screen anomaly—a new frontier in product discovery through functional metagenomics. Appl Microbiol Biotechnol [Internet]. 2012 Feb 22;93(3):1005–20. Available from: http://link.springer.com/10.1007/s00253-011-3804-3

5. Ferrer M, Golyshina O, Beloqui A, Golyshin PN. Mining enzymes from extreme environments. Curr Opin Microbiol [Internet]. 2007 Jun;10(3):207–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369527407000513

6. Reigstad LJ, Bartossek R, Schleper C. Preparation of High-Molecular Weight DNA and Metagenomic Libraries from Soils and Hot Springs. In: Methods in Enzymology [Internet]. 1st ed. Elsevier Inc.; 2011. p. 319–44. Available from: http://dx.doi.org/10.1016/B978-0-12-386489-5.00013-0

7. Angelov A, Mientus M, Liebl S, Liebl W. A two-host fosmid system for functional screening of (meta)genomic libraries

from extreme thermophiles. Syst Appl Microbiol [Internet]. 2009 May;32(3):177–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0723202009000137

8. DeCastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol [Internet]. 2016 Sep 27;7(SEP):1–21. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01521/abstract

9. Felczykowska A, Krajewska A, Zielińska S, Łoś JM, Bloch SK, Nejman-Faleńczyk B. The most widespread problems in the function-based microbial metagenomics. Acta Biochim Pol [Internet]. 2015 Feb 13;62(1):161–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25680374

10. Escuder-Rodríguez J-J, DeCastro M-E, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Advances of Functional Metagenomics in Harnessing Thermozymes. Metagenomics: Perspectives, Methods, and Applications. 2017.

11. Mardanov A V., Kadnikov V V., Ravin N V. Metagenomics: A Paradigm Shift in Microbiology. In: Metagenomics [Internet]. Elsevier; 2018. p. 1–13. Available from: https://doi.org/10.1016/B978-0-08-102268-9.00001-X

12. Armstrong Z, Rahfeld P, Withers SG. Discovery of New Glycosidases From Metagenomic Libraries. In: Methods in Enzymology [Internet]. 1st ed. Elsevier Inc.; 2017. p. 3–23. Available from: http://dx.doi.org/10.1016/bs.mie.2017.06.001

13. Decastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. Front Microbiol. 2016;7(September):1521.

14. Guazzaroni ME, Silva-Rocha R, Ward RJ. Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. Microb Biotechnol.

2015;8(1):52–64.

15. Taupp M, Mewis K, Hallam SJ. The art and design of functional metagenomic screens. Curr Opin Biotechnol [Internet]. 2011 Jun;22(3):465–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166911000371

16. Berini F, Casciello C, Marcone GL, Marinelli F. Metagenomics: novel enzymes from non-culturable microbes. FEMS Microbiol Lett [Internet]. 2017 Nov 15;364(21). Available from: https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnx211/4329276

17. Ferrer M, Martínez-Martínez M, Bargiela R, Streit WR, Golyshina O V, Golyshin PN. Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microb Biotechnol [Internet]. 2016 Jan;9(1):22–34. Available from: http://doi.wiley.com/10.1111/1751-7915.12309

18. Pandey S, Singh S, Yadav AN, Nain L, Saxena AK. Phylogenetic diversity and characterization of novel and efficient cellulase producing bacterial isolates from various extreme environments. Biosci Biotechnol Biochem [Internet]. 2013;77(7):1474–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23832366

19. Farrow MF, Arnold FH. High Throughput Screening of Fungal Endoglucanase Activity in Escherichia coli. J Vis Exp. 2011;21(54):1–4.

20. Zhang L, Fan Y, Zheng H, Du F, Zhang KQ, Huang X, et al. Isolation and characterization of a novel endoglucanase from a Bursaphelenchus xylophilus metagenomic library. PLoS One. 2013;8(12):1–8.

21. Berlemont R, Delsaute M, Pipers D, D'Amico S, Feller G, Galleni M, et al. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. ISME J [Internet]. 2009;3(9):1070–81. Available from:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19458657&retmode=ref&cmd=prlinks%5Cnpapers2://publication/doi/10.1038/ismej.2009.48

22. Ilmberger N, Streit WR. Metagenomics [Internet]. Streit WR, Daniel R, editors. Metagenomics: Methods and Protocols. Totowa, NJ: Humana Press; 2010. 177–188 p. (Methods in Molecular Biology; vol. 668). Available from: http://link.springer.com/10.1007/978-1-60761-823-2

23. Kim Y-K, Lee S-C, Cho Y-Y, Oh H-J, Ko YH. Isolation of Cellulolytic Bacillus subtilis Strains from Agricultural Environments. ISRN Microbiol. 2012;2012:1–9.

24. Berlemont R, Pipers D, Delsaute M, Angiono F, Feller G, Galleni M, et al. Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements. Rev Argent Microbiol. 2011;43(2):94–103.

25. Kasana RC, Salwan R, Dhar H, Dutt S, Gulati A. A rapid and easy method for the detection of microbial cellulases on agar plates using Gram's iodine. Curr Microbiol. 2008;57(5):503–7.

