UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Bioquímica y Ciencias Biológicas

Tesis para la obtención del Grado Académico de Doctor en Ciencias Biológicas

Caracterización del potencial biotecnológico de lagunas de estabilización de PYMES lácteas usando metagenómica.

José Matías Irazoqui

Director de Tesis: Ariel Amadio

Lugar de realización: Instituto de Investigación de la Cadena Láctea (INTA-CONICET)

-2022-

Agradecimientos

Este trabajo no podría haber sido posible sin la ayuda de una enorme cantidad de gente.

En primer lugar, quiero agradecerle a CONICET, INTA Rafaela y UNL, por darme la oportunidad de hacer este doctorado.

También quiero que agradecerle a Ariel por la posibilidad de hacer el doctorado con él y por todas esos momentos de charlas, consejos, y discusiones, pero sobre todo un montón de mucha buena onda.

A Flor, por haberme bancado todos estos años y haberme tenido toda la paciencia del mundo, no solo como aprendiz de laboratorio, sino también como compañero de oficina.

A toda la gente del edificio central del INTA Rafaela y del laboratorio de Sanidad Animal, por todos esos almuerzos, charlas y momentos lindos durante todos estos años.

A mi familia, que a pesar de estar a muchos muchos kilómetros, siempre me apoyaron en todo este recorrido desde el principio en absolutamente todo, desde estudiar algo llamado “Bioinformática” a 1000 km de cada, hasta hacer un doctorado en otra ciudad igual de lejos y en un tema igual de raro.

Y a todos los amigos y personas que fui conociendo en todos estos años, ya sea en laboratorios, cursos, congresos, charlas, viajes y demases, que de alguna u otra forma aportaron su granito de arena para que este trabajo pudiera ser llevado a cabo.

Muchas gracias!!!

Los resultados obtenidos en esta tesis han dado lugar a las siguientes publicaciones:

- Eberhardt, M. F., Irazoqui, J. M., & Amadio, A. F. (2020). β-galactosidases from a sequence-based metagenome: cloning, expression, purification and characterization. *Microorganisms*, *9*(1), 55.
- Irazoqui, J. M., Eberhardt, M. F., Adjad, M. M., & Amadio, A. F. (2022). Identification of key microorganisms in facultative stabilization ponds from dairy industries, using metagenomics. *PeerJ*, *10*, e12772.

INDICE

1 ABREVIATURAS

ADN: Ácido desoxirribonucleico.

EDTA: Ácido etilendiaminotetraacético.

HMM: Modelos ocultos de Markov (*Hidden Markov Models*)

HPLC: Cromatografía líquida de alta eficacia.

MAG: Genoma ensamblado desde un metagenoma.

p/v: peso en volumen.

PCR: Reacción en cadena de la polimerasa.

PyMES: Pequeñas y medianas empresas.

SDS: Dodecil sulfato sódico.

SDS-PAGE: Electroforesis en gel de poliacrilamida con dodecilsulfato sódico.

TCA: ácido tricloroacético

Tris: tris(hidroximetil)aminometano.

UV: Radiación ultravioleta.

v/v: volumen en volumen.

2 RESUMEN

El lactosuero es el principal subproducto de la industria láctea. Los enormes volúmenes producidos por año y su alta demanda química y biológica de oxígeno del mismo, debido a su alta carga orgánica, lo convierten en un potencial problema ambiental. Debido a su composición, principalmente lactosa y proteínas, el suero puede considerarse como materia prima para la elaboración de productos de mayor valor, mediante el uso de enzimas. Las enzimas llamadas β-galactosidasas son capaces de transformar la lactosa en galactooligosacáridos, que poseen actividad prebiótica. Por su parte, las proteasas son enzimas capaces de hidrolizar las proteínas del suero y liberar péptidos, que presentan una gran variedad de actividades, como antimicrobiana y antihipertensiva, entre otras.

Las lagunas de estabilización facultativas son la principal tecnología usada por pequeñas industrias lácteas para el tratamiento de sus efluentes, entre los cuales se pueden encontrar restos de suero. En estas lagunas, la comunidad microbiana es la encargada de estabilizar la carga orgánica de los efluentes, combinando procesos aeróbicos y anaeróbicos. Dado que la principal fuente de materia orgánica volcada en las lagunas de estabilización de industrias lácteas son diluciones de leche, entre las que se puede encontrar lactosuero, es esperable que miembros de su comunidad microbiana presenten β-galactosidasas y proteasas.

En el presente trabajo se plantea el estudio de las comunidades microbianas de dos sistemas de lagunas facultativas de dos pequeñas industrias lácteas del centro de la provincia de Santa Fe. Utilizando metagenómica basada en secuenciación y un enfoque centrado en genomas, se reconstruyeron 110 genomas microbianos completos, los cuales fueron caracterizados taxonómica y funcionalmente. Entre ellos se identificaron 7 grupos taxonómicos que serían claves para los sistemas facultativos, ya que fueron encontrados en ambos sistemas y presentan rutas metabólicas de interés que no fueron identificadas en otros organismos.

En la base de datos específica CAZy existen 5 familias descritas con actividad β-galactosidasa (EC: 3.2.1.23). Comparando los genes predichos en el metagenoma contra esta base de datos, se identificaron 379 β-galactosidasas candidatas, en 3 de las familias de interés (GH1, GH2, y GH42). Estos genes fueron clasificados tanto taxonómicamente como en familia y se seleccionaron 18 candidatos para su clonado y expresión. De ellos, 5 pudieron ser expresados y mostraron actividad con orto-nitrofenil-β-galactósido y lactosa. Entre estas 5 enzimas, βgal5 mostró una actividad superior a la del resto de las enzimas expresadas y a la de otras enzimas similares de origen metagenómico.

Por su parte, la base de datos MEROPS describe 4 familias de proteasas que han sido reportadas que poseen actividad sobre las proteínas del suero. Dentro de los metagenomas se identificaron 851 proteasas candidatas, en 3 de las familias de interés (C01, S01 y S08). Estos genes fueron clasificados tanto taxonómicamente como en familia y se seleccionaron 10 candidatos para su clonado

y expresión. De ellos, 1 pudo ser expresado y la proteína resultante mostró actividad proteasa sobre azocaseína y proteínas de leche.

El estudio de los miembros de una comunidad microbiana permite conocer con mayor profundidad el funcionamiento de los procesos metabólicos dentro de un ambiente y la identificación de microorganismos de importancia para ellos. Por otra parte, las enzimas de origen metagenómico poseen el potencial de modificar subproductos de distintas industrias, como la láctea, y obtener productos de valor agregado. Este trabajo describe la combinación de estas estrategias para contribuir al mejor entendimiento del funcionamiento de las lagunas de estabilización, así como a la identificación y producción de enzimas para el aprovechamiento de subproductos y su transformación para agregar valor.

3 ABSTRACT

Whey is the main by-product from dairy industries. The large volumes produced every year and its high chemical and biological oxygen demands, make it an important environmental problem. Due to its composition, mainly lactose and proteins, whey can be considered as a raw material for added-value products, using enzymes. β-galactosidases are enzymes capable of transforming lactose into galactooligosaccharides, compounds with prebiotic activity. On the other hand, proteases are enzymes able to hydrolyse whey proteins to obtain peptides, which have been reported to have a wide array of activities, such as antihypertensive and antimicrobial, among others.

Facultative stabilization ponds are the main technology used by small dairy industries to treat their effluents, which include small fractions of whey. In these ponds, the microbial community stabilize the organic matter present in effluents, combining both aerobic and anaerobic processes. Since the main source of organic matter discharged into these ponds are different dilutions of milk, that may contain traces of whey, the members of the microbial community would harbour both β-galactosidases and proteases.

In this work, two stabilization ponds systems, from two small dairy industries from Santa Fe province, were studied. Using a sequence-based and genome-centered metagenomics approach, 110 novel genomes were reconstructed. These were taxonomically and functionally characterized. We identified 7 key taxonomic groups that would be key for facultative ponds, since they were found in both systems and they presented important metabolic pathways that were not found in other members of the community.

There are 5 families with β-galactosidase activity reported in the CAZy database (EC: 3.2.1.23). Comparing the predicted genes in the metagenomes with the database, 379 gene candidates were found, in 3 of these 5 families (GH1, GH2, and GH42). These genes were classified both taxonomically and family-wise and 18 of them were selected for cloning and expression. A total of 5 enzymes were successfully expressed and showed activity with ortho-nitrophenyl-β-galactoside and lactose. One of these enzymes, βgal5, showed higher activity levels than the other 4, and higher than other enzymes obtained from metagenomes.

As for proteases, there are 4 families described in the MEROPS database than have been reported to be active with whey protein. When compared to MEROPS, a total of 851 putative proteases, in 3 families (C01, S01, and S08) were identified in the metagenomes. As with the β-galactosidases, these genes were taxonomically and family-wise classified, and 10 genes were selected for cloning and expression. One was successfully expressed and the recombinant enzyme was active towards azocasein and milk proteins.

The study of the members of a microbial community brings a better understanding of the metabolic processes occurring in an environment, and the identification of key microorganisms for these processes. On the other hand, novel enzymes from metagenomes could transform by-products from industries, such as dairy industries, to produce added-value products. This work describes the combination of strategies to contribute to a better understanding of facultative stabilization ponds, as for the identification and production of enzymes for the transformation of by-products into added-value products.

INTRODUCCIÓN

4 INTRODUCCIÓN

LACTOSUERO

El lactosuero es el principal subproducto de la industria láctea. En la producción de queso, la primera etapa consiste en añadir a la leche un cuajo, que es una mezcla de enzimas, entre las que se destaca la quimosina (Ryan y Walsh, 2016). El suero de queso o **lactosuero** es el líquido remanente después de cuajar y remover la caseína de la leche (figura 4.1). El mismo representa entre el 85% y el 95% del volumen total de la leche utilizada y retiene alrededor del 55% de los nutrientes de la leche (González Siso, 1996).

Figura Figura 4.1: Obtención de lactosuero durante la producción de queso. Credito de la imagen: https://www.acs.org

La tabla 4.1 muestra la composición por litro de lactosuero (Juliano y col., 2017). El agua es el principal componente (~93% del volumen), mientras que entre los sólidos orgánicos, lo que más se destacan son la lactosa (4,5-5% peso/volumen) y distintas proteínas, que representan entre el 0,6% y 0,8% de peso en volumen.

Tabla 4.1. Principales componentes del lactosuero

Componente	Cantidad (g/L)
Agua	935
Lactosa	47
Proteinas	3,7
Lipidos	1,85
Minerales	4,2
Vitaminas	0,0075
Nitrogeno no proteico	0,132
Elementos traza	0,0028

Según la Organización para la Cooperación y el Desarrollo Económicos (OECD, por sus siglas en inglés), la producción anual de leche entre 2017 y 2019 rondó los 839 Mt, mientras que la producción de queso rondó los 23,5 Mt (OECD y col., 2020). De acuerdo a Božanić y col. (2014), se puede estimar que para producir un 1 Kg de queso duro o semiduro son necesarios 10 L de leche, y como resultado se obtienen 9 L de lactosuero. De aquí es que se estima que la cantidad de lactosuero producido en el mundo supera ampliamente los 100 Mt anuales (Baldasso y col., 2011). Según datos de la OECD-FAO, en 2019, en el país se produjeron 429 kt de queso, siendo el segundo en la región detrás de Brasil (790 kt).

Como indican Ryan y Walsh (2016), el mayor problema asociado al lactosuero es su potencial impacto ambiental. Los efluentes de queserías suelen tener una alta proporción de suero (Carvalho y col., 2013), esto produce que la demanda biológica de oxígeno (DBO) puede variar entre los 40 y 60 g/L, mientras que la demanda química de oxígeno (DQO) se encuentra entre los 50 y 80 g/L (Chatzipaschali y Stamatis, 2012). Estos valores son entre 100 y 175 veces mayores que los de aguas residuales domésticas (Mockaitis y col., 2006; Smithers, 2008). Los enormes volúmenes producidos anualmente en todo el mundo y este gran potencial impacto ambiental hacen que el desarrollo y aplicación de tecnologías para su tratamiento sean de enorme interés.

En el pasado, muchas queserías arrojaban sus desechos y efluentes, entre los que se encontraba el lactosuero, directamente a la tierra o a cuerpos de agua (ríos, lagos, etc; Prazeres y col., 2012). Otras soluciones contemplaban disponerlos en sistemas de residuos municipales, o utilizarlos en la alimentación de animales (Schingoethe, 1976), aunque esta última práctica ha caído en desuso (Malaspina y col., 1996). Ninguna de estas alternativas ha logrado realmente resolver los problemas ambientales, y el manejo de efluentes de las pequeñas queserías es un desafío cada vez más

importante, debido a las cada vez más rigurosas restricciones legales (Mawson, 1994; Farizoglu y col., 2007).

COMPONENTES DEL LACTOSUERO

Dada su alta carga orgánica, el lactosuero es una excelente fuente de **proteínas** y **carbohidratos** para ser explotada por industrias como la alimenticia, biotecnológica, o médica, entre otras. En los últimos años muchos trabajos se han enfocado en el aprovechamiento de este subproducto (Smithers, 2008).

Los procesos más simples se enfocan en la utilización del suero para producir quesos o bebidas. En países como Suecia o los Países Bajos existen bebidas que enriquecen jugos de frutas con suero (Kosikowski 1968; Holsinger y col. 1974; Jelen, 2009), aunque estas bebidas no poseen mucho éxito fuera de sus regiones. Otra opción es la producción de bebidas alcohólicas, como cervezas, vinos y champagnes fermentando el lactosuero con levaduras, como *Kluyveromyces fragilis* or *Saccharomyces lactis*, con aditivos como sacarosa o malta (Sienkiewicz y Riedel 1990; Jeličić y col. 2008; Holsinger y col. 1974). El suero también puede usarse para hacer quesos, como la ricota o el mysost, calentándolo a altas temperaturas y usando aditivos, como ácido cítrico, para cuajar las proteínas que después se usarán para producir el queso (Jelen and Buchheim, 1976; Pintado y col., 2001). Sin embargo, todos estos productos tienen poco atractivo comercial y no representan una forma de tratar grandes cantidades de suero producido (Ryan y Walsh, 2016).

Otra alternativa es someter al suero a tratamientos físico químicos, que producen productos de mayor valor pero requieren equipamiento especial (figura 4.2). Mediante secado spray puede obtenerse suero en polvo (Kosikowski 1979; Yang y Silva 1995), que puede ser preservado por más tiempo y usado en alimentación tanto animal como humana (Gonzalez Siso, 1996). Si en lugar de secado, se utilizan membranas para ultrafiltrar el lactosuero pueden obtenerse un concentrado de proteínas de suero (WPC, por sus siglas en inglés). En este proceso también produce como resultado permeado de suero (Mollea y col. 2013). El permeado representa alrededor del 90% del volumen del suero y la lactosa representa alrededor del 85% de su parte sólida (Khorshid, 1974; Fenton-May et al., 1971). Mediante ósmosis inversa puede tratarse este subproducto para purificar y concentrar la lactosa (Prazeres y col., 2012). Si bien el WPC y, en menor medida, la lactosa son productos con valor agregado, es posible seguir tratando estos componentes del suero para producir otros productos de mayor valor.

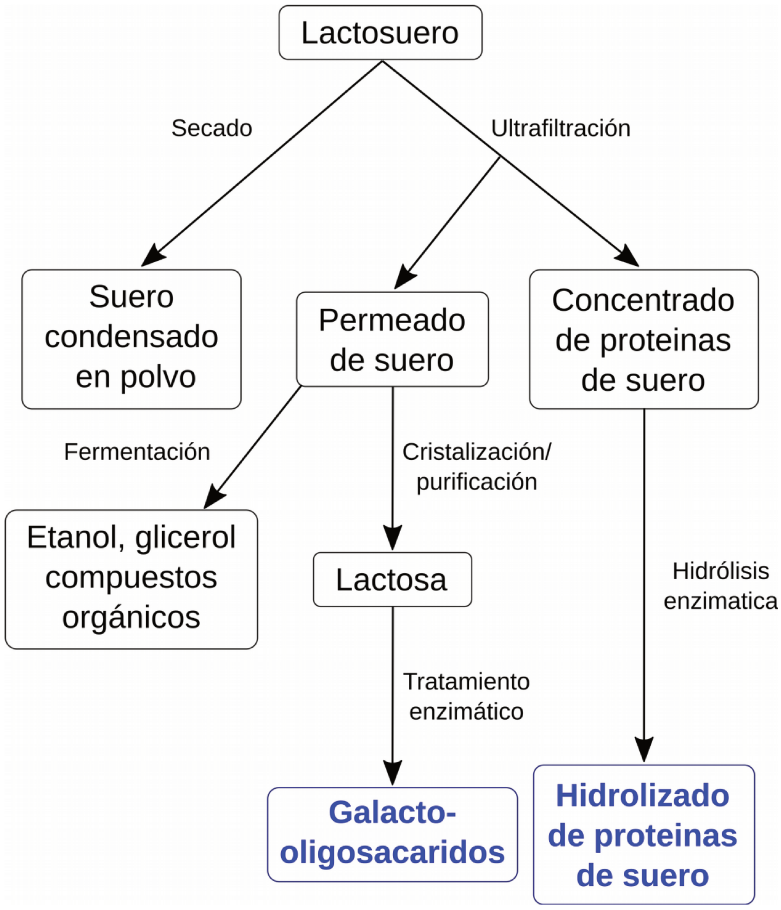

Figura 4.2: Procesos para obtener subproductos del lactosuero (adaptado de Božanić y col., 2014)

LACTOSA

La lactosa es un disacárido formado por los monosacáridos glucosa y galactosa, unidos por un enlace glicosídico β 1-4 (Paques y Lindner, 2019). La galactosa es un azúcar con la misma fórmula química que la glucosa, C6H12O6, y sus estructuras difieren solo en la posición de un grupo hidroxilo. Sin embargo, esta diferencia le confiere a la galactosa propiedades químicas y bioquímicas diferentes a las de la glucosa.

La lactosa también es conocida como el “azúcar de leche”, ya que solo se encuentra significativamente presente en la leche, y su única fuente natural conocida hasta el momento son las glándulas mamarias (Urashima y col., 1988). Su concentración puede variar entre el 0% y el 10%, peso en volumen (Fox, 2013). La concentración de lactosa en leche bovina ronda el 4,8%, aunque varía ligeramente entre las distintas etapas de lactancia.

Este compuesto es hidrolizado en el intestino delgado de los mamíferos por una enzima comúnmente denominada *lactasa*, que rompe el enlace glicosídico que une a los 2 monosacáridos que forman la lactosa (figura 4.3). Las enzimas capaces de romper este enlace (β 1-4) son denominadas **β-galactosidasas**. Un aspecto único en humanos respecto a esta enzima es la división evolutiva en dos fenotipos en la adultez. Alrededor de un tercio de la población mantiene la capacidad de digerir lactosa mediante la lactasa intestinal, mientras que el resto de la población la pierde (Catanzaro y col., 2021).