26. Knapik K, Becerra M, González-Siso M-I. Microbial diversity analysis and screening for novel xylanase enzymes from the sediment of the Lobios Hot Spring in Spain. Sci Rep [Internet]. 2019 Dec 1;9(1):11195. Available from: http://www.nature.com/articles/s41598-019-47637-z

27. Anderson I, Abt B, Lykidis A, Klenk H-P, Kyrpides N, Ivanova N. Genomics of Aerobic Cellulose Utilization Systems in Actinobacteria. Ben-Jacob E, editor. PLoS One [Internet]. 2012 Jun 18;7(6):e39331. Available from: https://dx.plos.org/10.1371/journal.pone.0039331

28. Nyyssönen M, Tran HM, Karaoz U, Weihe C, Hadi MZ, Martiny JBH, et al. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. Front Microbiol [Internet]. 2013;4(SEP):1–14. Available from:

http://journal.frontiersin.org/article/10.3389/fmicb.2013.00282/abstract

29. Blank S, Schröder C, Schirrmacher G, Reisinger C AG. Biochemical Characterization of a Recombinant xylanase from Thermus brockianus, Suitable for Biofuel Production. JSM Biotechnol Bioeng. 2014;2(1):1027.

30. Pinheiro GL, Correa RF, Cunha RS, Cardoso AM, Chaia C, Clementino MM, et al. Isolation of aerobic cultivable cellulolytic bacteria from different regions of the gastrointestinal tract of giant land snail Achatina fulica. Front Microbiol [Internet]. 2015 Aug 20;6(AUG):1–15. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00860/abstract

31. Ohlhoff CW, Kirby BM, Van Zyl L, Mutepfa DLR, Casanueva A, Huddy RJ, et al. An unusual feruloyl esterase belonging to family VIII esterases and displaying a broad substrate range. J Mol Catal B Enzym [Internet]. 2015 Aug;118:79–88. Available from: http://dx.doi.org/10.1016/j.molcatb.2015.04.010

32. Hårdeman F, Sjöling S. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. FEMS Microbiol Ecol. 2007;59(2):524–34.

33. Leis B, Angelov A, Mientus M, Li H, Pham VTT, Lauinger B, et al. Identification of novel esterase-active enzymes from hot environments by use of the host bacterium Thermus thermophilus. Front Microbiol [Internet]. 2015 Apr 8;6(April):275. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25904908

34. Selvin J, Kennedy J, Lejon DPH, Kiran S, Dobson ADW. Isolation identification and biochemical characterization of a novel halo-tolerant lipase from the metagenome of the marine sponge Haliclona simulans. Microb Cell Fact [Internet]. 2012;11(1):72. Available from: Microbial Cell Factories

35. Sang SL, Li G, Hu XP, Liu YH. Molecular Cloning, Overexpression and Characterization of a Novel Feruloyl Esterase from a Soil Metagenomic Library. J Mol Microbiol Biotechnol [Internet]. 2011;20(4):196–203. Available from: https://www.karger.com/Article/FullText/329833

36. Andersen A, Svendsen A, Vind J, Lassen S., Hjort C, Borch K, et al. Studies on ferulic acid esterase activity in fungal lipases and cutinases. Colloids Surfaces B Biointerfaces [Internet]. 2002 Sep;26(1–2):47–55. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0927776501003125

37. Chahiniana H, Sarda L. Distinction Between Esterases and Lipases: Comparative Biochemical Properties of Sequence-Related Carboxylesterases. Protein Pept Lett [Internet]. 2009 Oct 1;16(10):1149–61. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8665&volume=16&issue=10&spage=1149

38. Fojan P, Jonson PH, Petersen MTN, Petersen SB. What distinguishes an esterase from a lipase: A novel structural approach. Biochimie. 2000;82(11):1033–41.

39. Jo E, Kim J, Lee A, Moon K, Cha J. Identification and Characterization of a Novel Thermostable GDSL-Type Lipase from Geobacillus thermocatenulatus. J Microbiol Biotechnol [Internet]. 2021 Mar 28;31(3):483–91. Available from: http://www.jmb.or.kr/journal/view.html?doi=10.4014/jmb.2012.12036

40. NCBI_Resource_Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D7–19. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1290

41. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol [Internet]. 2011 Jan 11;7(1):539. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21988835

42. Akoh CC, Lee GC, Liaw YC, Huang TH, Shaw JF. GDSL family of serine esterases/lipases. Prog Lipid Res. 2004;43(6):534–52.

43. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res [Internet]. 2004 Jun;14(6):1188–90. Available from: ftp://ftp.ncbi.nih.gov/genomes/Bacteria

44. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol [Internet]. 2012 Apr;21(8):1794–805. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2012.05538.x

45. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics [Internet]. 2011 Mar 15;27(6):863–4. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr026

46. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics [Internet]. 2014 Mar 1;30(5):614–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt593

47. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods [Internet]. 2015 Jan 17;12(1):59–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25402007

48. Silva GGZ, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics [Internet]. 2016 Feb 1;32(3):354–61. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv584

49. Bethesta MD, National Library of Medicine US. National Center for Biotechnology Information (NCBI) [Internet] [Internet]. [cited 2016 Sep 28]. Available from: https://www.ncbi.nlm.nih.gov/

50. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics [Internet]. 2008 Dec 19;9(1):386. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-386

51. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D733–45. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189

52. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D457-62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26476454

53. Overbeek R, Begley T, Butler RM, Choudhuri J V., Chuang H-Y, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res [Internet]. 2005;33(17):5691–702. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16214803

54. Takai K, Kobayashi H, Nealson KH, Horikoshi K. Sulfurihydrogenibium subterraneum gen. nov., sp. nov., from a subsurface hot aquifer. Int J Syst Evol Microbiol [Internet]. 2003 May 1;53(3):823–7. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02506-0

55. Sahm K, John P, Nacke H, Wemheuer B, Grote R, Daniel R, et al. High abundance of heterotrophic prokaryotes in hydrothermal springs of the Azores as revealed by a network of 16S rRNA gene-based methods. Extremophiles [Internet]. 2013 Jul 26;17(4):649–62. Available from: http://link.springer.com/10.1007/s00792-013-0548-2

56. Dodsworth JA, Hungate BA, Hedlund BP. Ammonia oxidation, denitrification and dissimilatory nitrate reduction to ammonium in two US Great Basin hot springs with abundant ammonia-oxidizing archaea. Environ Microbiol [Internet]. 2011 Aug;13(8):2371–86. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2011.02508.x

57. Huber R, Wilharm T, Huber D, Trincone A, Burggraf S, König H, et al. Aquifex pyrophilus gen. nov. sp. nov., Represents a Novel Group of Marine Hyperthermophilic Hydrogen-Oxidizing Bacteria. Syst Appl Microbiol. 1992;15(3):340–51.