Esta división ocurre por mutaciones en la región del promotor del gen que codifica para la lactasa (LCT), que puede regular negativamente su expresión. De aquí que se obtienen 2 genotipos, uno dominante, capaz de expresar la lactasa en cantidades suficientes para hidrolizar grandes ingestas de lactosa, y otro recesivo y *wild-type*, que resulta en niveles de expresión bajos de la enzima y, por lo tanto, baja (o nula) tolerancia a la ingesta de lactosa (Anguita-Ruiz y col., 2020). Dado que la intolerancia a la lactosa afecta a un gran número de personas, muchos trabajos se han enfocado en reducir o eliminar su contenido de productos comerciales, principalmente mediante el uso de β-galactosidasas (Flickinger, 2010).

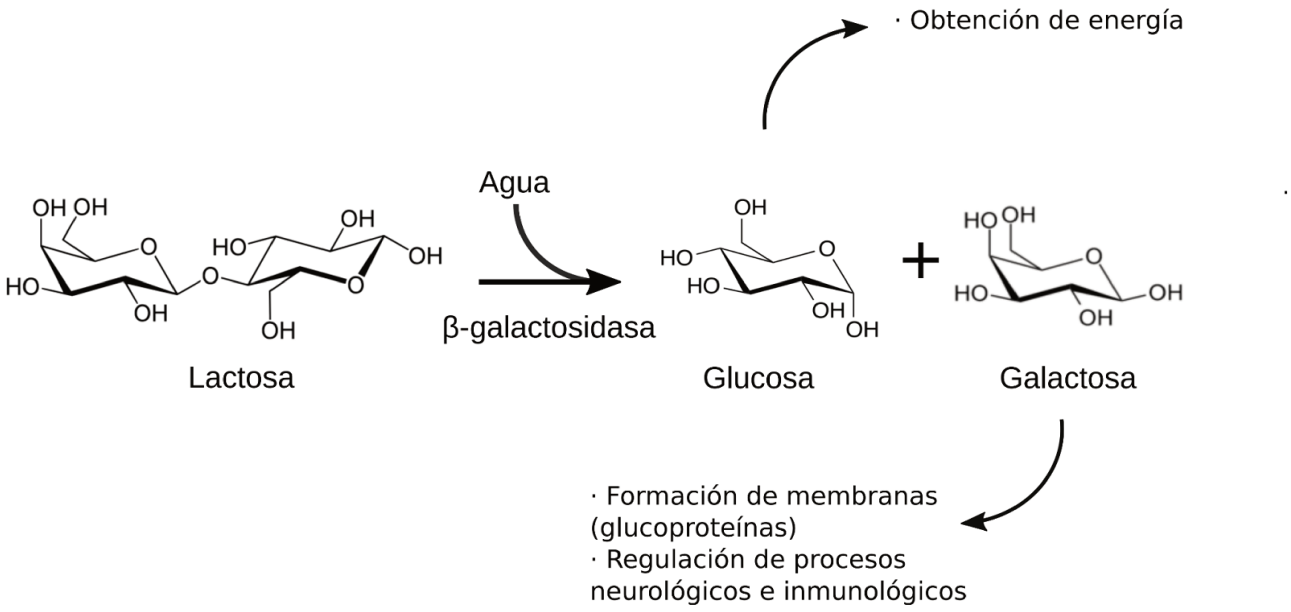

Figura 4.3: Degradación de la lactosa y uso de los monómeros.

GALACTOOLIGOSACÁRIDOS

Algunos miembros de las familias de β-galactosidasas también pueden usar la lactosa como precursor de otros compuestos más complejos, llamados **galactooligosacáridos** (GOS, Boyer y col., 1963), mediante una reacción de transgalactosilación.

Los GOS están formados por una glucosa terminal y un número variable de galactosas, normalmente, entre 2 y 7 (Figura 4.4), unidos por enlaces β-glicosídicos (Torres y col., 2010; Gänzle, 2012), usualmente β-(1-4) y β-(1-6), y en menor medida, β-(1-3) (Coulier y col., 2009). Si bien es posible obtener GOS mediante síntesis química, por la acción de ácidos minerales (Huh y col., 1991), no es posible controlar la especificidad de los GOS obtenidos. Por ello, la forma más utilizada es la síntesis enzimática (de Roode y col., 2003).

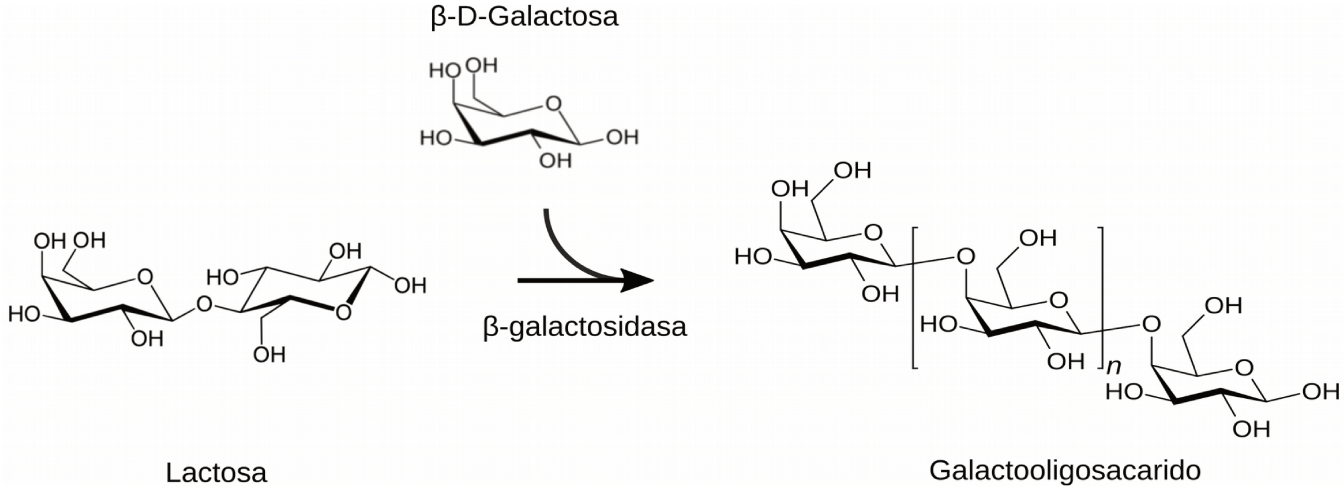

Figura 4.4: Síntesis de galactooligosacáridos

Al igual que otros oligosacáridos, los GOS tienen un sabor agradable, y pueden mejorar la textura y la sensación que producen los alimentos en boca. La α-amilasa presente en la saliva, las enzimas digestivas y la microbiota oral no pueden degradarlos, por lo tanto llegan sin sufrir modificaciones al colon (Van Loo y col., 1999; Tzortzis y Vulevic, 2009). Estos compuestos pueden actuar como prebióticos, favoreciendo la proliferación de microorganismos presentes en el intestino humano que confieran beneficios al bienestar y la salud de una persona, como los miembros de los grupos *Bifidobacterium* y *Lactobacillus* (Gibson y Roberfroid, 1995; Gibson y col., 2004; Macfarlane y col., 2006).

El mercado global de GOS ascendió a los 483,93 millones de dólares en el 2019, y se espera que crezca al 10,15% anual entre 2021 y 2028 (Research Nester, 2021). El principal uso de los GOS es en fórmula infantil y alimentos para niños (Playne y Crittenden, 2009). La fórmula infantil usualmente es enriquecida con entre 6,0 y 7,2 g/L de GOS, junto con 0,6-0,8 g/L de otro oligosacárido denominado fructo-oligosacáridos (FOS; Rastall, 2006; Playne y Crittenden, 2009). Estos derivados de lactosa intentan emular a los oligosacáridos que naturalmente sólo se encuentran en leche humana (Playne y Crittenden, 2009)

PROTEÍNAS DE SUERO

El suero es considerado una fuente de proteínas de gran calidad, ya que contiene altos niveles de todos los aminoácidos esenciales (Ismail y Gu, 2010). Las proteínas presentes en el suero se detallan en la Tabla 4.2. Entre ellas se destacan la β-lactoglobulina (β-Lg), la α-lactoalbúmina (α-La), la seroalbúmina bovina (SAB) y las inmunoglobulinas (Ig, Walstra y col., 2005). También puede encontrarse una buena proporción de péptidos resultantes del clivaje de la κ-caseína, durante la formación del cuajo de queso.

Tabla 4.2: Proteínas de suero. Adaptada de Ryan y Walsh, 2016.

Proteina	% del suero
β-lactoglobulina	50–55
α-lactoalbúmina	20–25
Inmunoglobulinas	10
Seroalbúmina bovina	5–10
Proteose peptone 3	12
Lactoferrina	~1–2
Lactoperoxidasa	~0.5

Desde un punto de vista nutricional, las proteínas del suero son superiores a otras proteínas animales, como la caseína, o las proteínas del huevo (Božanić y col., 2014), ya que son ricas en aminoácidos esenciales ramificados, como la leucina, isoleucina y valina (Devries y Phillips, 2015). Estos aminoácidos cumplen un rol crucial en distintos metabolismos, el control de glucosa en sangre y el funcionamiento neuronal. También se ha demostrado que las proteínas de suero son mejores para la supresión del crecimiento de tumores (Parodi, 2007), debido a componentes como la α-La, β-Lg y la SAB.

PÉPTIDOS

Las proteínas del suero son de gran importancia dietaria, ya que representan una fuente de aminoácidos esenciales, pero también pueden llevar a cabo otras funciones biológicas en la forma de **péptidos** activos. (Deeth y Bansal, 2018). Los péptidos son fragmentos específicos de una proteína, con un tamaño que varía entre los 2 y 20 aminoácidos, y que tienen un impacto positivo en el funcionamiento y condiciones del cuerpo (Kitts y Weiler, 2003).

La forma más usada para producir péptidos es la hidrólisis enzimática, mediante **proteasas**. La mayoría de los péptidos conocidos se han obtenido usando enzimas digestivas, como la pepsina y la tripsina, aunque también se han utilizado, solas o en forma combinada, otras enzimas digestivas, como la alcalasa, quimotripsina y la papina, y enzimas de origen bacteriano y fúngico (Korhonen y Pihlanto, 2003)

Los beneficios para la salud que brindan los péptidos los han convertido en un producto de interés comercial. Los hidrolizados de proteínas de suero (WPH, por sus siglas en inglés) son considerados ingredientes en la formulación de sustitutos para la leche materna, dado sus alto valor nutricional, baja amargura y baja antigenicidad (Brandelli y col., 2015). También se han lanzado al mercado productos con péptidos con actividad antihipertensiva y anticariogénica (Korhonen, 2009). Asimismo, nuevas investigaciones se están llevando a cabo para buscar péptidos que ayuden a combatir desórdenes

psicológicos, cognitivos y de comportamiento (Nongonierma y FitzGerald, 2015). Se han reportado numerosos efectos benéficos para el organismo, como ser antimicrobianos, antioxidantes, o antihipertensivos, entre otros (Brandelli y col., 2015).

ENZIMAS INDUSTRIALES

Como se comentó previamente, la obtención de GOS y péptidos es posible gracias a la acción de dos tipos de enzimas, llamadas β-galactosidasas y proteasas. El uso de enzimas para la transformación de productos no es algo novedoso: desde hace mucho tiempo, enzimas encontradas en la naturaleza han sido usadas para la producción de alimentos y la fabricación de productos, como el cuero (Kirk y col., 2002). El desarrollo de técnicas de fermentación permitieron la producción de enzimas en procesos industriales y los avances en las tecnologías de genes recombinantes han mejorado aún más estos procesos y permitieron la comercialización de enzimas que no pueden producirse a escala industrial naturalmente. Actualmente, el mercado de enzimas posee un tamaño estimado de 9.900 millones de dólares anuales (Grand View Research, 2020).

Entre las enzimas usadas a nivel industrial, el 50% son de origen fúngico, mientras que 35% provienen de bacterias y 15% de plantas (Saranraj y Naidu, 2014). Alrededor de 150 procesos industriales diferentes utilizan alguna enzima, o microorganismos vivos, siendo la industria alimenticia la predominante (Adrio y Demain, 2014; Liu y Kokare, 2017). Las enzimas representan herramientas de interés para bio-convertir subproductos de distintos procesos industriales en otros productos de mayor valor (Brandelli y col., 2015).

La mayor parte de las enzimas reportadas son utilizadas sólo en un número limitado de procesos industriales (Herbert, 1992; Elleuche y col., 2014). Esta limitación está dada por diversos factores. Por un lado, leves variaciones en condiciones fisicoquímicas (pH, temperatura, presencia de solventes) pueden afectar significativamente la estabilidad de las enzimas (Berini y col., 2017; Sysoev y col., 2021). Utilizando métodos tradicionales basados en cultivos, el número y la diversidad de organismos estudiados es limitado, ya que es difícil la obtención de cultivos puros (Staley y Konopka, 1985). Sin embargo, en las últimas dos décadas, se han desarrollado métodos independientes de cultivo para el estudio de muestras ambientales que demostraron el potencial de microorganismos no cultivados como fuente de enzimas novedosas (Handelsman y col., 1998; Chen y Pachter, 2005). Según Chen y Pachter (2005), la aplicación de técnicas modernas de genómica para el estudio de comunidades microbianas a partir de muestras ambientales, sin la necesidad de aislamientos y cultivos celulares, se denomina **metagenómica**. Ésta es generalmente reconocida como la metodología con mayor potencial para la identificación de enzimas novedosas (Lorenz y Eck, 2005). Entre 2014 y 2017 se han reportado más de 300 enzimas de origen metagenómico y aplicación industrial (Berini y col., 2017) y al menos 7 enzimas patentadas (Prayogo y col., 2020). Los avances en la metagenómica han sido de

gran ayuda para identificar enzimas novedosas, que no podrían estudiarse por métodos tradicionales, y representan un gran beneficio para las distintas industrias, como la alimenticia.

METAGENÓMICA

Obtener cultivos puros de muchos microorganismos ha sido la principal limitante para el estudio de comunidades microbianas (Handelsman, 2004). Para sobrepasar las limitaciones asociadas a las técnicas basadas en cultivos, se desarrollaron nuevos métodos basados en el estudio del ADN microbiano (Streit y Schmitz, 2004). Estos estudios generalmente se pueden dividir en 3 tipos: los basados en el estudio de un gen marcador o *metabarcoding*, los basados en expresión de fragmentos o "metagenómica funcional", y los basados en el estudio del ADN total de una muestra, o "**metagenómica basada en secuencia**".

El *metabarcoding* se basa en la caracterización de la comunidad estudiando a partir de un gen marcador, siendo el gen ribosomal de la subunidad pequeña 16S el más estudiado (Olsen y col., 1986). La ventaja de usar este gen es que permite establecer relaciones filogenéticas entre especies y proporciona un marco para el estudio de la ecología de una comunidad. Esta clase de estudios comienzan con la extracción del ADN ambiental de las muestras, y la posterior amplificación de alguna de las denominadas "regiones hipervariables" del gen (Tringe y Hugenholtz, 2008). La mayor limitante de este enfoque es que no es posible estudiar aspectos funcionales de la comunidad, ya que solo se amplifica un gen, o un fragmento del mismo, sin tener en cuenta el resto del genoma. Además, existen limitaciones en los análisis filogenéticos basados en el estudio de un solo gen, ya que no captura toda la variación genética que puede existir entre organismos.

Una alternativa que sí permite la caracterización funcional de una comunidad es la metagenómica funcional. La misma consiste en extraer y purificar el ADN de la comunidad estudiada, y luego utilizarlo para construir bibliotecas de expresión que serán clonadas en bacterias como *Escherichia coli* para su expresión. El tamaño del inserto de ADN metagenómico puede variar significativamente, yendo de fragmentos pequeños de menos de 10 Kb en vectores de secuenciación tradicional (Henne y col., 1999), a fragmentos de entre 25 y 35 Kb clonados utilizando fagos o cósmidos (Entcheva y col., 2001) o hasta insertos de 200 Kb en cromosomas bacterianos artificiales (Beja y col., 2000). Con los clones obtenidos se realizan ensayos de actividad con distintos sustratos para identificar qué insertos de ADN metagenómico codifican enzimas novedosas que se expresen activamente (Handelsman, 2004). Una vez identificado un clon positivo, este se puede secuenciar e identificar el gen que codifica para la proteína de interés, para profundizar su análisis.

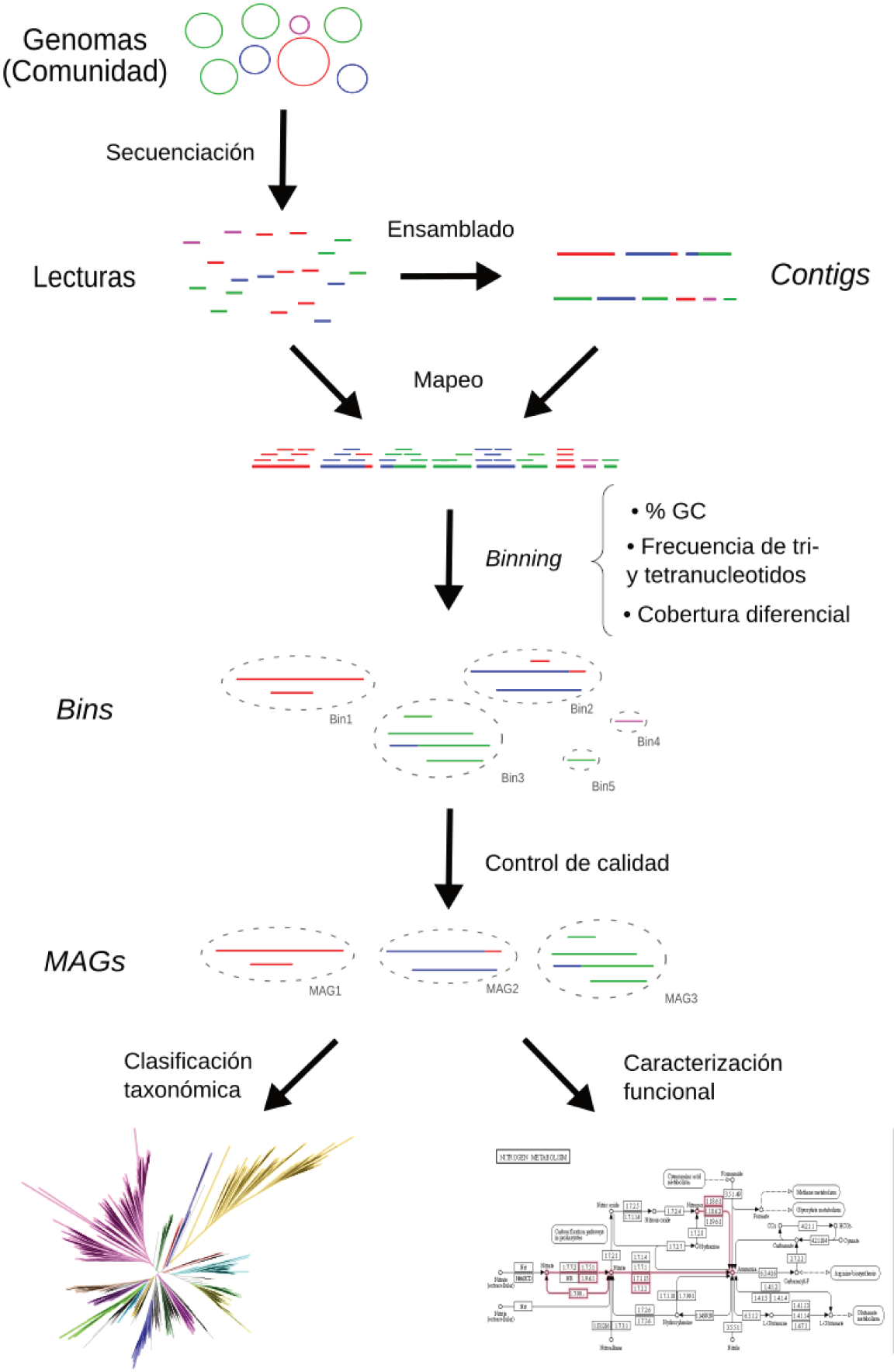

Figura 4.5: Diagrama del proceso de secuenciación, ensamblado y reconstrucción de genomas a partir de metagenomas.