58. Bonjour F, Aragno M. Growth of thermophilic, obligatorily chemolithoautotrophic hydrogen-oxidizing bacteria related to Hydrogenobacter with thiosulfate and elemental sulfur as electron and energy source. FEMS Microbiol Lett. 1986;35:11–5.

59. Alfredsson GA, Ingason A, Kristjansson JK. Growth of thermophilic obligately autotrophic hydrogen-oxidizing bacteria on thiosulfate. Lett Appl Microbiol [Internet]. 1986 Aug;2(2):21–3. Available from: http://doi.wiley.com/10.1111/j.1472-765X.1986.tb01507.x

60. Huber R, Eder W, Heldwein S, Wanner G, Huber H, Rachel R, et al. Thermocrinis ruber gen. nov., sp. nov., a Pink-Filament-Forming Hyperthermophilic Bacterium Isolated from Yellowstone National Park. Appl Environ Microbiol [Internet]. 1998 Oct 1;64(10):3576–83. Available from: https://aem.asm.org/content/64/10/3576

61. Dodsworth JA, Ong JC, Williams AJ, Dohnalkova AC, Hedlund BP. Thermocrinis jamiesonii sp. nov., a thiosulfate-oxidizing, autotropic thermophile isolated from a geothermal spring. Int J Syst Evol Microbiol [Internet]. 2015 Dec 1;65(Pt_12):4769–75. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.000647

62. Massello FL, Chan CS, Chan K-G, Goh KM, Donati E, Urbieta MS. Meta-Analysis of Microbial Communities in Hot Springs: Recurrent Taxa and Complex Shaping Factors beyond pH and

Temperature. Microorganisms [Internet]. 2020 Jun 16;8(6):906. Available from: https://www.mdpi.com/2076-2607/8/6/906

63. Lin K-H, Liao B-Y, Chang H-W, Huang S-W, Chang T-Y, Yang C-Y, et al. Metabolic characteristics of dominant microbes and key rare species from an acidic hot spring in Taiwan revealed by metagenomics. BMC Genomics [Internet]. 2015 Dec 3;16(1):1029. Available from: http://dx.doi.org/10.1186/s12864-015-2230-9

64. Moreira D, Amils R. Phylogeny of Thiobacillus cuprinus and Other Mixotrophic Thiobacilli: Proposal for Thiomonas gen. nov. Int J Syst Bacteriol [Internet]. 1997 Apr 1;47(2):522–8. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-47-2-522

65. Nakagawa S, Nakamura S, Inagaki F, Takai K, Shirai N, Sako Y. Hydrogenivirga caldilitoris gen. nov., sp. nov., a novel extremely thermophilic, hydrogen- and sulfur-oxidizing bacterium from a coastal hydrothermal field. Int J Syst Evol Microbiol [Internet]. 2004 Nov 1;54(6):2079–84. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.03031-0

66. Nunoura T, Miyazaki M, Suzuki Y, Takai K, Horikoshi K. Hydrogenivirga okinawensis sp. nov., a thermophilic sulfur-oxidizing chemolithoautotroph isolated from a deep-sea hydrothermal field, Southern Okinawa Trough. Int J Syst Evol Microbiol [Internet]. 2008 Mar 1;58(3):676–81. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.64615-0

67. Lovley DR, Giovannoni SJ, White DC, Champine JE, Phillips EJP, Gorby YA, et al. Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. Arch Microbiol [Internet]. 1993 Apr;159(4):336–44. Available from:

http://link.springer.com/10.1007/BF00290916

68. Henry EA, Devereux R, Maki JS, Gilmour CC, Woese CR, Mandelco L, et al. Characterization of a new thermophilic sulfate-reducing bacterium. Arch Microbiol [Internet]. 1994 Jan;161(1):62–9. Available from: http://link.springer.com/10.1007/BF00248894

69. Sekiguchi Y, Muramatsu M, Imachi H, Narihiro T, Ohashi A, Harada H, et al. Thermodesulfovibrio aggregans sp. nov. and Thermodesulfovibrio thiophilus sp. nov., anaerobic,thermophilic, sulfate-reducing bacteria isolated from thermophilic methanogenic sludge, and emended description of the genus Thermodesulfovibrio. Int J Syst Evol Microbiol. 2008;58(11):2541–8.

70. Zhou E-M, Adegboruwa AL, Mefferd CC, Bhute SS, Murugapiran SK, Dodsworth JA, et al. Diverse respiratory capacity among Thermus strains from US Great Basin hot springs. Extremophiles [Internet]. 2020 Jan 18;24(1):71–80. Available from: https://doi.org/10.1007/s00792-019-01131-6

71. Balkwill DL, Kieft TL, Tsukuda T, Kostandarithes HM, Onstott TC, Macnaughton S, et al. Identification of iron-reducing Thermus strains as Thermus scotoductus. Extremophiles. 2004;8(1):37–44.

72. Willems A, Falsen E, Pot B, Jantzen E, Hoste B, Vandamme P, et al. Acidovorax, a New Genus for Pseudomonas facilis, Pseudomonas delafieldii, E. Falsen (EF) Group 13, EF Group 16, and Several Clinical Isolates, with the Species Acidovorax facilis comb. nov., Acidovorax delafieldii comb. nov., and Acidovorax temperans sp. . Int J Syst Bacteriol [Internet]. 1990 Oct 1;40(4):384–98. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-40-4-384

73. Saxena R, Dhakan DB, Mittal P, Waiker P, Chowdhury A, Ghatak A, et al. Metagenomic Analysis of Hot Springs in Central India Reveals Hydrocarbon Degrading Thermophiles and Pathways Essential for Survival in Extreme Environments. Front Microbiol [Internet]. 2017 Jan 5 [cited

2018 Feb 26];7:2123. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2016.02123/full

74. Padilla-Del Valle R, Morales-Vale LR, Ríos-Velázquez C. Unraveling the microbial and functional diversity of Coamo thermal spring in Puerto Rico using metagenomic library generation and shotgun sequencing. Genomics Data [Internet]. 2017 Mar;11:98–101. Available from: http://dx.doi.org/10.1016/j.gdata.2016.12.010

75. Kozubal MA, Macur RE, Jay ZJ, Beam JP, Malfatti SA, Tringe SG, et al. Microbial Iron Cycling in Acidic Geothermal Springs of Yellowstone National Park: Integrating Molecular Surveys, Geochemical Processes, and Isolation of Novel Fe-Active Microorganisms. Front Microbiol [Internet]. 2012;3(MAR):1–16. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2012.00109/abstract

76. Selvarajan R, Sibanda T, Tekere M. Thermophilic bacterial communities inhabiting the microbial mats of “indifferent” and chalybeate (iron-rich) thermal springs: Diversity and biotechnological analysis. Microbiologyopen [Internet]. 2018 Apr;7(2):e00560. Available from: http://doi.wiley.com/10.1002/mbo3.560

77. Raposo P, Viver T, Albuquerque L, Froufe H, Barroso C, Egas C, et al. Transfer of Meiothermus chliarophilus (Tenreiro et al.1995) Nobre et al. 1996, Meiothermus roseus Ming et al. 2016, Meiothermus terrae Yu et al. 2014 and Meiothermus timidus Pires et al. 2005, to Calidithermus gen. nov., as Calidithermus chliarophilus com. Int J Syst Evol Microbiol [Internet]. 2019 Apr 1;69(4):1060–9. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003270

78. Udaondo Z, Duque E, Ramos J-L. The pangenome of the genus Clostridium. Environ Microbiol [Internet]. 2017 Jul;19(7):2588–603. Available from: http://doi.wiley.com/10.1111/1462-2920.13732

79. Liu L, Jiao J-Y, Fang B-Z, Lv A-P, Ming Y-Z, Li M-M, et al. Isolation of Clostridium from Yunnan-Tibet hot springs and description of Clostridium thermarum sp. nov. with lignocellulosic ethanol production. Syst Appl Microbiol [Internet]. 2020 Sep;43(5):126104. Available from: https://doi.org/10.1016/j.syapm.2020.126104

80. Rainey FA, Donnison AM, Janssen PH, Saul D, Rodrigo A, Bergquist PL. Description of Caldicellulosiruptor saccharolyticus. FEMS Microbiol Lett. 1994;120:263–6.

81. Heylen K, Lebbe L, De Vos P. Acidovorax caeni sp. nov., a denitrifying species with genetically diverse isolates from activated sludge. Int J Syst Evol Microbiol [Internet]. 2008 Jan 1;58(1):73–7. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.65387-0

82. Zeytun A, Sikorski J, Nolan M, Lapidus A, Lucas S, Han J, et al. Complete genome sequence of Hydrogenobacter thermophilus type strain (TK-6T). Stand Genomic Sci [Internet]. 2011 Apr 29;4(2):131–43. Available from: http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.1463589

83. Nishihara A, Thiel V, Matsuura K, McGlynn SE, Haruta S. Phylogenetic Diversity of Nitrogenase Reductase Genes and Possible Nitrogen-Fixing Bacteria in Thermophilic Chemosynthetic Microbial Communities in Nakabusa Hot Springs. Microbes Environ [Internet]. 2018;33(4):357–65. Available from: https://www.jstage.jst.go.jp/article/jsme2/33/4/33_ME18030/_article

84. Hamilton TL, Koonce E, Howells A, Havig JR, Jewell T, de la Torre JR, et al. Competition for ammonia influences the structure of chemotrophic communities in geothermal springs. Appl Environ Microbiol. 2014;80(2):653–61.

85. Pester M, Rattei T, Flechl S, Gröngröft A, Richter A, Overmann J, et al. AmoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA

genes from soils of four different geographic regions. Environ Microbiol. 2012;14(2):525–39.

86. Revsbech NP, Trampe E, Lichtenberg M, Ward DM, Kühl M. In Situ Hydrogen Dynamics in a Hot Spring Microbial Mat during a Diel Cycle. Spormann AM, editor. Appl Environ Microbiol [Internet]. 2016 Jul 15;82(14):4209–17. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.00710-16

87. Otaki H, Everroad RC, Matsuura K, Haruta S. Production and Consumption of Hydrogen in Hot Spring Microbial Mats Dominated by a Filamentous Anoxygenic Photosynthetic Bacterium. Microbes Environ [Internet]. 2012;27(3):293–9. Available from: https://www.jstage.jst.go.jp/article/jsme2/27/3/27_ME11348/_article

88. Lindsay MR, Colman DR, Amenabar MJ, Fristad KE, Fecteau KM, Debes R V., et al. Probing the geological source and biological fate of hydrogen in Yellowstone hot springs. Environ Microbiol. 2019;21(10):3816–30.