La última alternativa es el uso de secuenciación masiva de todo el material genético extraído, sin necesidad de pasar por una etapa de clonado ni de amplificación. Esta alternativa ofrece una mirada global de la comunidad, sin los sesgos de la amplificación por PCR que implica el *metabarcoding* o de la expresión heteróloga, permitiendo un mejor entendimiento de la diversidad filogenética y el estudio de rutas metabólicas completas (Chen y Pachter, 2005). Las limitaciones en estos casos se basan en la posibilidad de identificar la función de cada uno de los genes a partir de la comparación de los mismos con bases de datos de secuencias.

En un principio, los análisis de metagenómica basados en secuencias se centraban solo en la identificación de genes individuales a partir de un ensamblado, la identificación de su función y luego en la caracterización global de las funciones identificadas. Pero los avances en las tecnologías de secuenciación y el desarrollo de nuevas herramientas bioinformáticas, permitió tomar un enfoque centrado en la reconstrucción del genoma completo de los distintos miembros de la comunidad (Albertsen y col., 2013; McMahon, 2015). Este enfoque permite tener un mejor entendimiento de la capacidad metabólica y estimar el rol que cumpliría cada especie dentro de una comunidad (Strous y col., 2012).

El proceso de reconstrucción de genomas o *binning* consiste en agrupar los *contigs* resultantes de un ensamblado de acuerdo a características intrínsecas de sus secuencias, como el porcentaje de guanina y citosina (%GC; Dick y col., 2009), la frecuencia de tri- y tetra-nucleótidos, y la cobertura obtenida luego de mapear las lecturas de las distintas muestras contra el ensamblado (Figura 4.5; Albertsen y col., 2013).

Una vez formados estos grupos de *contigs* o *bins*, es crítico realizar un análisis de su calidad. Tradicionalmente, se calidad de genomas obtenidos de aislamientos se determinó utilizando estadísticas de ensamblado como el N50 (Salzberg y col., 2012). Sin embargo, los estudios de células únicas (*single-cell*) y metagenómica han usado otras estrategias, como la identificación de genes marcadores de copia única (Raes y col., 2007; Wrighton y col., 2012; Swan y col., 2013). Herramientas automáticas como CheckM (Parks y col., 2015) permite estimar la completitud y contaminación de un genoma ensamblado de un metagenoma a partir de estos genes. La completitud se define como el porcentaje de los genes de copia única propios de un linaje que fueron encontrados en un *bin*, mientras que la contaminación se define como el porcentaje de estos genes que están presentes en más de una copia en el *bin*.

Para definir la completitud y contaminación de un bin es necesario primero determinar su linaje, para asi conocer el conjunto de genes marcadores a buscar. Brevemente, los autores de CheckM definieron un conjunto de 43 genes presentes en copia única en el 97% de los genomas disponibles para los dominios *Bacteria* y *Archaea* a la fecha de publicación de la herramienta. Estos genes son buscados en cada *bin* y comparados con el conjunto de referencia para asignarle un linaje taxonómico. Luego,

se buscan todos los genes marcadores específicos del linaje asignado dentro del *bin* y, de acuerdo con el número de genes encontrados, se calculan la completitud y la contaminación.

La obtención de genomas de alta calidad a partir de metagenómica permite expandir nuestro entendimiento sobre la evolución microbiológica, la diversidad metabólica y el rol que cumplen los microorganismos en procesos naturales e industriales (Parks y col., 2017).

LAGUNAS FACULTATIVAS

Las lagunas facultativas son la principal alternativa utilizada por pequeñas queserías para el tratamiento de sus efluentes, principalmente en países en desarrollo (von Sperling, 2007). Estos efluentes son retenidos en lagunas el tiempo necesario para que la materia orgánica presente en ellos sea estabilizada por la comunidad microbiana allí presente. Si bien estos sistemas requieren largos tiempos de para la estabilización de los efluentes y su eficiencia no es tan alta como los otros métodos, no tienen un alto costo operativo ya que no necesitan equipamiento extra y los costos de construcción son accesibles si se cuenta con las extensiones de tierra necesaria para instalarlas.

En la parte superior de la laguna se encuentra la zona aeróbica, donde la materia orgánica que queda diluida en el agua sin asentarse en el fondo de la laguna es oxidada mediante respiración aeróbica (figura 4.6). El oxígeno requerido es provisto por algas y microorganismos fotosintéticos. Dado que es necesaria luz solar para llevar adelante la fotosíntesis, la eficiencia de este tipo de sistemas depende directamente del clima del lugar. A mayor profundidad, la penetración de luz solar es menor, lo que causa que haya más consumo de oxígeno (respiración aeróbica) que producción, habiendo zonas en las que no hay oxígeno disuelto. Además, durante la noche no se produce fotosíntesis, por lo que la ausencia de oxígeno es más marcada. En estos casos, otros aceptores de electrones deben ser usados, como es el caso de los nitratos. A las zonas donde puede haber tanto presencia como ausencia de oxígeno se las denomina facultativas. Por último, los sólidos en suspensión tienden a asentarse en el fondo de la laguna y son degradados mediante procesos anaeróbicos hasta obtener dióxido de carbono, metano y otros compuestos, dejando solo la fracción inerte en el fondo de la laguna.

Normalmente, estos sistemas se organizan en más de una laguna, dispuestas en serie. Cada laguna está conectada con la siguiente, y el flujo de agua es provocado por la altura de la columna de agua. De esta manera se evita que el agua fluya en dirección contraria, o se mezclen los contenidos de distintas lagunas. En cada etapa, se recibe un efluente con menor carga orgánica que en la etapa anterior, debido a la acción de las comunidades microbianas presentes.

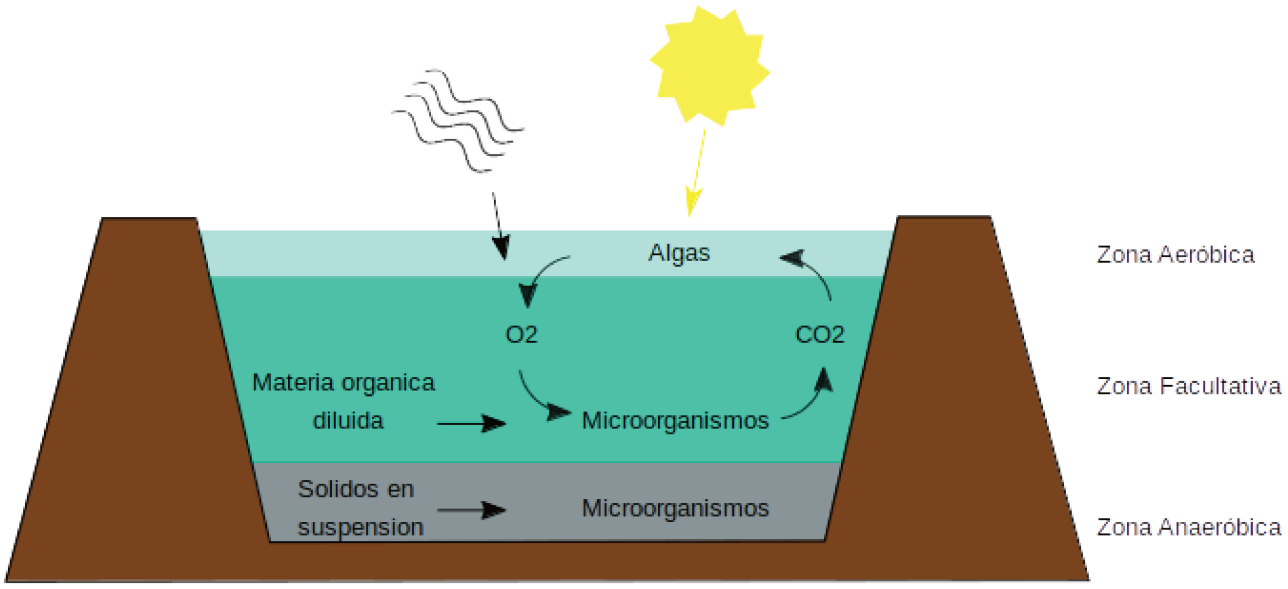

Figura 4.6: Esquema del funcionamiento de lagunas de estabilización facultativas.

Durante los últimos años, se han caracterizado las comunidades de distintos tipos de sistemas de tratamientos de distintos efluentes, tanto domiciliarios (Guo y col., 2017; Huang y col., 2021; Jankowski y col., 2022) o de residuos de sistemas de producción animal (MacGarvey y col., 2015, MacGarvey y col., 2017) Aunque la dinámica de las comunidades microbianas en estos sistemas está influenciada, en parte, por la carga de los efluentes volcados en ellas y las condiciones ambientales (Cydzik-Kwiatkowska y Zielińska, 2016), se han detectado grupos taxonómicos conservados en la mayoría de los sistemas. En sistemas aeróbicos se ha visto una gran proporción de miembros del *phylum Proteobacteria*, principalmente de las clases *Gammaproteobacteria* y *Betaproteobacteria*, y en menor medida de las *phyla Bacteroidetes* y *Actinobacteria* (MacGarvey y col., 2005, MacGarvey y col., 2007; Belila y col., 2013; Guo y col., 2017; Huang y col., 2021; Jankowski y col., 2022), mientras que en sistemas anaeróbicos se ha detectado una fuerte dominancia del phylum Firmicutes (MacGarvey y col., 2005, MacGarvey y col., 2007). Por lo tanto, el conocimiento sobre la composición de comunidades en ambientes facultativos es limitado.

Los efluentes de las queserías incluyen el agua y los químicos usados para el lavado de las máquinas y los restos de diluciones de leche y suero que hayan quedado en las mismas (Carvalho y col., 2013). Dado que la principal fuente de materia orgánica dentro de estas lagunas son diluciones de leche, es de esperar que los microorganismos que viven en ellas presenten enzimas que le permitan modificar sus componentes.

5 OBJETIVOS

El objetivo principal del presente trabajo es estudiar las comunidades microbianas en lagunas de estabilización de PYMES lácteas utilizando herramientas metagenómicas, e identificar enzimas para la revalorización de la lactosa y las proteínas del lactosuero.

Para conseguirlo, se plantean 3 objetivos particulares:

1) Analizar y comparar las comunidades microbianas de lagunas de estabilización de PYMES lácteas mediante secuenciación masiva de ADN metagenómico.

2) Identificar genes de interés biotecnológico para la transformación de los principales componentes de lactosuero: proteínas y lactosa.

3) Expresar y evaluar la actividad de proteasas y β-galactosidasas recombinantes.

MATERIALES Y MÉTODOS

6 MATERIALES Y MÉTODOS

LAGUNAS DE ESTABILIZACIÓN

Las lagunas de tratamientos de efluentes pertenecen a 2 PyMES lácteas ubicadas en el centro de la provincia de Santa Fe. Las empresas fueron denominadas AUR y CYC. AUR posee un sistema de 2 lagunas conectadas en serie, la primera posee un tamaño de 80 x 40 m y 1,5 m de profundidad, mientras que el tamaño de la segunda es de 90 x 40 m y 2.5 m de profundidad. Por su parte, el sistema de CYC cuenta con 4 lagunas interconectadas dispuestas en serie, todas de un tamaño de 35 x 50 m; las 2 primeras poseen una profundidad de 2,5 m, mientras que la profundidad de las 2 últimas es de 1,5 m. Por cada laguna, se tomaron 3 muestras en distintos puntos, para conocer la variabilidad espacial. Las muestras se tomaron intentando representar toda la columna de líquido desde el fondo a la superficie.

La temperatura, el pH y la conductividad fueron medidos con el medidor multiparamétrico Horiba U-50. Los parámetros restantes fueron determinados mediante los análisis fisicoquímicos de agua y efluentes, realizados por la Universidad Tecnológica Nacional (Rafaela, Santa Fe; Tabla S1.).

MUESTREO Y EXTRACCIÓN DE ADN METAGENÓMICO

De cada una de las 6 lagunas a estudiar se tomaron 3 muestras de 1 l de agua. Para la extracción de ADN, 20 ml de agua fueron pasados por un filtro de queso y centrifugados a 1000 rpm, por 10 minutos. Los pellets fueron resuspendidos en 1 ml de Tris-EDTA 50 mM pH 8 y congelados. Se agregaron 60 µl de lisozima (10 mg/ml) en Tris 250 mM pH 8, se lo incubó hasta descongelar y luego se lo incubó por 45 minutos en hielo. Se agregó una solución de 0.5% SDS, 50 mM Tris pH 7,5, 0,4 M EDTA y 1 mg/ml de proteinasa K (120 µl) y se incubó por 1 hora a 50ºC, mezclando por inversión cada 10 minutos. La suspensión fue transferida a un tubo conteniendo las perlas del kit de purificación PowerWater Isolation kit (MOBIO) y agitado en vortex por 5 minutos. Luego se centrifugó esta mezcla durante un minuto a 13000 rpm y el sobrenadante fue transferido a un tubo limpio. Finalmente, se realizó una extracción con fenol-cloroformo y una precipitación con etanol (Sambrook y Russell, 2006). El ADN resultante fue cuantificado utilizando el espectrofotómetro Nanodrop 2000 (Thermo).

SECUENCIACIÓN Y ANÁLISIS DEL AMPLICÓN DEL 16S

Para las 18 muestras tomadas se amplifico la región V4 del gen ribosomal del 16S, usando los cebadores F515-TAG: 5’ **CACGACGTTGTAAAACGAC**GTGCCAGCMGCCGCGGTAA 3’ y R806-TAG: 5’ **CAGGAAACAGCTATGACC**GGACTACVSGGGTATCTAAT 3’. Estos cebadores son versiones modificadas de F515 y R806 (Caporaso y col., 2011), a los que se le agregó en el comienzo sitios del plásmido M13 (en negrita). La amplificación fue hecha con el sistema FastStart High Fidelity PCR (Roche). Para cada muestra se realizó una segunda ronda de PCRs para agregar un identificador

único y los adaptadores A y B de Roche, siguiendo las instrucciones del fabricante. Estas bibliotecas de secuenciación fueron amplificadas nuevamente mediante una PCR basada en emulsión (emPCR) y secuenciadas en la plataforma Genome Sequencer FLX (Roche Applied Science) en INDEAR (Argentina), utilizando la química Titanium, de acuerdo con las indicaciones del fabricante.

Las lecturas obtenidas fueron analizadas con QIIME2 (version 2017.2; Bolyen y col., 2019). Brevemente, primero se demultiplexaron las lecturas usando usearch (Edgar, 2010) y luego fueron cargadas a QIIME2 como lecturas *singles* demultiplexadas. Con el programa cutadapt se eliminaron adaptadores y con la implementación disponible en QIIME2 de dada2 (Callahan y col., 2015) se eliminó el ruido, utilizando la opción “denoise-pyro”. A continuación, se realizó un agrupamiento *de novo* de las lecturas, con 97% como umbral de identidad. Con las secuencias obtenidas se calculó la diversidad de Shannon para estimar la diversidad alfa y construir curvas de rarefacción. Luego, las secuencias fueron alineadas con la implementación de MAFFT (Katoh y Standley, 2013) en QIIME2 y se construyó un árbol filogenético con la implementación de RAxML (Stamatakis, 2014) en QIIME2. Esta filogenia fue utilizada para calcular el índice de UniFrac ponderado como estimador de la diversidad beta. Con estos últimos resultados se calcularon las distancias entre muestras utilizando el escalado multidimensional no métrico (nMDS).

SECUENCIACIÓN Y ANÁLISIS DE WGS

La secuenciación de genoma completo se hizo a partir de una mezcla compuesta para cada laguna. Ya que no se observó variación significativa entre las muestras pertenecientes a una misma laguna, se juntaron las 3 réplicas para formar una solo muestra compuesta por laguna, utilizando la misma masa (ng) de cada una de ellas. Las 6 muestras resultantes fueron secuenciadas utilizando la plataforma Illumina HiSeq 1500 (INDEAR), generando lecturas pareadas de 150 bp. Estas fueron recortadas utilizando Trimmomatic (v0.36; Bolger et al., 2014), con parámetros por defecto, para remover adaptadores y secuencias de baja calidad. Las estadísticas de lecturas obtenidas se muestran en la tabla 7.2. Luego se estimó la cobertura lograda para cada muestra con Nonpareil (v2.4r1; Rodriguez-R y Konstantinidis, 2014a).

Con las lecturas preprocesadas, se llevó a cabo una clasificación taxonómica utilizando KRAKEN2 (v2.0.6; Wood et al, 2019) contra la base de datos estándar y contra la base de datos de plantas, con el fin de identificar la presencia de algas fotosintéticas.