89. Fenchel T, King GM, Blackburn TH. Bacterial Metabolism. In: Bacterial Biogeochemistry [Internet]. Elsevier; 2012. p. 1–34. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780124158368000013

90. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol [Internet]. 2015 Apr 2;13(4):217–29. Available from: http://dx.doi.org/10.1038/nrmicro3400

91. Yang S, Winkel M, Wagner D, Liebner S. Community structure of rare methanogenic archaea: insight from a single functional group. FEMS Microbiol Ecol [Internet]. 2017 Nov 1;93(11):1–10. Available from: https://academic.oup.com/femsec/article/doi/10.1093/femsec/fix126/4331631

92. Wang Y, Hatt JK, Tsementzi D, Rodriguez-R LM, Ruiz-Pérez

CA, Weigand MR, et al. Quantifying the Importance of the Rare Biosphere for Microbial Community Response to Organic Pollutants in a Freshwater Ecosystem. Stams AJM, editor. Appl Environ Microbiol [Internet]. 2017 Apr 15;83(8):1–19. Available from: https://aem.asm.org/lookup/doi/10.1128/AEM.03321-16

93. Boyd ES, Leavitt WD, Geesey GG. CO2 Uptake and Fixation by a Thermoacidophilic Microbial Community Attached to Precipitated Sulfur in a Geothermal Spring. Appl Environ Microbiol [Internet]. 2009 Jul 1;75(13):4289–96. Available from: https://aem.asm.org/content/75/13/4289

94. Maeder DL, Anderson I, Brettin TS, Bruce DC, Gilna P, Han CS, et al. The Methanosarcina barkeri genome: Comparative analysis with Methanosarcina acetivorans and Methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes. J Bacteriol. 2006;188(22):7922–31.

95. Galloway-Peña J, Hanson B. Tools for Analysis of the Microbiome. Dig Dis Sci [Internet]. 2020 Mar 31;65(3):674–85. Available from: http://link.springer.com/10.1007/s10620-020-06091-y

96. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS. Microbial Cellulose Utilization: Fundamentals and Biotechnology. Microbiol Mol Biol Rev [Internet]. 2002 Sep;66(3):506–77. Available from: https://journals.asm.org/doi/10.1128/MMBR.66.3.506-577.2002

97. Barcelos MCS, Ramos CL, Kuddus M, Rodriguez-Couto S, Srivastava N, Ramteke PW, et al. Enzymatic potential for the valorization of agro-industrial by-products. Biotechnol Lett [Internet]. 2020 Oct 9;42(10):1799–827. Available from: https://doi.org/10.1007/s10529-020-02957-3

98. Wilson DB. Cellulases and biofuels. Curr Opin Biotechnol [Internet]. 2009 Jun;20(3):295–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166909000676

99. Wohlgemuth R, Littlechild J, Monti D, Schnorr K, van Rossum T, Siebers B, et al. Discovering novel hydrolases from hot environments. Biotechnol Adv [Internet]. 2018 Dec;36(8):2077–100. Available from: https://doi.org/10.1016/j.biotechadv.2018.09.004

100. Escuder-Rodríguez J-J, DeCastro M-E, Cerdán M-E, Rodríguez-Belmonte E, Becerra M, González-Siso M-I. Cellulases from Thermophiles Found by Metagenomics. Microorganisms [Internet]. 2018 Jul 10;6(3):66. Available from: http://www.mdpi.com/2076-2607/6/3/66

101. Basit A, Liu J, Rahim K, Jiang W, Lou H. Thermophilic xylanases: from bench to bottle. Crit Rev Biotechnol [Internet]. 2018 Oct 3;38(7):989–1002. Available from: https://doi.org/10.1080/07388551.2018.1425662

102. Tomazini A, Higasi P, Manzine LR, Stott M, Sparling R, Levin DB, et al. A novel thermostable GH5 β-xylosidase from Thermogemmatispora sp. T81. N Biotechnol [Internet]. 2019 Nov;53(July):57–64. Available from: https://doi.org/10.1016/j.nbt.2019.07.002

103. Liu Y, Huang L, Zheng D, Xu Z, Li Y, Shao S, et al. Biochemical characterization of a novel GH43 family β-xylosidase from Bacillus pumilus. Food Chem [Internet]. 2019 Oct;295(May):653–61. Available from: https://doi.org/10.1016/j.foodchem.2019.05.163

104. Saha BC. Hemicellulose bioconversion. J Ind Microbiol Biotechnol [Internet]. 2003 May 1;30(5):279–91. Available from: https://academic.oup.com/jimb/article/30/5/279-291/5992143

105. Bosetto A, Justo PI, Zanardi B, Venzon SS, Graciano L, dos Santos EL, et al. Research Progress Concerning Fungal and Bacterial β-Xylosidases. Appl Biochem Biotechnol [Internet]. 2016 Feb 4;178(4):766–95. Available from: http://link.springer.com/10.1007/s12010-015-1908-4

106. Bajaj P, Mahajan R. Cellulase and xylanase synergism in industrial biotechnology. Appl Microbiol Biotechnol

[Internet]. 2019 Nov 18;103(21–22):8711–24. Available from: http://link.springer.com/10.1007/s00253-019-10146-0

107. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics [Internet]. 2012 Jun 1;28(11):1420–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22495754

108. Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. Rodriguez-Valera F, editor. PLoS One [Internet]. 2011 Mar 9;6(3):e17288. Available from: https://dx.plos.org/10.1371/journal.pone.0017288

109. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook [Internet]. Totowa, NJ: Humana Press; 2005. p. 571–608. Available from: http://link.springer.com/10.1385/1592598900

110. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–303.