Luego se diseñó la estrategia para la reconstrucción de genomas. Dado que en este tipo de lagunas el agua fluye de una laguna a la siguiente, llevando tanto materia orgánica como microorganismos, se planteó la hipótesis de que gran parte de la diversidad microbiológica está conservada a lo largo de todo el sistema. Esta hipótesis fue confirmada tanto por el análisis de beta diversidad hecho en base a las lecturas de 16S, como por la clasificación taxonómica de lecturas de WGS. Por lo tanto, se combinaron las lecturas en dos conjuntos de datos, AUR y CYC, para aumentar la profundidad y

utilizar la estrategia de cobertura diferencial (Albertsen y col., 2013) para mejorar el ensamblado. Se consideró a cada laguna como una muestra del sistema, para darle sustento estadístico al proceso de *binning*. Los dos sets de datos fueron ensamblados con IDBA_UD (v1.1.2; Peng et al., 2012), y la cobertura de cada *contig* obtenido fue calculada con Bowtie2 (v2.2.4; Langmead y Salzberg, 2012) y Samtools (v1.3.1; Li, 2011). Los *contigs* de al menos 1000 bp fueron agrupados en *bins* con MaxBin2 (v2.2.5; Wu et al., 2016) y la calidad de los mismos fue evaluada con CheckM (v1.0.11; Parks et al., 2015). De acuerdo con el planteo de los autores de CheckM, se consideraron como genomas ensamblados de un metagenoma (MAG, por las siglas en inglés) a aquellos *bins* que tuvieran por lo menos 70% de completitud y menos de 20% de contaminación. Aquellos MAGs con más de 90% de completitud y menos de 5% de contaminación se los denominó “de alta calidad”, mientras que el resto fueron denominados en borrador (*draft*). Los genomas en borrador fueron examinados manualmente para eliminar *contigs* que hayan sido mal asignados, teniendo en cuenta genes marcadores, %GC y cobertura diferencial. Finalmente, la calidad del grupo definitivo de *bins* fue nuevamente evaluado con CheckM y clasificado taxonómicamente usando GTDB-tk (Chaumeil et al., 2019) y la base de datos GTDB (versión 95; Parks et al., 2018). A aquellos MAGs clasificados como miembros del filo *Patescibacteria* se los analizó con el clasificador de ANVI’O (Eren et al., 2021), para chequear su pertenencia al filo, completitud y contaminación. Finalmente, los MAGs fueron comparados entre sí, para encontrar microorganismos presentes en ambos sistemas. Para ello se calculó la identidad nucleotídica promedio (ANI, Rodriguez-R y Konstantinidis, 2014b) como fue propuesto por Goris y col (2007). Se utilizó un tamaño de fragmento de 1000 bp y una ventana de 700 bp, considerando sólo pares de genomas que tuvieran al menos 50% de los fragmentos alineados.

El árbol filogenético se construyó a partir del alineamiento publicado por Hug y col (2016). Brevemente, se descargaron de pfam (Mistry y col., 2021) los perfiles de HMM para 16 genes ribosomales: L2 (PF00181), L3 (PF00297), L4 (PF00573), L5 (PF00281), L6 (PF00347), L14 (PF00281), L15 (PF00827), L16 (PF00252), L18 (PF00861), L22 (PF00237), L24 (TIGR1079 y TIGR1080), S3 (PF00410), S8 (PF00410), S10 (PF00338), S15 (PF00366) y S19 (PF00203). Para cada MAG se buscaron los 16 genes y los resultados fueron agregados al alineamiento base usando MAFFT (versión 7.205). Este alineamiento fue utilizado para construir el árbol filogenómico, usando Iqtree (versión 1.6.7; Nguyen y col., 2015), con el modelo LG, 1000 aLRTs y 1000 bootstraps. La figura final fue construida usando Dendroscope3 (versión 3.7.3; Huson y Scornavacca, 2012) e Inkscape (versión 0.92; Bah, 2007).

Se predijeron genes para cada MAG y para los *contigs* no clasificados usando PRODIGAL (versión 2.6.3; Hyatt et al., 2010) y estos fueron anotados comparando contra las bases de datos KEGG (Kanehisa y Goto, 2000; Aramaki y col., 2019), MEROPS (Rawlings y col., 2016) y CAZy (Drula y col., 2021). Para esta última, se utilizaron los perfiles de HMM disponibles en la base de datos dbCAN (Yin y col., 2012) y la herramienta HMMer3 (hmmer.org), descartando todas las secuencias cuyo

alineamiento sea menor al 60% del tamaño del perfil, para eliminar falsos positivos. En el caso de KEGG, solo se usaron los módulos listados en la categoría de "*Metabolism*" y se consideró completo a un módulo cuando al menos el 50% de los genes que lo componen estaban presentes y si se encontraron genes claves para ese proceso (por ejemplo, genes RuBisCo para la fijación de carbono).

IDENTIFICACIÓN Y CLASIFICACIÓN DE B-GALACTOSIDASAS

A partir de los genes predichos como *CAZymes*, se buscaron todos aquellos que pertenecieran a las familias con actividad β-galactosidasa reportada (GH1, GH2, GH35, GH42 y GH59). Uno de los objetivos fue capturar la mayor diversidad posible, tanto a nivel de familia de glicosil hidrolasa como a nivel taxonómico, para elegir los candidatos a clonar. Por ello, estas secuencias fueron comparadas contra la base de datos nr del NCBI (O'Leary y col., 2016) y el mejor *hit* encontrado se usó para determinar la clasificación taxonómica de cada secuencia.

En el caso de las GH2, se realizó un análisis de las secuencias basado en la clasificación propuesta por Talens-Perales y col. (2016), donde se identificaron 2 dominios presentes en todas las GH2 y 4 dominios adicionales que definirían la actividad de la enzima. En primer lugar se descargaron los alineamientos de cada uno de estos 6 dominios de Pfam: GH2N (PF02837), GH2d (PF00703), GH2C (PF02836), DUF4981 (PF16353), DUF4982 (PF16355) y Bgal_Small_N (PF02929). Usando HMMer, se buscaron estos 6 perfiles en todas las secuencias clasificadas como GH2 y, en base a la presencia o ausencia de estos perfiles, se definió qué arquitectura de dominio correspondía a cada secuencia. Además, se utilizó la herramienta web de HMMer (https://www.ebi.ac.uk/Tools/hmmer/) para comparar las secuencias contra la base de datos completa de Pfam y determinar si otros dominios adicionales estaban presentes en el extremo C-terminal.

IDENTIFICACIÓN Y CLASIFICACIÓN DE PROTEASAS

Dado que MEROPS no posee una base de datos de perfiles de HMM como es el caso de las β-galactosidasas, la búsqueda de proteasas se realizó usando BLAST. En primer lugar, desde MEROPS se descargaron todas las secuencias reportadas para cada familia reportada. Luego, para cada archivo se corrió una búsqueda con BLAST usando los genes del metagenoma como *query* y las secuencias de MEROPS como *subject*. Se descartaron todos los resultados que tuvieran menos de 50% de cobertura de *query* y, en caso de que un mismo un gen *query* diera resultados con más de una familia, se lo clasificó en aquella para la cual el porcentaje de identidad fuera mayor. Una vez que se contó con el set definido de proteasas, fueron comparadas con la base de datos *nr*, usando BLAST, y se usó el mejor hit para determinar su clasificación taxonómica.

AMPLIFICACIÓN DESDE METAGENOMA

A partir de los genes identificados, se seleccionaron 18 β-galactosidasas y 10 proteasas candidatas para su clonado y expresión, buscando capturar diversidad tanto de familia como taxonómica. Una vez

seleccionados los candidatos, se diseñaron cebadores para su amplificación (Tabla A.2). Los genes fueron amplificados usando el ADN metagenómico como molde y una ADN polimerasa con corrección de errores (AccuPrime, Thermo Scientific, Waltham, MA, USA). Todas las amplificaciones contaron con un paso inicial de desnaturalización a 94°C de 3 minutos, seguido por 35 ciclos compuestos por un paso de desnaturalización a 94°C durante 30 segundos, un paso de anillado de 30 segundos a la temperatura adecuada para cada par de cebadores (Tabla suplementaria S5) y un paso de extensión a 72°C de 3 minutos. Se concluyó con un paso de extensión a 72°C durante 10 minutos. Los productos de PCR obtenidos fueron chequeados por electroforesis en geles de agarosa al 1% y luego purificados con el kit Wizard® PCR Clean Up (PROMEGA).

CLONADO Y EXPRESIÓN EN *ESCHERICHIA COLI*

Se utilizaron dos cepas de *E. coli* para el clonado y la expresión de las enzimas. En primer lugar, para la propagación de plásmidos se utilizó *E. coli* DH5α (Thermo fisher, Waltham, MA, USA) y para la expresión de las enzimas *E. coli* BL21 (DE3). Las cepas fueron hechas competentes químicamente y transformadas de acuerdo a los protocolos descritos en Sambrook y Russell (2006).

Los productos de PCR purificados fueron nuevamente amplificados con cebadores que permitieran incorporar sitios de restricción (Tabla S3) y fueron clonados en el vector pGEM-T-easy (Promega, Madison, WI, USA). Para cada gen, se cortó el plásmido utilizando las enzimas de restricción correspondientes a los sitios incorporados, siguiendo las instrucciones del proveedor (Thermo Scientific, Waltham, MA, USA). El gen con los sitios de restricción y el plásmido fueron ligados usando ligasa T4 (Fermentas) siguiendo instrucciones del fabricante y la construcción obtenida fue transformada en *E. coli* DH5α para su propagación. Las colonias fueron incubadas toda la noche a 37°C en placas de agar y caldo de Lisogenia (LB), con ampicilina (100 µg/ml) y X-gal (0,2 mg/ml). Finalmente, las colonias fueron seleccionadas utilizando el sistema de blancas/azules del vector pGEM-T-Easy y chequeadas por PCR para verificar la presencia del gen usando los cebadores específicos de cada gen.

Usando el kit de miniprep (PROMEGA), los plásmidos pGEM-T-easy modificados fueron extraídos de las colonias de *E. coli* DH5α y luego digeridos con las enzimas de restricción correspondientes. Los genes liberados fueron clonados en el vector pET-TEV, un vector basado en pET28 incorporando un sitio de reconocimiento para la proteasa TEV cadena abajo de la cola de histidina, en el extremo N terminal. Esta construcción se utilizó para transformar *E. coli* BL21, la cual fue incubada toda la noche a 37°C en LB. Este cultivo se utilizó para inocular 10 ml de medio mínimo M9 (Miller, 1972), con kanamicina (30 µg/ml) a una DO600 inicial de 0,1. Este nuevo cultivo se incubó a 37°C y 200 rpm y la expresión de los genes se indujo cuando la DO600 alcanzó el valor de 1, usando 100 µM isopropil β-D-1-galactopiranosido (IPTG). El cultivo inducido fue incubado 6 h más a 22°C.

Las células y el sobrenadante fueron separados por centrifugación, a 5000 rpm durante 15 minutos. Los *pellets* fueron resuspendidos en buffer fosfato (fosfato pH 8 100 mM, NaCl 300 mM). Las células resuspendidas fueron lisadas por sonicado en frío y los extractos crudos fueron clarificados por centrifugación durante 15 minutos, a 15.000 rpm.

Finalmente, las enzimas fueron purificadas por cromatografía de afinidad, usando resina de agarosa Ni2+-NRA (Invitrogen, Carlsbad, CA, USA), de acuerdo al protocolo sugerido por el fabricante. Tanto extractos crudos como las enzimas purificadas fueron analizadas por SDS-PAGE en geles de poliacrilamida 12% (Laemmli, 1970). La concentración de proteínas purificadas fue medida con Nanodrop.

CHEQUEO POR SECUENCIACIÓN DE CAPILARES

Se tomaron colonias transformadas con el plásmido pGEM-T-easy y se realizó una extracción del plásmido utilizando el kit (PROMEGA). Estos plásmidos fueron secuenciados mediante electroforesis capilar (Unidad de Genómica – CICVyA - INTA). Los electroferogramas obtenidos fueron preprocesados con preGAP4 y luego analizados con GAP4 (Staden y col., 2000). Las secuencias obtenidas fueron alineadas con los genes predichos a partir del ensamblado metagenómico, para identificar la presencia de mutaciones.

CLONADO Y EXPRESIÓN EN *SACCHAROMYCES CEREVISIAE*

Para aquellas enzimas que no pudieran ser expresadas en forma soluble en *E. coli*, se utilizó como sistema alternativo el de *Saccharomyces cerevisiae BJ3505* (Eastman Kodak Company, Rochester, NY, USA). A partir de un producto de PCR purificado del gen candidato, se incorporaron por PCR sitios homólogos al vector YEp (Eastman Kodak Company; Tabla A.S3). Luego, las construcciones fueron clonadas utilizando las recombinasas de *S. cerevisiae*, como fue reportado en Becerra y col. (2001). Brevemente, las células fueron transformadas, incorporando el vector y el gen con los sitios homólogos mediante electroporación. Brevemente, para que las células entren en estado de electrocompetencia primero se las cultivó toda la noche en medio YPD (extracto de levadura 1% p/v, peptona 2% p/v, glucosa 2% p/v) a 30°C y 250 RPM. Luego se realizó una dilución 1/7 de este cultivo y se lo incubó nuevamente a 30°C y 250 RPM por 3 hs, hasta alcanzar una DO600 inicial de 1. Este cultivo fue centrifugado 5 minutos a 5000 RPM. Se descartó el sobrenadante, el *pellet* fue resuspendido en igual volumen de agua esteril y se centrifugó nuevamente por 5 minutos a 5000 RPM. Este lavado se repitió 3 veces. Luego se hicieron 2 lavados adicionales, pero con la mitad de volumen y usando glicerol 10% en lugar de agua. Luego de estos lavados, se centrifugó nuevamente por 5 minutos a 5000 RPM y el pellet resultante fue resuspendido en 1 ml de glicerol por cada 60 µl de pellet. Para la transformación, se usaron 50 µl de células electrocompetentes, 2 µl de plásmido y 4 µl de inserto. La electroporación se realizó a 1,5 kV, 25 µF y 200 Ω. Luego, se agregó 1 ml de YPD y las células transformadas se incubaron por 20 minutos a 30°C. Por último, las células fueron cultivadas en

placas de medio mínimo y agar (YNB 1X Invitrogen Q30007, dropout sin triptófano 1X, glucosa 2% p/v, agar 2% p/v) durante 48 horas a 30°C.

Las colonias resultantes fueron chequeadas por PCR para verificar que hayan incorporado el gen. Aquellas colonias positivas fueron cultivadas 24hs en medio mínimo (YNB 1X Invitrogen Q30007, dropout sin triptófano 1X, glucosa 2% p/v) y luego se realizó una dilución 1/100 en el medio de expresión YPHSM (glucosa 2% p/v, glicerol 3% v/v, extracto de levadura y peptona 8% p/v) y se las cultivó a 30°C durante 7 días.

Dado que estas construcciones incorporan a la proteína un péptido señal que permite su secreción al medio, para obtener las proteínas recombinantes se centrifugaron alícuotas de 1 ml durante 15 minutos a 5000 RPM y se filtró el sobrenadante utilizando filtros de 3 kDa. La expresión de la enzima fue controlada por SDS-PAGE en geles de poliacrilamida 12%.

CARACTERIZACIÓN DE B-GALACTOSIDASAS

La actividad β-galactosidasa fue medida usando 2-nitrofenil-β-D-galactopiranosida (ONPG). Las enzimas purificadas fueron diluidas en buffer Z (Na2HPO4 100 mM, NaH2PO4 40 mM,

KCl 10 mM y MgSO4 1,6 mM) e incubadas a 30°C por 4 minutos. Se dió inicio a la reacción agregando a esa solución un igual volumen de sustrato (4 mg/ml) en buffer Z. Se alicuotó la mezcla y se detuvo la reacción agregando igual volumen de Na2CO3 1M. Finalmente, se midió el p-nitrofenol liberado por absorbancia de UV a 405 nm. Se definió una unidad de actividad β-galactosidasa (U) como la cantidad de enzima capaz de liberar 1 µmol de p-nitrofenol por minuto (µmol min−1 mL−1) en las condiciones experimentales.

Para determinar el pH óptimo de la enzima, se realizó el mismo ensayo con ONPG, usando buffer de Briton-Robinson a distintos pH, entre 4 y 10 (20 mM ácido acético, 20 mM ácido fosfórico y 20 mM ácido bórico boric acid titulado con 1 M NaOH para alcanzar el pH deseado). Por otra parte, para determinar la temperatura óptima se evaluó la actividad con ONPG como sustrato, entre 30°C y 65°C. Una vez determinada la temperatura óptima, se evaluó la termoestabilidad incubando la enzima en buffer Z a dicha temperatura durante 24h y midiendo la actividad regularmente, usando ONPG como sustrato. Finalmente, se estudió el efecto de diferentes iones en la actividad enzimática: MgCl2, ZnSO4, CaCl2 y KCl. Para realizar este ensayo, se agregaron diferentes concentraciones de las sales (1 y 10 nM) al ensayo estándar de actividad con ONPG.

La hidrólisis de lactosa fue determinada midiendo la producción de glucosa. La enzima purificada fue diluida en buffer Z y se agregó igual volumen de sustrato. La solución se incubó a temperatura óptima, entre 20 y 960 minutos (16 hs). Para detener la reacción, se incubó la solución a 96°C durante 5 minutos. Se definió una unidad de actividad β-galactosidasa (U) como la cantidad de enzima capaz de

liberar 1 µmol de D-glucosa por minuto (µmol min−1 mL−1) en las condiciones experimentales. La concentración de glucosa fue medida utilizando el kit comercial D-Glucose GOD-POD (Wiener)

Por último, la concentración de GOS fue determinada por HPLC (HPLC Waters Breeze I), utilizando la columna Sugar Pack Waters (6.5 mm × 300 mm) y como fase sólida 100 µM EDTA-Calcio (Sigma Aldrich, St. Louis, MO, USA). La temperatura de la columna se fijó en 80°C, la temperatura del sensor en 37°C, con una sensibilidad de 32 y un flujo de 0,5 ml/min. Los azúcares eluidos fueron detectados usando el detector de índice de refracción Waters 2414. Los estándares usados para la identificación y cuantificación de los azúcares fueron una mezcla de estaquiosa, rafinosa, sacarosa y galactosa.

Los parámetros cinéticos de la hidrólisis fueron determinados realizando el ensayo de actividad descrito anteriormente, con diferentes concentraciones de sustrato (de 0 a 20 mM). Las reacciones se llevaron a cabo a 30°C de 6 a 20 minutos. Las medidas fueron tomadas por triplicado.

CARACTERIZACIÓN DE PROTEASAS

La actividad proteasa fue determinada usando como sustrato azocaseína (Sigma). Se tomaron 100 µl de sobrenadante de cultivo de *S. cerevisiae* y 100 µl de azocaseína (1% en Tris-HCl pH 8), y fueron incubados por 30 minutos a 30°C. Se detuvo la reacción agregando 200 µl de TCA 10%. Las muestras fueron centrifugadas por 5 minutos a 10.000 rpm y se transfirieron 100 µl del sobrenadante a 100 µl de NaOH 1N. Finalmente, se midió el sustrato hidrolizado por absorbancia de UV a 405 nm. Una unidad de actividad proteasa fue definida como la cantidad de enzima necesaria para aumentar la absorbancia 0.01 a A405 en las condiciones experimentales anteriores (Pushpam y col., 2011). El cálculo se hizo de la siguiente manera:

$$U = (Ae - An) / (Ve * 0.01)$$

donde Ae es la absorbancia de la reacción con la enzima Pr10, An es la absorbancia de la reacción realizada con el sobrenadante de un cultivo de *S. cerevisiae* sin transformar (negativo) y Ve es el volumen de enzima utilizado (100 µl).

Para determinar la temperatura óptima se utilizó azocaseína como sustrato, incubando la reacción durante 30 minutos a diferentes temperaturas entre 25°C y 55°C. Para determinar el pH óptimo se realizó el mismo ensayo, pero disolviendo la azocaseína en buffer de Briton-Robinson a distintos pH, entre 5 y 9 (20 mM ácido acético, 20 mM ácido fosfórico y 20 mM ácido bórico boric acid titulado con 1 M NaOH para alcanzar el pH deseado).

Para evaluar la actividad proteasa con sustrato natural, se utilizaron placas de Agar leche (1%). Se sembraron 10 µl del sobrenadante utilizado para la determinación con azocaseína, junto con 10 µl de células sin transformar como control negativo y la enzima Novozym® (10028, Novozymes) como control positivo.

FIGURAS

Todos las figuras presentadas fueron hechas con el paquete ggplot2 (Wickham, 2006), de R (R Core Team, 2022), e Inkscape.

COMUNIDADES MICROBIANAS EN LAGUNAS DE ESTABILIZACIÓN FACULTATIVAS

7 COMUNIDADES MICROBIANAS EN LAGUNAS DE ESTABILIZACIÓN FACULTATIVAS

INTRODUCCIÓN

Si bien el uso de técnicas independientes de cultivo, como el estudio del gen ribosomal 16S, para el estudio de comunidades fue propuesto hace más de 30 años (Woese, 1987), solo recientemente se empezaron a reportar los primeros genomas obtenidos desde metagenomas (Tyson y col., 2004). Los avances en las técnicas de secuenciación han dado lugar a una explosión en el número de estudios sobre muestras ambientales usando metagenómica basada en secuencia y, sobretodo, en el número de genomas publicados a partir de este tipo de estudios (Reddy y col., 2015; Mitchell y col, 2018). El acceso a organismos novedosos, solo descritos por análisis metagenómicos (Hug y col., 2016; Parks y col., 2017), ha llevado incluso a replantear los sistemas de clasificación taxonómica utilizados (Parks y col., 2018).

Como se mencionó anteriormente, la reconstrucción de genomas permite conocer mejor el rol que cumplen los microorganismos en procesos metabólicos naturales e industriales (Parks y col., 2017). Particularmente, en lagunas de estabilización los principales procesos que ocurren son el ciclo del carbono y del nitrógeno (Grady y col., 2011). Si bien hay numerosos reportes sobre sistemas de estabilización de efluentes (McGarvey y col.,2005; McGarvey y col., 2007; Belila y col., 2013; Cydzik-Kwiatkowska y Zieliińska, 2016), muchos se basan en el estudio del gen 16S, lo cual limita la predicción funcional sobre los grupos taxonómicos identificados.

El presente capítulo plantea el estudio de la comunidad microbiana de dos sistemas de lagunas de estabilización de PyMES lácteas, denominadas AUR y CYC. El primer sistema, AUR está compuesto por dos lagunas dispuestas en serie (AUR1 y AUR2), mientras que el segundo sistema está compuesto por 4 lagunas en serie (CYC1-4). En primer lugar se realizó una primera caracterización a partir de la secuenciación del gen ribosomal 16S. A partir de los resultados obtenidos, se realizó la secuenciación masiva (WGS, por sus siglas en inglés) de todo el ADN metagenómico, con la intención de realizar una caracterización taxonómica y funcional de la comunidad, utilizando una estrategia basada en la reconstrucción de genomas completos y la identificación de genes y rutas metabólicas completas.

ANÁLISIS DEL AMPLICÓN DEL GEN 16S

La secuenciación del amplicón 16S generó en total 178.940 lecturas, de las cuales 3.217 fueron destacadas por no contar con un *barcode* específico. Las muestras contaron con un promedio de 9.762 lecturas cada una, con 17.918 como valor máximo (M3 de CYC4) y 3.092 como mínimo (M2 de

CYC3, Tabla 7.1). Después del control de calidad, 68.537 (39%) lecturas fueron descartadas por baja calidad o ser quimeras.

Tabla 7.1: Resultados de la secuenciación del amplicón del gen 16S

Laguna	M1		M2		M3	
	Lecturas crudas	Lecturas finales	Lecturas crudas	Lecturas finales	Lecturas crudas	Lecturas finales
AUR1	14.438	9.041	5.911	4.084	7.602	5.088
AUR2	4.752	3.636	6.870	5.400	5.646	4.553
CYC1	16.985	9.111	9.499	5.579	13.026	7.415
CYC2	16.919	9.328	5.526	3.358	10.985	6.093
CYC3	8.059	4.838	3.092	2.092	10.674	6.271
CYC4	5.278	3.527	12.543	7.571	17.918	10.201

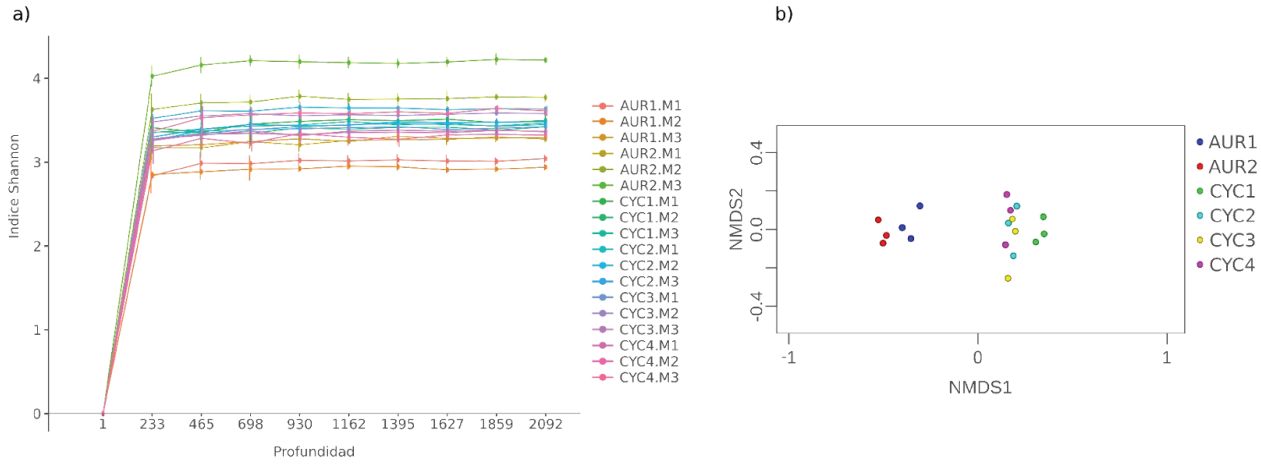

Figura 7.1: Resultados de la secuenciación del amplicón del gen del 16S. a) Curvas de rarefacción de las 18 muestras secuenciadas, utilizando el índice de Shannon. b) Escalado multidimensional no métrico (nMDS) de las muestras, utilizando el índice UniFrac ponderado.

Las curvas de rarefacción mostraron que todas las muestras llegaron a un *plateau*, indicando que la profundidad de secuenciación fue suficiente (figura 7.1a). El análisis de beta-diversidad indicó que las muestras de una misma empresa (AUR o CYC) contaban con composiciones similares, y que las composiciones entre empresas se diferenciaban (figura 7.1b). Basado en este resultado, para la secuenciación completa del ADN metagenómico, se decidió combinar las réplicas y secuenciar una muestra compuesta para cada laguna.

Análisis de Secuenciación *Shotgun*

En la secuenciación de tipo *shotgun* se generaron más de 200 millones de lecturas, sumando más de 60 Gb (tabla 7.2). Después del control de calidad, alrededor del 1% de las lecturas de cada muestra fueron descartadas. De acuerdo al software Nonpareil, la cobertura media lograda fue del 86%, lo que sugiere que el esfuerzo de secuenciación fue suficiente para lograr un ensamblado de calidad (figura 7.2). Es decir, el muestreo debería cubrir la mayor parte de la diversidad presente en las muestras.

Tabla 7.2: Resultados de la secuenciación tipo *shotgun*. Cantidad de lecturas obtenidas por la secuenciación de genoma completo y las lecturas que se mantuvieron luego del filtrado por calidad.

Muestra	Lecturas crudas	Lecturas post filtrado por calidad					
		Pareadas	% del total	No pareadas	% del total	Descartadas	% del total
AUR1	31.790.925	23.152.130	72,38%	8.299.000	26,10%	339.795	1,07%
AUR2	27.509.873	20.216.656	73,49%	6.912.935	25,13%	380.282	1,38%
CYC1	31.475.370	21.674.483	68,86%	9.563.122	30,38%	237.765	0,76%
CYC2	32.632.974	21.416.110	65,63%	10.960.256	33,59%	256.608	0,79%
CYC3	40.218.397	29.596.551	73,59%	10.281.985	25,57%	339.861	0,85%
CYC4	36.510.040	23.346.096	63,94%	12.796.972	35,05%	366.972	1,01%

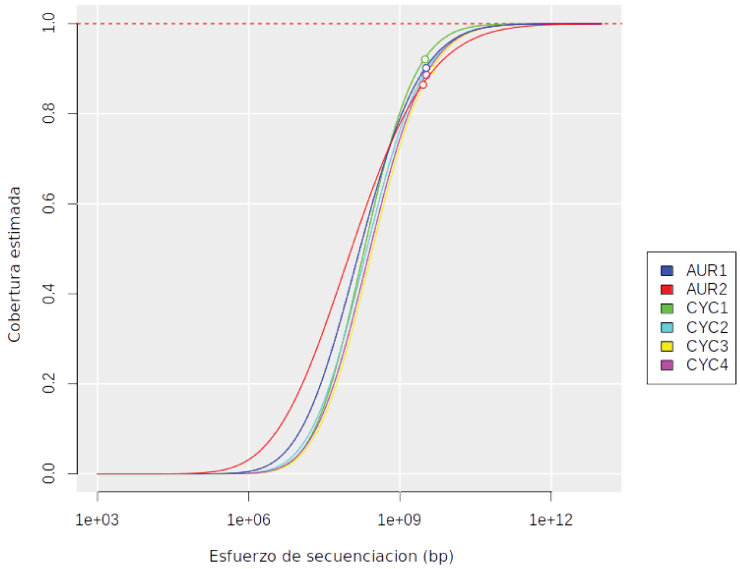

Figura 7.2: Estimación de cobertura de la secuenciación tipo *shotgun*. Cobertura estimada utilizando Nonpareil para cada muestra.

Las lecturas fueron clasificadas taxonómicamente utilizando KRAKEN2, contra la base de datos estándar. En promedio, el 22% de las lecturas de cada muestra pudieron ser clasificadas, siendo CYC1 la muestra con mayor porcentaje de lecturas clasificadas (~30%, Figura 7.3). En total, se identificaron 58 *phyla* distintas, pero 50 de ellas no representaban más del 1% de la muestra (figura 7.3). El *phylum* dominante en las muestras CYC fue *Proteobacteria*, representando el 70% de cada muestra en promedio. En cambio, en las muestras AUR no hubo un solo *phylum* dominante en ambas muestras. En AUR1, *Proteobacterias* (38%) y *Actinobacterias* (32%) fueron los dos grupos taxonómicos más representados, mientras que en AUR2 se vio una clara predominancia de *Actinobacterias* (62%). Otras taxa con una representación significativa en todas las muestras son *Firmicutes* (entre 2% y 12% de cada muestra) y *Bacteroidetes* (entre 2% y 7%). En todas las muestras se encontraron lecturas asignadas a linajes de *Archaea,* pero siempre en muy baja proporción: CYC4 fue la muestra con mayor porcentaje de *Archaea*, donde representaban el 1% del total. Así mismo, en todas las muestras hubo lecturas clasificadas como *Uroviricota*, un *phylum* de virus compuesto principalmente por bacteriofagos.

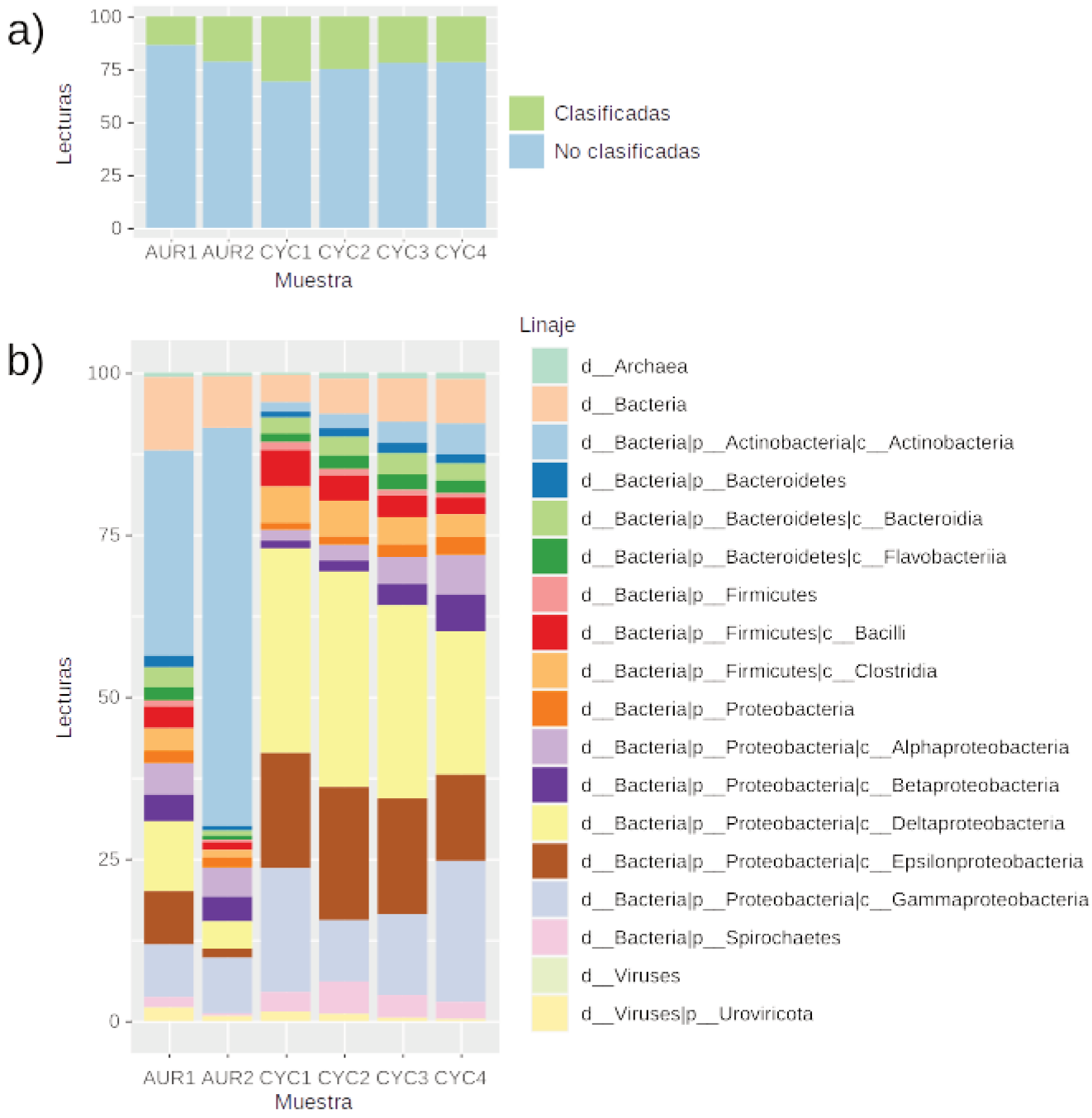

Figura 7.3: Clasificación taxonómica basada en lecturas. Porcentaje de lecturas asignadas por muestra a cada uno de los principales linajes encontrados, utilizando KRAKEN2 contra la base de datos estándar. a) Porcentaje de lecturas clasificadas. b) Perfil taxonómico de cada muestra. Los linajes que representan menos del 1% en todas las muestras no fueron considerados.

También se hizo una búsqueda contra la base de datos de plantas, para corroborar la presencia de algas en las lagunas. En todas las muestras se encontraron lecturas clasificadas dentro del *phylum Chlorophyta*, que incluye a las algas verdes, pero siempre se encontraron en proporciones poco significativas, rondando entre el 0.03% y el 0.08% del total de la muestra.

ENSAMBLADO Y RECONSTRUCCIÓN DE GENOMAS

El ensamblado de lecturas se realizó mediante una estrategia que permitiera el *binning* por cobertura diferencial. Por esto, se hizo un ensamblado combinando todas las lagunas correspondientes a un mismo sistema, ya que los resultados de diversidad beta del amplicón del gen 16S y la clasificación taxonómica de las lecturas de la secuenciación tipo *shotgun* sugieren que no hay diferencias significativas entre ellas. Esto permite considerar a cada laguna como una "muestra" dentro del sistema, lo que da mayor sustento estadístico al proceso de *binning* y permite una mejor reconstrucción de genomas.

Al ensamblar las lecturas se obtuvieron, en total, 238.725 *contigs* de más de 1 k (Tabla 7.3). Más del 60% de las lecturas de cada muestra mapeo contra el ensamblado correspondiente.

Tabla 7.3: Estadísticas del ensamblado y predicción de genes por sistema de lagunas.

Sistema	Contigs totales	≥ 1 kb	≥ 10 kb	≥ 100 kb	Mas largo (b)	N50	Genes predichos
AUR	461.955	91.676	3.304	25	313.235	1.430	764.778
CYC	845.042	147.049	6.051	103	458.380	1.413	1.369.301
Total	1.306.997	238.725	9.355	128	458.380	1.419	2.134.079

El *binning* de los contigs de más de 1 kb dio lugar a 264 *bins*, 95 para el set de datos AUR y 169 para CYC (Tabla 7.4; Tabla S3). Para ambos sets, el porcentaje de contigs asignados a *bins* es muy alto: 94.5% para AUR y 93% para CYC (figura 7.5). Después de la inspección manual, la calidad de 25 *bins* mejoró, eliminando *contigs* que mostraban un comportamiento diferente en su cobertura diferencial o en su porcentaje de GC. También se encontraron 3 *bins* quiméricos, en los que *contigs* con distintas características habían sido agrupados (Figura 7.4). Estos *bins* fueron divididos en dos, obteniendo 6 nuevos *bins.*

Tabla 7.4: Estadísticas del *binning* para cada muestra.

Sistema	*Bins* totales	MAGs totales	MAGs de alta calidad	MAGs en borrador
AUR	95	43	14	29
CYC	169	67	21	46
Total	264	110	35	75

Luego de todos los controles, se obtuvieron 35 MAGs de alta calidad, con más de 90% de completitud y baja contaminación. También se reconstruyeron 75 MAGs con más de 70% de completitud y menos de 20% de contaminación: estos genomas fueron clasificados como MAGs borradores.