111. SchrödingerLCC. The PyMOL Molecular Graphics System, Version 1.2r3pre [Internet]. Available from: https://pymol.org/2/#page-top

112. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D506–15. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

113. Shi H, Li X, Gu H, Zhang Y, Huang Y, Wang L, et al. Biochemical properties of a novel thermostable and highly xylose-tolerant β-xylosidase/α-arabinosidase from Thermotoga thermarum. Biotechnol Biofuels [Internet]. 2013;6(1):27. Available from: Biotechnology for Biofuels

114. Li Q, Wu T, Qi Z, Zhao L, Pei J, Tang F. Characterization of a

novel thermostable and xylose-tolerant GH 39 β-xylosidase from Dictyoglomus thermophilum. BMC Biotechnol [Internet]. 2018 Dec 21;18(1):29. Available from: https://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-018-0440-3

115. Margolles-Clark E, Tenkanen M, Nakari-Setälä T, Penttilä M. Cloning of genes encoding alpha-L-arabinofuranosidase and beta-xylosidase from Trichoderma reesei by expression in Saccharomyces cerevisiae. Appl Environ Microbiol [Internet]. 1996 Oct;62(10):3840–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8837440

116. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem [Internet]. 1976 May 7;72:248–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/942051

117. Silveira MHL, Aguiar RS, Siika-aho M, Ramos LP. Assessment of the enzymatic hydrolysis profile of cellulosic substrates based on reducing sugar release. Bioresour Technol [Internet]. 2014;151:392–6. Available from: http://dx.doi.org/10.1016/j.biortech.2013.09.135

118. Suleiman M, Schröder C, Klippel B, Schäfers C, Krüger A, Antranikian G. Extremely thermoactive archaeal endoglucanase from a shallow marine hydrothermal vent from Vulcano Island. Appl Microbiol Biotechnol [Internet]. 2019 Feb 13;103(3):1267–74. Available from: http://link.springer.com/10.1007/s00253-018-9542-z

119. Joshi N, Kaushal G, Singh SP. Biochemical characterization of a novel thermo-halo-tolerant GH5 endoglucanase from a thermal spring metagenome. Biotechnol Bioeng [Internet]. 2021 Apr 21;118(4):1531–44. Available from: https://onlinelibrary.wiley.com/doi/10.1002/bit.27668

120. Shi R, Li Z, Ye Q, Xu J, Liu Y. Heterologous expression and characterization of a novel thermo-halotolerant endoglucanase Cel5H from Dictyoglomus thermophilum. Bioresour Technol [Internet]. 2013 Aug;142:338–44.

Available from:
http://dx.doi.org/10.1016/j.biortech.2013.05.037

121. Gaur R, Tiwari S. Isolation, production, purification and characterization of an organic-solvent-thermostable alkalophilic cellulase from Bacillus vallismortis RG-07. BMC Biotechnol [Internet]. 2015 Dec 19;15(1):19. Available from: https://bmcbiotechnol.biomedcentral.com/articles/10.1186/s12896-015-0129-9

122. Wang H-J, Hsiao Y-Y, Chen Y-P, Ma T-Y, Tseng C-P. Polarity Alteration of a Calcium Site Induces a Hydrophobic Interaction Network and Enhances Cel9A Endoglucanase Thermostability. Kelly RM, editor. Appl Environ Microbiol [Internet]. 2016 Mar 15;82(6):1662–74. Available from: https://journals.asm.org/doi/10.1128/AEM.03326-15

123. Chauvaux S, Beguin P, Aubert JP, Bhat KM, Gow LA, Wood TM, et al. Calcium-binding affinity and calcium-enhanced activity of Clostridium thermocellum endoglucanase D. Biochem J [Internet]. 1990 Jan 1;265(1):261–5. Available from: https://portlandpress.com/biochemj/article/265/1/261/24767/Calciumbinding-affinity-and-calciumenhanced

ANNEX II

Curriculum Vitae

DATOS ACADÉMICOS

Licenciado en Biotecnología por la Universidad de León (2006/2011).

Máster Oficial Interuniversitario en Biotecnología Avanzada por la Universidade da Coruña y la Universidade de Vigo (2011/2013).

Idioma inglés de nivel avanzado (B2) certificado por la Escuela Oficial de Idiomas (curso 2014/2015).

EXPERIENCIA PROFESIONAL E INVESTIGADORA

Prácticas profesionales en el Laboratorio Municipal de Ponferrada (55 días, 7h/día, 2010). Laboratorio de análisis de aguas de consumo.

Trabajo de fin de Máster en el Laboratorio de Bioquímica y Biología Molecular de la Facultad de Ciencias de la Universidad de A Coruña ("Caracterización bioquímica de la enzima α-galactosidasa de *Saccharomyces cerevisiae* y de su mutante R412L", 6 meses durante el curso 2012/2013).

Estancia en el Departamento de Ciencias de la Computación de la San Diego State University (California, EEUU). 3 meses durante el curso 2017/2018.

MERITOS CIENTÍFICOS

CAPÍTULOS DE LIBRO

Escuder-Rodríguez J-J, López-López O, Becerra M, Cerdán M-E, González-Siso M-I. Extremophilic esterases for bioprocessing of lignocellulosic feedstocks. Extremophilic Enzymatic Processing of Lignocellulosic Feedstocks to Bioenergy. 2017.

Escuder-Rodríguez J-J, DeCastro M-E, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Advances of Functional Metagenomics in Harnessing Thermozymes. Metagenomics: Perspectives, Methods, and Applications. 2017.