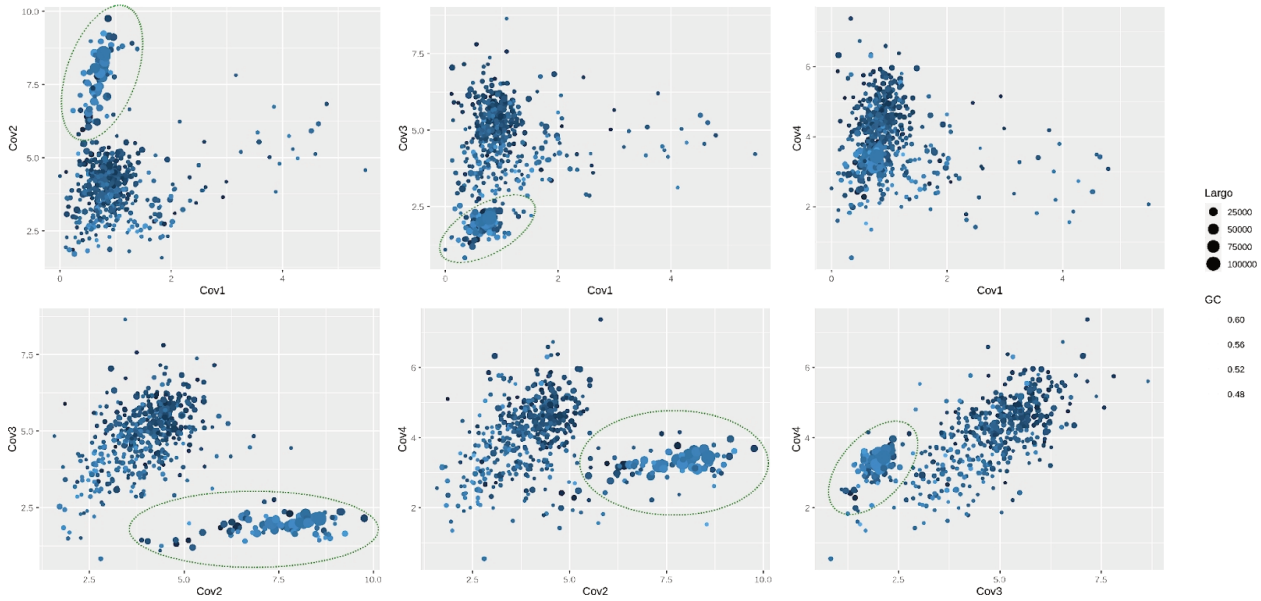

Figura 7.4: Ejemplo de *bin* quimérico. Gráfico de dispersión construido a partir de la cobertura de los *contigs* del *bin* INTA.CYC.085 en cada muestra. El tamaño de los puntos está definido en base al tamaño de los *contigs* y el color en relación al contenido de GC. En verde se señala un grupo de *contigs* que luego formarán el MAG INTA.CYC.1002.

La clasificación taxonómica obtenida con GTDBtk mostró que los MAGs estaban distribuidos en 16 *phyla* y 23 clases (Figura 7.6a; Tabla S4). Solo un MAG, INTA.CYC.1002, fue clasificado como *Archaea*, todos los demás MAGs fueron clasificados dentro del dominio *Bacteria*. El *phylum* con más MAGs asignados fue *Firmicutes_A*, con 17 genomas, seguido por *Patescibacteria* y *Bacteroidetes*, con 10 MAGs cada uno. En general, no se encontró diversidad dentro de un mismo *phylum:* para 12 de las 16 *phyla* encontradas solo se encontraron miembros de una clase. *Patescibacteria* resultó ser el más diverso, con miembros de 5 clases distintas.

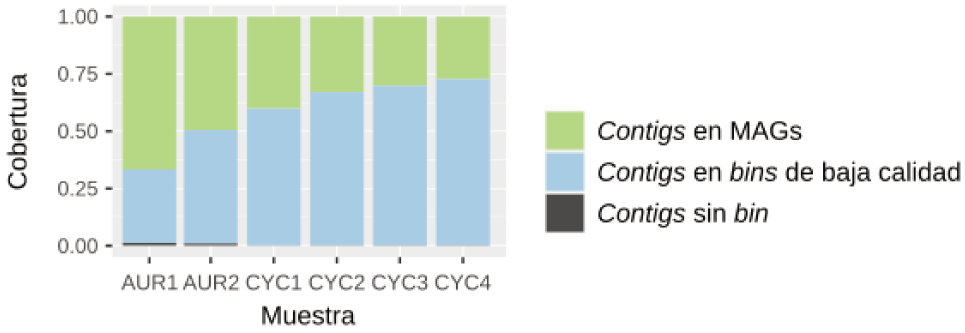

Figura 7.5: Proporción de reads en *contigs* mayores a 1000b ubicados en MAGs (verde), en *bins* de baja calidad (azul) y que no forman parte de un *bin* (negro).

Al analizar la cobertura de los MAGs (Figura 7.6b), se pudo observar que *Saccharimonadia*, miembro de *Patescibacteria*, fue el taxón dominante dentro de las lagunas CYC1 y CYC2, principalmente debido a la gran abundancia del MAG INTA.CYC.001 en esas muestras. Sin embargo, para las muestras CYC3 y CYC4, la clase más abundante pasó a ser *Bacteroidetes* (*phylum Bacteroidota*). Por su parte, en AUR1 *Bacteroidetes* fue nuevamente la taxa dominante, pero para AUR2 la clase más abundante fue *Actinobacteria*, con 6 veces más cobertura que la taxa siguiente.

El siguiente paso consistió en identificar que MAGs eran compartidos entre ambos sistemas, calculando la identidad promedio de nucleótidos (ANI) entre todos ellos. Solo 6 pares de MAGs mostraron un alto porcentaje de alineamiento (>50%) y porcentajes de identidad muy altos (>98%). La mitad de estos pares corresponden a la clase *Clostridia*, mientras que el resto fueron clasificados como *Chromatiales*, *Bacilli* y *Spirochaetes*.

Tabla 7.5: Identidad Promedio de Nucleótidos (ANI) entre MAGs.

MAG 1	MAG 2	Linaje de GTDB	% Identidad	% Alineado
INTA.AUR.063	INTA.CYC.029	c__Clostridia	99.93	89.61
INTA.AUR.026	INTA.CYC.020	c__Clostridia	99.98	86.58
INTA.AUR.070	INTA.CYC.166	o__Chromatiales	99.97	83.48
INTA.AUR.027	INTA.CYC.017	o__Sphaerochaetales	99.72	74.78
INTA.AUR.040	INTA.CYC.052	c__Bacilli	99.97	74.62
INTA.AUR.036	INTA.CYC.091	c__Clostridia	98.20	68.56

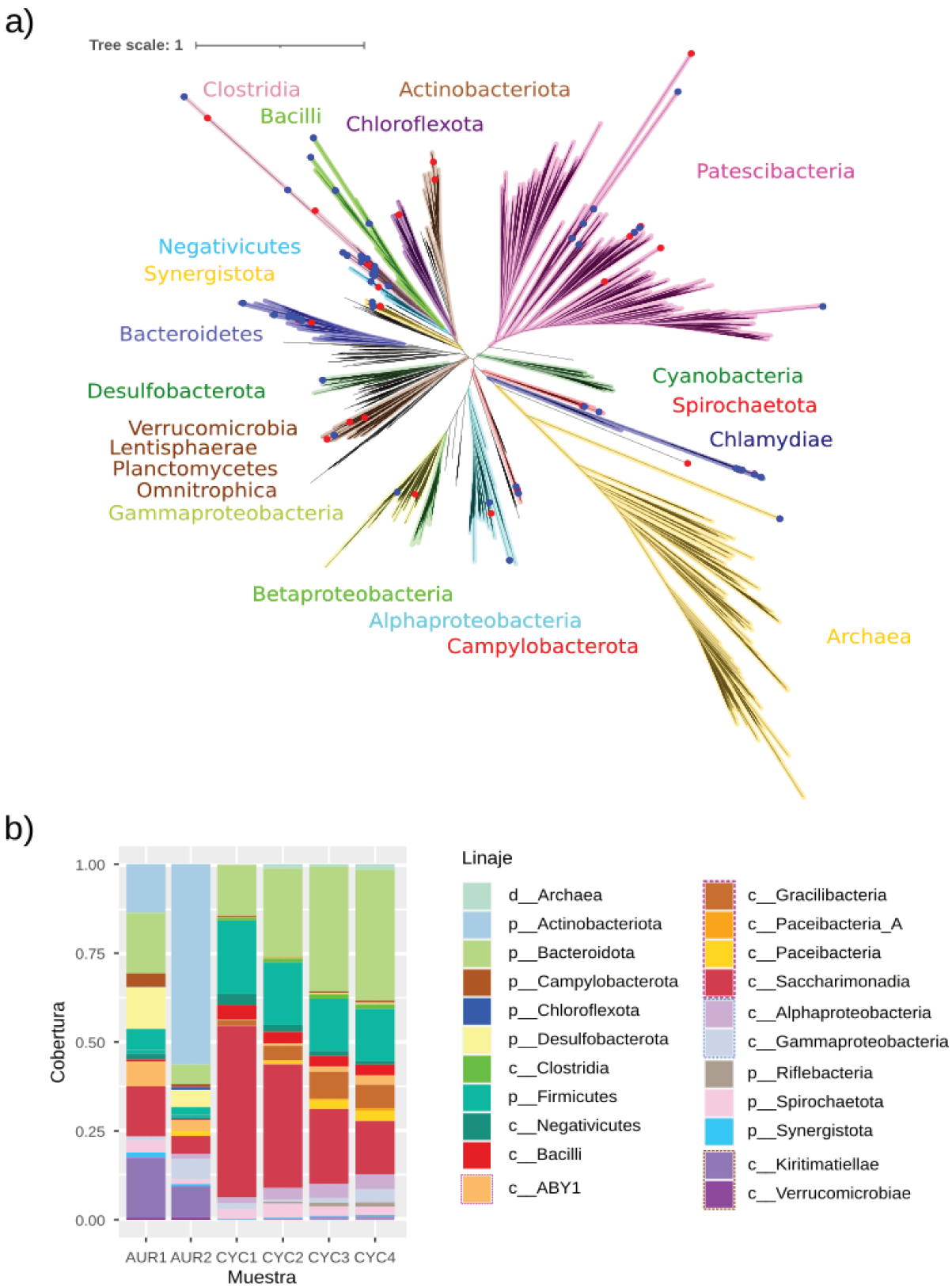

Figura 7.6: Clasificación taxonómica basada en MAGs. Perfiles taxonómicos de las muestras obtenidos comparando los MAGs de alta calidad y borradores contra la base de datos GTDB. a) Árbol filogenómico construido a partir de 16 genes ribosomales (adaptado de Hug y col., 2016) Los puntos azules corresponden a los MAGs de CYC, mientras que los puntos rojos corresponden a los MAGs de AUR. b) Abundancia de cada grupo taxonómico identificado. La altura de cada barra indica la proporción de lecturas asignadas a cada MAG. Las *taxa* dentro de la línea punteada rosa corresponden al *phylum Patescibacteria*, la línea punteada azul al *phylum Proteobacteria* y la linea punteada marrón al *phylum Verrucomicrobia*.

Más de 2 millones de genes fueron predichos en ambos sistemas, aunque CYC cuenta con casi el doble de genes predichos que AUR (Tabla 7.3). Los MAGs, que fueron construidos solo con *contigs* de largo ≥1kb, representan sólo el 15% del total de genes predichos. Por ello, se consideraron todos los genes para la anotación funcional.

Usando las bases de datos CAZy, MEROPS y KEGG, se anotaron alrededor del 40% de los genes en MAGs. Por su parte, menos del 15% de los genes por fuera de los MAGs pudieron ser anotados.

La Figura 7.6 resume las principales rutas metabólicas analizadas. Estas corresponden a los procesos metabólicos más reportados en lagunas facultativas, que incluyen la hidrólisis de macromoléculas y la degradación de los monómeros obtenidos, siguiendo principalmente los denominados ciclos del carbono y del nitrógeno.

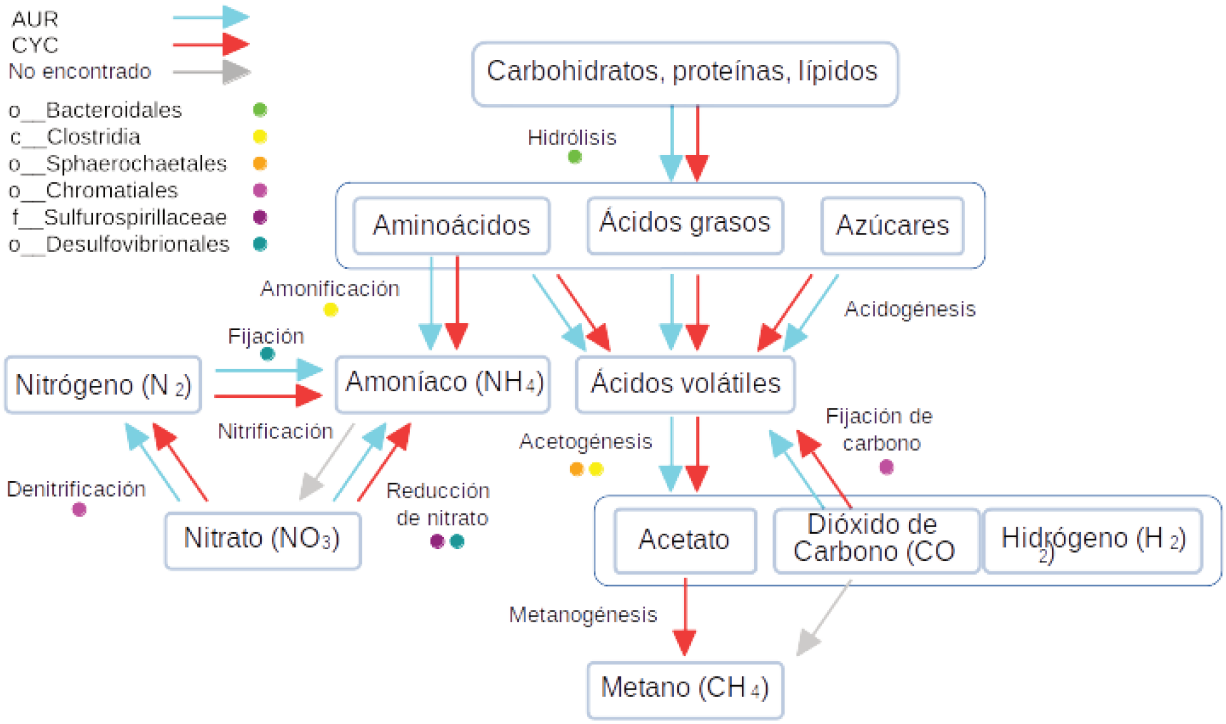

Figura 7.6: Principales rutas metabólicas en lagunas de estabilización facultativas. Esquema de las principales rutas detalladas en los sistemas analizados. Los genes predichos en cada sistema fueron comparados contras las bases de datos CAZy, MEROPS y KEGG. Las flechas indican si el proceso metabólico fue hallado en cada sistema, y los puntos coloreados indican la participación de algún organismo clave en ese proceso.

Para identificar enzimas que participen en la hidrólisis de macromoléculas, se buscaron enzimas degradadoras de carbohidratos en CAZy (CAZymas) y proteasas en MEROPS. En total, se identificaron 21.196 CAZymas, siendo las glicosil transferasas (GT) el grupo más abundante (47,4% del total). El segundo grupo más abundante fue el de las glicosil hidrolasas (GH), contabilizando el 38.3% de las secuencias anotadas, seguido por las carbohidrato esterasas (CE, 12,7%) y las

polisacárido liasas (PL, 1,6%). Un análisis más detallado de las CAZymas identificadas se realiza en el capítulo 8.

Por su parte, comparando con MEROPS, se identificaron 24.145 proteasas putativas distribuidas en 9 clases distintas. Dado que la clasificación de este tipo de enzimas depende principalmente de su residuo catalítico o de la presencia de iones metálicos, no es posible hacer un análisis funcional directo a partir de estos resultados. En el capítulo 9 se profundiza más sobre este tipo de enzimas.

En primer lugar se analizó el ciclo del nitrógeno. En muchos MAGs (63/110) se encontraron 1 o más enzimas o rutas para obtener amoníaco a partir de aminoácidos, uno de los compuestos del ciclo de nitrógeno. Para la etapa de nitrificación, la única enzima que no se encontró en ninguno de los dos sistemas es la encargada de catalizar el primer paso (metano/amonio monooxigenasa). Para la etapa de reducción desasimilatoria de nitrato, el módulo completo de KEGG (M00530) fue encontrado en 14 MAGs, pero en ninguno se encontró el módulo para la reducción asimilatoria de nitrato. Por otro lado, en 4 MAGs se encontró completo el módulo para la denitrificación de nitrato (M00529) y en otros 4 MAGs el módulo completo para la fijación de nitrógeno (M00175).

Luego, el análisis se centró en los genes relacionados al metabolismo del carbono, comenzando por la hidrólisis de polisacáridos. Si bien todos los MAGs presentaron CAZymas, los organismos clasificados como *Bacteroidetes* son los que mostraron mayor cantidad y variedad de secuencias asignadas a las familias GH, CE y PL, relacionadas con la degradación de polisacáridos y la liberación de monosacáridos.

Los monómeros son convertidos en ácidos grasos volátiles en la etapa denominada acidogénesis, que fue identificada en la mayoría de los MAGs. La segunda etapa del ciclo del carbono, la acetogénesis, puede ser dividida en dos partes: la conversión de piruvato en acetil-CoA (M00307) y la conversión de acetil-CoA en acetato (mediada por las enzimas K01026, K01067, K01905+K22224, K18118 o K24012). La primer parte de esta etapa fue encontrada en 41 MAGs, pero en tan solo 7 se encontró además alguno de los genes necesarios para la segunda mitad.

Para la etapa de metanogénesis hay descriptos 4 módulos en KEGG, dependiendo del compuesto de partida: metanol (M00356), acetato (M00357), metilamina (M00563) y dióxido de carbono y agua (M00567). Ninguno de los módulos fue encontrado completo. Sin embargo, evidencia de la metanogénesis a partir de metanol fue encontrado en INTA.CYC.1002, el único MAG clasificado como *Archaea*, y algunos genes de la metanogénesis a partir de dióxido de carbono y agua, y de la metanogénesis a partir de acetato fueron encontrados entre los genes de CYC fuera de los MAGs.