DeCastro M-E, Escuder-Rodríguez J-J, Becerra M, Rodríguez-Belmonte E, González-Siso M-I. Archaeal Biocommunication

in Hot Springs Revealed by Metagenomics. In: Witzany G, editor. Biocommunication of Archaea [Internet]. Cham: Springer International Publishing; 2017. p. 85–101. Available from: https://doi.org/10.1007/978-3-319-65536-9_6

PUBLICACIONES EN REVISTAS CIENTÍFICAS

Rico-Díaz A, Álvarez-Cao M-E, Escuder-Rodríguez J-J, González-Siso M-I, Cerdán ME, Becerra M. Rational mutagenesis by engineering disulphide bonds improves *Kluyveromyces lactis* beta-galactosidase for high-temperature industrial applications. Sci Rep. 2017;7.

Gomri M, Rico-Díaz A, Escuder-Rodríguez J-J, El Moulouk Khaldi T, González-Siso M-I, Kharroub K. Production and Characterization of an Extracellular Acid Protease from Thermophilic *Brevibacillus* sp. OA30 Isolated from an Algerian Hot Spring. Microorganisms [Internet]. 2018 Apr 12;6(2):31. Available from: http://www.mdpi.com/2076-2607/6/2/31

Escuder-Rodríguez J-J, DeCastro M-E, Cerdán M-E, Rodríguez-Belmonte E, Becerra M, González-Siso M-I. Cellulases from Thermophiles Found by Metagenomics. Microorganisms [Internet]. 2018 Jul 10;6(3):66. Available from: http://www.mdpi.com/2076-2607/6/3/66

DeCastro M-E, Escuder-Rodriguez J-J, Cerdan M-E, Becerra M, Rodriguez-Belmonte E, Gonzalez-Siso M-I. Heat-Loving β-Galactosidases from Cultured and Uncultured Microorganisms. Curr Protein Pept Sci. 2018;19(12):1224–34.

Rai DK, Gurusaran M, Urban V, Aran K, Ma L, Li P, et al. Structural determination of Enzyme-Graphene Nanocomposite Sensor Material. Sci Rep. 2019;9(1).

Adjeroud M, Escuder-Rodríguez J-J, González-Siso M-I, Kecha M. Metagenomic Investigation of Bacterial and Archaeal Diversity of Hammam Essalihine Hot Spring from Khenchela, Algeria. Geomicrobiol J [Internet]. 2020;37(9):804–17. Available from: https://doi.org/10.1080/01490451.2020.1783035

Medjemadj M, Escuder-Rodríguez J-J, Boudemagh A, González-Siso M-I. Actinobacteria isolated from Algerian hot spring waters: A potential source of important enzymes. Eco Env Cons. 2020;26(3):1145–57.

PARTICIPACIÓN EN CONGRESOS CIENTÍFICOS

Póster en el XXXVI Congreso de la SEBBM (Madrid), septiembre 2013. "Caracterización bioquímica de un mutante termoestable de la α-galactosidasa de *Saccharomyces cerevisiae*".

Póster en el XXXVII Congreso de la SEBBM (Granada), septiembre 2014. "Búsqueda de termoenzimas hidrolíticas en una metagenoteca de aguas termales".

Póster en el XXXVIII Congreso de la SEBBM (Valencia), septiembre 2015. "Caracterización de una comunidad microbiana procedente de la fuente termal de Muíño da Veiga (Ourense) mediante metagenómica basada en secuencia".

Póster en el XL Congreso de la SEBBM (Barcelona), octubre 2017. "Comparison of two metagenomic strategies for the identification of hydrolases from a hot spring."

Póster en el 41 Congreso de la SEBBM (Santander), septiembre 2018. "Identificación y caracterización bioquímica de hidrolasas de fuentes termales".

Póster en el Congreso Nacional de Biotecnología BIOTEC2019 (Vigo), junio 2019. "Characterization of a novel thermostable lypolitic hydrolase found by functional metagenomics from a hot spring".

Póster en el 43 Congreso de la SEBBM (Barcelona), julio 2021. "Expresión y caracterización bioquímica de una celulasa obtenida mediante metagenómica basada en secuencia de aguas termales."

ASISTENCIA A CURSOS Y CONGRESOS

Asistencia al III Congreso Interuniversitario de Biotecnología, Universidad de León, julio 2008.

Asistencia al I Congreso de Biotecnología y Empresa, Universidad de Salamanca, marzo 2009.

Curso teórico, “Recursos y utilidades documentales para la Tesis y otros trabajos académicos”, Universidad de A Coruña (Servicio de Biblioteca Universitaria, Sección de Información y Coordinación), marzo 2014, 10 horas.

Curso teórico-práctico, “Biotechnology and genetics of thermophiles: from lectures to lab”, Universidad Autónoma de Madrid, mayo 2014.

Curso del Plan de Apoyo a Nuevos Investigadores 2014/2015, “Análisis de datos con SPSS: Nivel Inicial”, Universidad de A Coruña (Vicerreitoría de Títulos, Calidad y Nuevas Tecnologías y Centro Universitario de Formación e Innovación Educativa), diciembre 2014, 15 horas.

Curso teórico-práctico, “Microfluidic fundamentals and applications; Innovation and Entrepreneurship for Scientists”, HOTDROPS Workshop Cambridge UK, 23-25 de marzo 2015.

Curso del Plan de Apoyo a Nuevos Investigadores 2015/2016, “Metodología de la investigación básica y aplicada”, Universidad de A Coruña (Vicerreitoría de Oferta Académica e Innovación Docente y Centro Universitario de Formación e Innovación Educativa), febrero 2016, 12 horas.

Curso del Centro de Supercomputación de Galicia (CESGA): Taller para usuarios: Acceso a FinisTerrae-2, junio 2016, 4 horas.

Asistencia al I Congreso Iberoamericano ACEM-AEBE: Biotecnología y Nanotecnología (organizado por internet). Diciembre 2020.