La última parte del ciclo de carbono es la fijación de dióxido de carbono. Una de las vías de fijación presente en bacterias es el Ciclo de Calvin, un sistema que puede dividirse en dos etapas: la primer mitad, que incluye al gen RuBisCo, está descrita en el módulo M00166 y la segunda mitad en el

módulo M00167. En total, 4 MAGs presentaron completa la primer etapa, pero ninguno poseía completa la segunda, dos en cada set de datos. Ambos genomas de AUR presentaron 4 de los 7 genes que constituyen esta segunda etapa, mientras que los dos genomas de CYC presentaron 5 de los 7 genes. Interesantemente, estos mismos 4 MAGs fueron los únicos que en los que se encontró completo un módulo de fotorreceptores (fotosistema anoxigénico II, M00597). Otras formas de fijación de carbono fueron buscadas, como el ciclo de Arnon-Buchanan o la ruta de Wood-Ljungdahl, pero la cantidad de genes encontrados para estos sistemas fue muy baja para considerarlos completos.

DISCUSIÓN

Las industrias lácteas, como pequeñas queserías, tratan sus efluentes en lagunas de estabilización facultativas. Las comunidades microbianas presentes en este tipo de lagunas se encargan de degradar los restos de materia orgánica presente en los efluentes, combinando procesos aeróbicos y anaeróbicos. El análisis presentado aquí se basa en la reconstrucción de genomas completos y la identificación de rutas metabólicas completas, en lugar de genes marcadores.

En primer lugar, se hizo un análisis basado en el gen marcador 16S, para estimar el nivel de complejidad del sistema, para determinar la variabilidad entre lagunas del mismo sistema y además la variabilidad entre ambos sistemas. La distancia entre muestras de una misma laguna fue menor que la distancia entre muestras de distintas lagunas. Por ello, para la secuenciación mediante estrategia de *shotgun*, se optó por secuenciar un pool de las tres réplicas de cada laguna, considerando cada laguna como una muestra dentro del sistema. Esto permitió obtener una mayor cobertura en cada muestra individual. Por otro lado, tanto la distancia de UniFrac entre muestras como la clasificación taxonómica de las lecturas de WGS mostraron que las comunidades en cada laguna de un mismo sistema eran similares. Este era un resultado esperado, ya que las lagunas están conectadas entre sí, permitiendo el paso de agua y de microorganismos. De aquí que se decidió realizar un co-ensamblado de todas las lecturas de un mismo sistema, lo que significó aumentar la cobertura de los microorganismos compartidos para el ensamblado. Además, permitió, al contar con distintas muestras dentro de un ensamblado, implementar una estrategia de binning que utilice la cobertura diferencial.

En total se reconstruyeron 110 MAGs, 35 de alta calidad (>90% de completitud; <5% de contaminación) y 75 borradores (>70% de completitud; <20% de contaminación). Cabe destacar que el incremento en el número de estudios metagenómicos ha sido muy reciente. Las muestras analizadas en el portal de datos metagenómicos del *European Bioinformatics Institute* (EBI-EMBL) ha visto el un incremento de 10 veces entre los años 2015 y 2017, alcanzando las 75.000 muestras (Mitchell y col., 2018); en 2019, el número de muestras ascendía a 140.000 (Ayling y col., 2020) y hoy supera las 300.000. Así mismo, el número de MAGs también vio un incremento en los últimos años. En 2015, la base de datos GOLD (*Genome OnLine Database*) contaba con alrededor de 4.600 genomas provenientes de metagenomas (Reddy y col., 2015) y en ese año se publicó el primer estudio a gran

escala, donde se ensamblaron más de 1.750 genomas (Brown y col., 2015). Sin embargo, hoy en día es posible encontrar grandes estudios que reportan miles de genomas: Youngblut y col. (2020) que obtuvieron más de 5.000 genomas de intestino de animales salvajes y, por su parte, Chen y col. (2021) obtuvieron más de 6.000 genomas del tracto digestivo de cerdos domésticos.

La caracterización taxonómica de las comunidades se hizo utilizando dos aproximaciones: por una lado la clasificación de las lecturas sin ensamblar, utilizando KRAKEN2, y por otro la clasificación filogenómica de los MAGs reconstruidos, con la herramienta GTDB-tk.

Los dos MAGs más abundantes en las lagunas AUR son INTA.AUR.005 e INTA.AUR.009: ambos fueron clasificados como *Actinobacterias* y ambos muestran un marcado incremento en su abundancia en la segunda laguna. Este aumento en la abundancia del *phylum Actinobacteria* también se observó en los resultados de KRAKEN2. La presencia de miembros de este *phylum* en barros activados ha sido relacionada a problemas operacionales en estos sistemas (Seviour y col., 2008). Por lo tanto, se podría suponer que la alta abundancia de estos dos MAGs y los altos valores de DQO y DBO estarían relacionados. Esta coincidencia entre KRAKEN2 y GTDB-tk no se observó en las lagunas CYC. En estas, el MAG más abundante fue INTA.CYC.001, clasificado como *Saccharimonadia* (anteriormente denominado TM7). Sin embargo, esta clase, y el *phylum Patescibacteria* en general, están casi ausentes de las lagunas de acuerdo con los resultados de KRAKEN2. Dado que la base de datos para KRAKEN2 se basa en las secuencias disponibles en la base de datos de genomas de NCBI, estos grupos descritos recientemente pueden estar subrepresentados, y esto dificultaría su identificación.

Es importante notar que el porcentaje de lecturas clasificadas por KRAKEN2 es bajo (entre el 13,9% y el 30,8%), es por ello que estos resultados pueden ser útiles para tener una estimación de la variabilidad taxonómica entre muestras, pero no sería prudente trazar conclusiones fuertes a partir de ellos. Se consideró que una aproximación basada en reconstrucciones de genomas y su clasificación filogenómica, con herramientas como GTDB-tk, es una estrategia superadora para determinar la diversidad microbiológica de un ambiente (Strous y col., 2012, Sczyrba y col., 2017).

Luego del análisis taxonómico, se realizó una caracterización funcional de cada sistema de lagunas. La predicción de genes arroja dos tipos de secuencias, no parciales y parciales, dependiendo si el algoritmo pudo identificar un codón de inicio y uno de fin en el gen. Entre ambos conjuntos de datos, se predijeron un total de 2.134.079 genes, de los cuales solo 558.265 (26,2%) son genes no parciales. Dado que una gran proporción de los genes predichos eran parciales y no se cuenta con la secuencia total del gen, puede ocurrir que no se tenga suficiente información para anotarlo. Es por ello que los porcentajes de anotación de genes parciales (12,5%) y no parciales (40%) difieren tanto. Una solución para limitar el número de genes parciales sería imponer un umbral fijo (por ejemplo, genes en *contigs* ≥1kb). Sin embargo, de esa forma se pierden secuencias completas y no se elimina el problema de las secuencias parciales. Dado que el principal objetivo de este capítulo es describir las rutas metabólicas

que pudieran ocurrir en estos ambientes y que el número de genes no parciales es bajo (26% del número total), se consideraron todos los genes predichos a la hora de reconstruir rutas metabólicas.

La anotación funcional de los genes fue basada en módulos de KEGG, que se definen como una colección de genes dentro de una ruta metabólica, definidos manualmente (Kanehisa y col., 2021). Los dos ciclos metabólicos de mayor importancia dentro de una laguna de estabilización son el ciclo de carbono y del nitrógeno, pero también resultan de interés el metabolismo del azufre, ya que el sulfato puede ser usado como aceptor de electrones en condiciones anaeróbicas, y la formación de compuestos de reserva como el ácido polihidroxibutírico (PHB, Grady y col., 2011).

Tanto el ciclo del carbono como el del nitrógeno comienzan con la degradación de polímeros (proteínas, polisacáridos) en monómeros (aminoácidos, monosacáridos). Los *Bacteroidetes* son un grupo de organismos que han sido encontrados en distintos tipos de reactores y se ha propuesto que poseen la capacidad de degradar compuestos orgánicos en monómeros (Sun y col., 2015). Esta clase es la segunda más representada dentro de ambos conjuntos de datos, con un total de 12 MAGs, superada sólo por *Clostridia* (13). La característica distintiva de los genomas reconstruidos para esta clase es que presentan un gran número de CAZymas. Los MAGs INTA.CYC.034, INTA.CYC.039, INTA.CYC.074 e INTA.CYC.095 son los que tienen más GHs en ambos sistemas, lo cual sugiere que estarían adaptados para incorporar carbono del ambiente, presente principalmente en forma de lactosa. Por ello, se propone que los *Bacteroidetes* serían miembros claves de la comunidad microbiana en lagunas de estabilización (tabla 7.6)

Tabla 7.6: Especies claves en lagunas facultativas. Organismos observados en ambas lagunas que poseen completo módulos para alguno de los procesos metabólicos buscados en las lagunas facultativas.

Taxón	# MAGs en AUR	# MAGs en CYC	Rol en lagunas facultativas
Bacteroidales	5	12	Degradación de polímeros
Clostridia	8	18	Degradación de aminoácidos, acetogénesis
Sphaerochaetales	3	3	Acetogenesis
Chromatiales	2	2	Fotosíntesis, fijación de carbono, reducción de sulfato a sulfuro, denitrificación
Sulfurospirillaceae	1	1	Reducción de nitrato a amoníaco.
Desulfovibrionales	2	1	Reducción de nitrato a amoníaco, reducción de sulfato a sulfuro, fijación de nitrógeno
Saccharimonadales	3	4	Desconocida

El ciclo del carbono puede dividirse en 2 etapas, la degradación de monómeros en dióxido de carbono o metano y la fijación de dióxido de carbono para obtener ácidos volátiles. A su vez, la degradación de monómeros puede dividirse en 3 pasos: acidogénesis (formación de ácidos volátiles), acetogénesis (formación de acetato) y metanogénesis (formación de metano).

Virtualmente, todos los MAGs poseen algún módulo o gen relacionado con la obtención de piruvato a partir de monómeros, como la glucólisis, que es la principal forma de obtención de energía en las células. Por otro lado, solo pocos MAGs poseen las enzimas necesarias para la acetogénesis. Muchos MAGs presentan genes para la oxidación del piruvato en acetil-CoA y algunos las enzimas necesarias para convertir el acetil-CoA en acetato, pero solo 7 MAGs poseen ambos. La mayoría de estos genomas fueron clasificados como *Clostridia*, aunque también se encontraron miembros de la clase *Sphaerochaetales* (perteneciente al *phylum Spirochaetes*), que han sido previamente reportados capaces de oxidar piruvato (Narihiro y col., 2015). Dado que ambos grupos de organismos se encontraron en ambos conjuntos de datos, incluyendo 4 de las especies compartidas entre ambas lagunas, se propone que ambos serían miembros claves de las lagunas facultativas. Es importante notar que la cantidad de genomas de *Sphaerochaetales* reconstruidos es menor que los de *Clostridia* y en los primeros no se encontraron todas las enzimas para la conversión del acetil-CoA en acetato, por lo cual su rol sería menor que el de las *Clostridia*.

Como fue planteado por Grady y col. (2011), puede ser considerado como positivo que las lagunas de estabilización produzcan metano, ya que puede usarse como fuente de energía. La metanogénesis ha sido solo descrita en *Archaea* (Buan, 2018). El número de lecturas clasificadas en este dominio fue siempre menor al 1%, salvo para la muestra CYC4, donde fue del 1,03%, y se logró reconstruir un solo MAG arqueal: INTA.CYC.1002. Este genoma presenta algunos genes relacionados a la metanogénesis a partir de metanol (M00356) y a partir de metilamina (M00563). Por otro lado, dentro de los genes de CYC no asignados a un *bin* se encontraron casi todos los genes involucrados en la formación de metano a partir de acetato y a partir de dióxido de carbono. Estos resultados sugieren que la comunidad de CYC sería capaz de producir metano, a diferencia de la comunidad de AUR. Sin embargo, dado el diseño de las lagunas facultativas, sería imposible recolectar el metano producido y este sería liberado al ambiente. Es por ello, que en un contexto de calentamiento global, la inhabilidad de una comunidad de producir metano no sería un aspecto negativo.

La etapa restante dentro del ciclo del carbono es la fijación de dióxido de carbono. De los 6 sistemas principales utilizados por autótrofos (Montoya y col., 2012), se encontraron genes de 4 de ellos. Si bien fueron encontrados genes del ciclo de Arnon-Buchanan, de la ruta de Wood-Ljungdahl y del ciclo del dicarboxilato-hidroxibutirato, ninguna de estas rutas estaba completa. El ciclo de Calvin se encontró casi completo en 4 MAGs (INTA.AUR.070, INTA.AUR,082, INTA.CYC.166, e INTA.CYC.169), incluidas las dos subunidades de la RuBisCo. Estos 4 MAGs fueron los únicos en también tener un módulo fotosintético completo y ser capaces de oxidar tiosulfato. Ha sido reportado que el tiosulfato puede ser

usado como dador de electrones durante la fotosíntesis (Falkenby y col., 2011). Además, estos organismos fueron los únicos en presentar los 3 genes, más el amplificador, para la producción de PHB como reserva de carbono. Todos estos organismos fueron clasificados dentro del orden de los *Chromatiales*, teniendo dos de ellos (INTA.AUR.070 e INTA.CYC.166) una ANI mayor al 99,9%. Dado su contribución crucial en la fijación del carbono y su conservación en ambos sistemas, los *Chromatiales* también son propuestos como organismos claves dentro de las lagunas.

El segundo ciclo importante en este tipo de ambiente es el ciclo del nitrógeno. Para poder comenzar el ciclo es necesario primero degradar los aminoácidos para obtener amoníaco. Luego, ese amoníaco puede ser convertido en nitrato en una etapa denominada nitrificación. Ese nitrato puede ser reducido nuevamente a amoníaco o convertido en nitrógeno gaseoso, mediante el proceso de desnitrificación. El último paso del ciclo comprende la fijación del nitrógeno gaseoso en amoníaco.

Como se indica en la tabla suplementaria S2, en KEGG hay numerosos genes y módulos para la transformación de distintos aminoácidos en amoníaco. Si bien, una gran proporción de MAGs (63/110) poseen al menos uno de estos mecanismos, los miembros de la clase *Clostridia* son los que muestran mayor cantidad. Como se comentó anteriormente, estos organismos serían claves para los sistemas facultativos por ser capaces de convertir ácidos volátiles en acetato, pero también por su habilidad de fermentar aminoácidos, como ya ha sido reportado (Militon y col., 2015).

El primer paso de la etapa de nitrificación es llevado a cabo por la enzima metano/amonio monooxigenasa (K10944, K10945, y K10946). Esta enzima no se encontró en ninguno de los 2 sistemas, por lo que esta etapa no podría llevarse a cabo en las lagunas estudiadas. Los microorganismos capaces de oxidar amoníaco pertenecen a un limitado grupo taxonómico (Arp y col., 2007; Zhou y col., 2015) y, aunque estos grupos fueron encontrados en la caracterización hecha con KRAKEN2, el número de lecturas es muy bajo (0.0001%).

El módulo para la fijación de nitrógeno fue encontrado completo en 4 MAGs. Es importante notar que otros 6 MAGs poseen el gen *nifH*, que se utiliza como marcador para identificar microorganismos fijadores de nitrógeno (Zehr y col., 2003), pero no poseen los otros genes del módulo. Es por ello que limitar la búsqueda únicamente a genes marcadores podría resultar insuficiente, y es más apropiada la búsqueda basada en rutas metabólicas o *cassettes* de genes completos cuando se tienen datos de secuenciación de genoma completo.

El metabolismo de nitrato puede seguir dos caminos. La desnitrificación, que permite la obtención de nitrógeno gaseoso, solo se encontró completo en un MAG (INTA.CYC.169, perteneciente al orden de los *Chromatiales*) y faltando un gen en otro (INTA.AUR.082, también perteneciente a los *Chromatiales*). La otra forma de metabolismo del nitrato es su reducción a amoníaco. En ambos sistemas se encontraron MAGs con el módulo completo para la reducción disimilatoria (M00530), Estos MAGs pertenecen principalmente a *Bacteroidetes* y a la familia *Sulfurospirillaceae*, con

representantes de ambas taxa en ambas lagunas. Dado que solo 4 *Bacteroidetes* de los 17 reconstruidos poseen estos genes, la reducción disimilatoria no sería una ruta relacionada al linaje completo, sino solo a algunos de sus miembros. Por su parte, los *Sulfurospirillaceae* no están descritos en la base de datos de taxonomía del NCBI (es propuesta por la base de datos GTDB), en cambio sus miembros aparecen como miembros de la clase *Epsilonproteobacteria*, uno de los linajes más abundantes de acuerdo a los resultados KRAKEN2 en las lagunas de CYC. Dada su alta abundancia según el análisis taxonómico y que los dos genomas reconstruidos son capaces de reducir el nitrato, este grupo de organismos también sería clave para las lagunas facultativas. Otro linaje muy abundante según KRAKEN2 y con el módulo de la reducción disimilatoria presente, son los *Desulfovibrionales*. Si bien en ninguno de los 3 MAGs asignados a este orden se encontró el módulo completo, en todos los casos fue por ausencia de una subunidad de algún gen y no por la ausencia de un paso completo. Dada su presencia en ambas lagunas, su alta abundancia según KRAKEN2 y su participación en este paso del ciclo, también serían organismos claves, pero en menor medida que la familia *Sulfurospirillaceae*.

Se encontraron 26 MAGs con genomas pequeños (<2Mb). Muchos de ellos pertenecen al recientemente reportado *Candidate Phyla Radiation* (CPR, Brown y col., 2015) o *Patescibacteria* (Rinke y col., 2013; Parks y col., 2017), pero también se encontraron otras *phyla*, como *Firmicutes* (7 MAGs con <2Mb). Como ha sido reportado (Nelson y Stegen, 2015), los miembros del *phylum Patescibacteria* poseen un metabolismo muy limitado, con pocas rutas metabólicas relacionadas a la síntesis de lípidos y de ácidos nucleicos. Así mismo, tampoco se encontraron genes del ciclo del ácido cítrico (Wrighton y col., 2012; Castelle y col., 2017), pero sí se encontraron partes para la obtención de energía a partir de compuestos de 3 carbonos y, en el caso de INTA.CYC.027 (el único MAG clasificado en la clase *Gracillibacteria*), enzimas relacionadas al metabolismo de ribosa. De las 5 clases halladas, 3 se encuentran en ambas lagunas. Llamativamente, la clase *Saccharimonadia* es uno de los linajes más abundantes en las muestras de CYC, representando más del 48% de la muestra CYC1, siendo INTA.CYC.001 el MAG más abundante en la primer laguna del sistema. La presencia de este tipo de organismos se ha reportado en otros bioreactores en proporciones similares (Remmas y col., 2017), lo cual indica que serían de importancia en las lagunas, pero su limitado genoma, la falta de aislamientos y el poco conocimiento bioquímico del metabolismo de este linaje dificulta la elaboración de una hipótesis sobre su rol.

Finalmente, existen otras taxa compartidas entre ambos sistemas, pero en baja abundancia. En ambos conjuntos de datos se encontraron organismos que han sido reportados como degradadores de compuestos orgánicos, como es el caso de *Verrucomicrobia* (Cardman y col., 2014), *Synergistetes* (Militon y col., 2015) y *Erysipelotrichales* (Seyedi y col., 2020). Sin embargo, ninguno de los MAGs reconstruidos poseen completo alguna ruta de interés. Por ello, consideramos que la presencia de estos no sería tan crucial como la de los grupos previamente descritos (Tabla 3).