ASISTENCIA A JORNADAS Y CONFERENCIAS

Asistencia a "Encuentro Agrobiotech Innovación" 26 y 27 de noviembre de 2015 Parque Tecnolóxico de Galicia (Tecnópole), Ourense.

Asistencia a la conferencia “Bioingeniería de Beta-galactosidasas” (agrupación CICA-INIBIC) de 1,5 horas de duración, 22 junio de 2017, A Coruña.

Asistencia a “I Xornadas de colaboración entre a Agrupación Estratéxica CICA-INIBIC e o Centro Singular de Investigación CITIC” 12 de abril de 2018, A Coruña.

Ponente con conferencia en el “International Minisymposium: Metagenomics of Thermophiles and Thermozymes” – “Hydrolitic enzymes found by metagenomics” 12 de mayo de 2017, A Coruña.

Ponente con conferencia en la “XIV Reunión de la Red Nacional de Microorganismos Extremófilos (RedEX)” – “Búsqueda de hidrolasas termófilas mediante metagenómica funcional y basada en secuencia” 3 y 4 de noviembre de 2017, Ourense.

Ponente con conferencia en la “Happy Hour CICA-INIBIC” (agrupación CICA-INIBIC) – “Bioprospección de fuentes hidrotermales mediante metagenómica” 25 de enero de 2019, A Coruña.

Asistencia a la conferencia “Una endoxilanasa de un hongo aislado en la Antártica: estudios bioquímicos y aplicación en panificación” (agrupación CICA-INIBIC) de 1 hora de duración, 14 de mayo de 2019, A Coruña.

Ponente con conferencia (presentación oral) en la I Jornada de Biología Celular y Molecular, Facultad de Ciencas, UDC. Septiembre 2019.

ACTIVIDADES DE DIVULGACIÓN CIENTÍFICA

Participación en el evento de divulgación ”World Biotech Tour 2017 – Festival Biotecnológico” (DOMUS-CICA) 30 de septiembre de 2017 A Coruña (8 horas de duración).

Ponente con conferencia en la Jornada Divulgativa “Stranger Science” (CRE, Sociedad de Científicos Retornados a España), 23 de marzo de 2018 A Coruña.

Colaborador en las actividades de divulgación Univerisidade da Coruña en la XXIII edición del “Día da Ciencia na Rúa” (Asociación de Amigos da Casa das Ciencias) 5 de mayo de 2018, A Coruña.

Colaborador en la Feria Científica UDCiencia (Fundación Barrié y grupo UDCiencia d+i de la Universidade da Coruña) del 25 al 29 de marzo de 2019, A Coruña.

Colaborador en las actividades de divulgación Univerisidade da Coruña en la XXIV edición del “Día da Ciencia na Rúa” (Asociación de Amigos da Casa das Ciencias), 4 de mayo de 2019, A Coruña.

Colaborador en los talleres D'tec UDCiencia (Agencia Gallega de Innovación, GAIN), Centro de Innovación y Servicios de la Tecnología y el Diseño en Ferrol. Noviembre 2019, Ferrol.

DOCENCIA

Conferencia impartida “Recursos biotecnológicos en fuentes geotermales gallegas” como parte del encuentro de la Universidad Internacional Menéndez Pelayo “Organismos termófilos: fuente de recursos para la biotecnología del siglo XXI” - junio de 2018, A Coruña.

Presentación de póster y asistencia a las IV Jornadas de Innovación Docente (Facultad de Ciencias de la Educación, UDC), CUFIE. Noviembre 2019.

PROYECTOS DE INVESTIGACIÓN

Miembro del equipo de investigación de los siguientes proyectos:

Ayuda para la consolidación y la estructuración de unidades de investigación competitivas del sistema universitario de Galicia (modalidad grupos de referencia competitiva). Código: CN 2012/118. Investigadora principal: María Esperanza Cerdán Villanueva. Entidad financiadora: Xunta de Galicia. Concellería de Educación e Ordenación Universitaria. Duración del proyecto: 2012-2015.

Ayuda para la consolidación y la estructuración de unidades de investigación competitivas del sistema universitario de Galicia (modalidad grupos de referencia competitiva). Investigadora

principal: María Esperanza Cerdán Villanueva. Entidad financiadora: Xunta de Galicia. Concellería de Educación e Ordenación Universitaria. Duración del proyecto: 2017-2019.

Ayuda para la consolidación y la estructuración de unidades de investigación competitivas del sistema universitario de Galicia (modalidad grupos de referencia competitiva). Investigadora principal: María Esperanza Cerdán Villanueva. Entidad financiadora: Xunta de Galicia. Concellería de Educación e Ordenación Universitaria. Duración del proyecto: 2020-2023.

Ayuda para la creación, reconocimiento y estructuración de agrupaciones estratégicas del sistema universitario de Galicia (agrupación CICA-INIBIC). Código: AGRUP2015/05. Investigador principal: Jaime Rodríguez González. Entidad financiadora: Xunta de Galicia. Duración del proyecto: 01/01/2015 - 30/11/2017.

Ayuda para la creación, reconocimiento y estructuración de agrupaciones estratégicas del sistema universitario de Galicia (agrupación CICA-INIBIC). Código: ED431E 2018/03. Investigador principal: Jaime Rodríguez González. Entidad financiadora: Xunta de Galicia. Duración del proyecto: 2018-2020.

Ultrahigh-throughput platform for the screening of thermostable proteins by thermophilic in vitro transcription-translation and microflucidics (HotDrops). Investigador principal: Mª Isabel González Siso. Entidad financiadora: Unión Europea. Marie-Curie Industry-Academia Partnerships and Pathways (IAPP). Call: FP7-PEOPLE-2012-IAPP. Duración del proyecto: 01/06/2013 - 01/06/2017.