CONCLUSIONES

Se realizó un análisis de las comunidades microbianas en las lagunas, utilizando una estrategia centrada en la reconstrucción de genomas completos y la identificación de rutas metabólicas en ellos. Este tipo de análisis permite un entendimiento más profundo de los procesos metabólicos en un ambiente que el que permiten los estudios centrados en marcadores moleculares, ya sean funcionales o taxonómicos.

Como resultados, se reconstruyeron 110 MAGs y se identificaron todos los genes y rutas metabólicas relacionadas a los principales procesos que se llevan a cabo en lagunas facultativas. Se identificaron 7 linajes que serían claves en la degradación de la materia orgánica en lagunas de estabilización facultativas. Un mejor entendimiento de los microorganismos presentes, su potencial metabólico y cómo interactúan entre sí es necesario para identificar más actores de importancia y permitiría mejorar el funcionamiento de este tipo de sistemas, tan utilizados en países en desarrollo, y minimizar el impacto ambiental de los residuos arrojados en ellos.

IDENTIFICACIÓN, CLONADO Y EXPRESIÓN DE B-GALACTOSIDASAS

8 IDENTIFICACIÓN, CLONADO Y EXPRESIÓN DE B-GALACTOSIDASAS

INTRODUCCIÓN

Las β-galactosidasas pertenecen a la familia de las glicosil hidrolasas (GH, EC 3.2.1.-), enzimas que rompen de manera específica, enlaces glicosídicos en carbohidratos (Davies y Henrissat, 1995). En el caso de la lactosa, esta hidrólisis resulta en la liberación de un grupo galactosil y un grupo glucosil. Si el aceptor del grupo galactosil es una molécula de agua, el resultado de la reacción es la hidrólisis de la lactosa; sin embargo, si el grupo aceptor es otro sacárido, el resultado es la síntesis de GOS en una reacción denominada de transgalactosidación (Gosling y col., 2010). Muchos estudios se han centrado en el mecanismo hidrolítico de esta familia de enzimas (Mahoney, 1998), y de las GH en general (Vocadlo y Davies, 2008), pero poca información hay disponible sobre su capacidad de transgalactolizar.

Dada la gran diversidad de carbohidratos, existe una enorme variedad de enzimas capaces de metabolizarlos. La base de datos CAZy (Lombard y col., 2014) provee acceso a secuencias e información actualizada de las glicosil hidrolasas, clasificadas en familias de acuerdo a la identidad de su secuencia (Henrissat, 1991; Henrissat y Bairoch, 1996). Sin embargo, esta clasificación posee sus limitaciones, ya que los miembros de una familia pueden tener actividades muy diferentes (diferentes sustratos o diferentes transformaciones químicas), y una misma actividad puede ser llevada a cabo por enzimas de distintas familias (Lombard y col., 2014). Por ejemplo, según CAZy, para la familia de las GH1 hay más de 45,000 secuencias reportadas, con más de 20 actividades distintas. Por otro lado, la actividad β-galactosidasa (EC:3.2.1.23) fue reportada en miembros de 5 familias distintas: GH1, GH2, GH35, GH42 y GH59.

La cantidad de secuencias disponible para cada una de esta familia es muy dispar (Tabla 8.1). La familia GH1 es la más representada en CAZy con más de 45.000 secuencias, seguida por GH2 con más de 26.000. Estas dos familias también son las que más estructuras disponibles tienen. A partir del análisis de estas secuencias y estructuras disponibles, para GH2, se ha propuesto que existen 3 dominios que caracterizan a todos los miembros de la familia, pero que diferentes combinaciones de otros dominios adicionales en el extremo C terminal de la enzima pueden contribuir a la diferenciación funcional (Talens-Perales y col., 2016). No hay reportes de este estilo para ninguna de las otras 4 familias con actividad β-galactosidasa.

En el siguiente capítulo se describe la identificación, selección, clonado y expresión exitosa de β-galactosidasas a partir de ADN metagenómico, y su caracterización funcional.

Tabla 8.1: Familias con actividad β-galactosidasa reportadas en CAZy. Número de secuencias y actividades reportadas para cada familia con actividad β-galactosidasa. Para cada familia se indica la cantidad de secuencias discriminadas por origen, *Bacteria*, *Archaea* y *Eucariota*, el número de enzimas cuya actividad fue caracterizada y el número de estructuras disponibles.

	GH1	GH2	GH35	GH42	GH59
Actividades reportadas	24	11	4	2	2
Secuencias totales	45.425	26.162	5.301	6.122	262
Bacteria (origen)	43.028	25.378	4.265	6.089	231
Archaea (origen)	175	91	25	24	0
Eucariota (origen)	2.162	661	1.007	8	31
Caracterizadas	326	180	75	63	5
Estructuras	76	55	16	11	1
Disponible comercialmente	SI	SI	SI	NO	NO

A partir de los ensamblados y los genes predichos obtenidos en el capítulo anterior, se realizó un análisis en profundidad de las CAZymas en general, y de las β-galactosidasas en particular. Este análisis permitió la selección de un grupo de genes candidatos que fueron clonados, expresados y caracterizados funcionalmente.

IDENTIFICACIÓN DE CAZYMAS

Como se describió en el capítulo anterior, de un total de 2.134.079 genes predichos, 21.196 fueron clasificados como CAZymas. Se encontraron candidatos para los 4 principales tipos de enzimas descritos en CAZy (Tabla 8.2), siendo GT (glicosil transferasas) el más abundante, y GH el tipo de *CAZyma* más diverso, con 107 familias distintas identificadas.

De las 221 familias identificadas, 159 (>70%) estaban presentes en ambos sistemas. Para las GH, que son el tipo de enzima de interés, de las 107 familias totales identificadas, 86 (>80%) se encontraron en ambos metagenomas. Las familias de GH que estaban en un metagenoma pero no en el otro se encontraron en cantidades muy bajas, entre 1 y 3 secuencias por familia.

Tabla 8.2 Predicción de CAZymas en ambos metagenomas. Clasificación de las CAZymas putativas identificadas en cada muestra. GH: Glicosil hidrolasas; GT: Glicosil transferasas; PL: polisacárido liasas; CE: carbohidrato esterasas

	AUR		CYC		Total	
Tipo	Genes	Familias	Genes	Familias	Genes	Familias
GH	2.970	93	5.156	100	8.126	107
GT	3.896	46	6.132	60	10.028	66
PL	104	13	236	17	340	17
CE	940	15	1.762	15	2.702	15

Dentro de GH, sólo hay reportadas 5 familias con actividad β-galactosidasa: GH1, GH2, GH35, GH42 y GH59. De estas, GH2 fue la más abundante entre ambos metagenomas, con 183 genes candidatos, seguida por GH1, con 142 secuencias, y GH42, con 54 genes (Tabla 8.3). Dado que el número de secuencias no parciales para las familias GH35 y GH59 fue muy bajo, se decidió no incluirlas en para los análisis posteriores y para la expresión heteróloga.

Tabla 8.3 Número de secuencias para las familias con actividad β-galactosidasa. Clasificación de las CAZymas

Familia	Cantidad de genes	Genes no parciales
GH1	142	121
GH2	183	161
GH35	18	6
GH42	54	35
GH59	1	0

Para analizar la diversidad taxonómica de las enzimas se utilizó el linaje asignado al MAG al cual pertenecían o, para aquellas secuencias que no se encontraban en algún MAG, se las clasificó utilizando el mejor *hit* de BLAST contra la base de datos nr. Hubo 10 secuencias para las cuales no se encontró ningún *hit* y quedaron como “no clasificadas” (Figura 8.1). Si bien se encontraron GH en 14 *phyla* distintas, más de la mitad de las secuencias pertenecen a *Bacteroidetes* y *Firmicutes*. Es interesante remarcar que las secuencias clasificadas como *Bacteroidetes* pertenecen casi exclusivamente a la familia GH2 (92%).

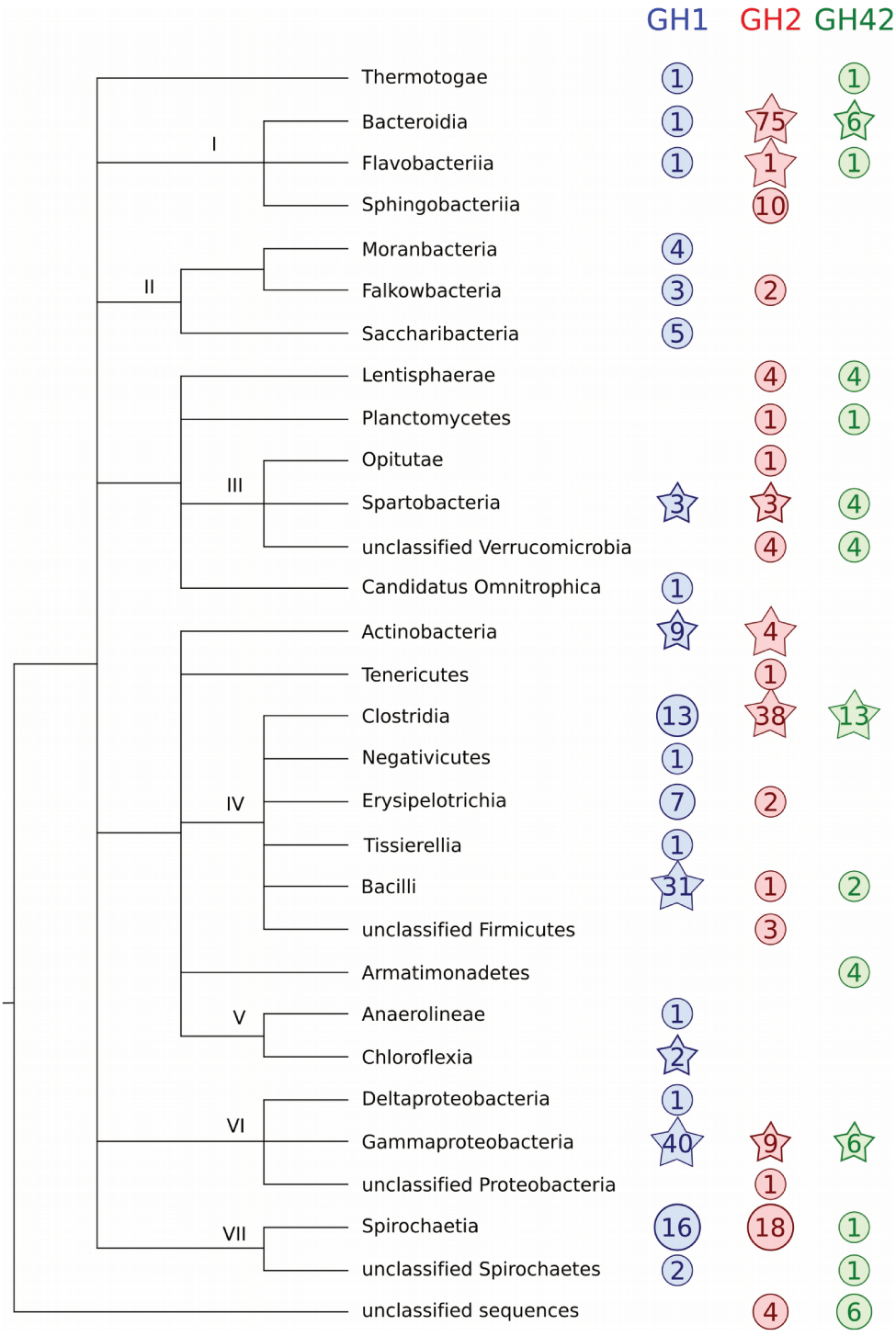

Figura 8.1: Distribución taxonómica de las GH predichas. Las secuencias fueron clasificadas a nivel de clase, usando la clasificación de GTDBtk, en caso que se encontraran en un MAG, o usando el mejor *hit* de BLAST con la base de datos nr, en caso contrario. El número de secuencias para cada familia y cada linaje se indica dentro de un círculo. Las estrellas indican los grupos en los que se seleccionó un candidato para clonar y expresar. I: *Bacteroidetes*; II: *Patescibacteria*; III: *Verrucomicrobia*; IV: *Firmicutes*; V: *Chloroflexi*; VI: *Proteobacteria*; VII: *Spirochaetes*.

La familia GH42 fue la más equitativamente distribuida entre los distintos linajes, mientras que tanto para GH1 y GH2 se encontraron dos clases dominantes: *Gammaproteobacteria* y *Bacilli* para la primera, *Bacteroidetes* y *Clostridia* para la segunda.

Para la familia GH2, Talens-Perales y col. (2016) describieron 5 distintas arquitecturas de dominio (AD), de las cuales 3 tienen actividad galactosidasa sugerida (AD2, AD3 y AD5). La AD está definida por la combinación de dominios presente en una secuencia. Los autores reportaron que 2 dominios están presentes en todas las GH2: un dominio N-terminal de unión a azúcares (GH2N, PF02837) y un barril TIM (GH2C, PF02836). La presencia o ausencia de otros 4 dominios define la AD de una GH2 (Figura 8.2).

De las 183 secuencias clasificadas como GH2, solo 79 (43%) pudieron ser asignadas a alguno de los grupos propuestos (Tabla S8). El tipo más abundante fueron las AD3, con 39 secuencias, seguido por la AD1, con 20. De las AD con actividad galactosidasa, se encontraron 12 para la AD5 y ninguna para la AD2. También se encontró un grupo de 6 secuencias que presentaban una AD no reportada por Talles-Perales y col. Estas tienen el dominio DUF4981, como las AD3, pero carecen el dominio "BGal Small N".

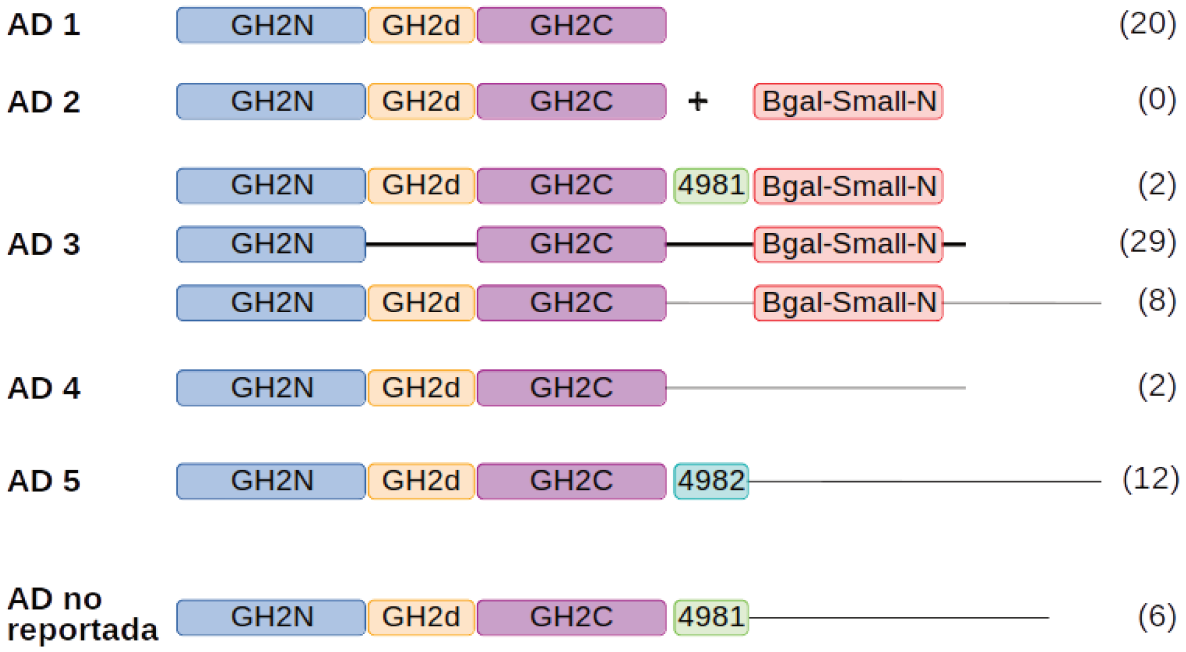

Figura 8.2: Arquitecturas de dominio (AD) de las GH2. Clasificación de las secuencias de la familia GH2, de acuerdo a los dominios presentes en ellas. Entre paréntesis se encuentra el número de secuencias encontradas para esa AD. GH2N: PF02837; GH2d: PF00703; GH2C: PF02836; 4981: PF16353; 4982: PF16355; Bgal-Small-N: PF02929. Figura adaptada de Talens-Pelares y col (2016).

AMPLIFICACIÓN, CLONADO Y EXPRESIÓN DE GENES

Teniendo en cuenta la diversidad taxonómica y de familia de las secuencias encontradas, se seleccionaron 18 genes para intentar su clonado y expresión (Tabla 8.4). De estos 18 genes, 15 fueron amplificados correctamente; solo las βgal2, βgal3 y βgal13 no pudieron ser amplificadas.

Tabla 8.4: Genes elegidos para el clonado y expresión. La elección de las secuencias se realizó intentando representar la diversidad funcional y taxonómica de las B-galactosidasas encontradas. El largo está expresado en número de aminoácidos y el peso molecular (PM) en KDa. La arquitectura de dominio de las GH2 está indicada por el número junto al nombre de la familia (nc: no clasificada)

Gen	Familia	Largo	PM (kDa)	Linaje	Nombre enzima
AUR.contig-100_513_12	GH1	453	51,64	*Spartobacteria*	βgal1
AUR.contig-100_17147_1	GH1	449	49,94	*Chloroflexia*	βgal2
AUR.contig-100_3427_6	GH1	477	51,99	*Actinobacteria*	βgal3
AUR.contig-100_2450_8	GH1	468	54,15	*Tissierella*	βgal4
AUR.contig-100_1620_7	GH42	677	77,67	*Clostridia*	βgal5
AUR.contig-100_1130_5	GH2_nc	619	71,33	*Bacteroidia*	βgal6
CYC.contig-100_2881_9	GH1	469	52,92	*Gammaproteobacteria*	βgal7
CYC.contig-100_4237_11	GH42	698	80,42	*Bacteroidia*	βgal8
CYC.contig-100_9789_2	GH1	480	54,69	*Bacilli*	βgal9
CYC.contig-100_1083_6	GH2_1	605	67,99	*Gammaproteobacteria*	βgal10
CYC.contig-100_2840_6	GH42	670	76,14	*Gammaproteobacteria*	βgal11
AUR.contig-100_220_22	GH2_3	1036	117,69	*Spartobacteria*	βgal12
AUR.contig-100_5793_4	GH2_3	1022	113,33	*Actinobacteria*	βgal13
CYC.contig-100_1928_8	GH2_3	1035	118,82	*Clostridia*	βgal14
AUR.contig-100_1432_6	GH2_3	1025	118,96	*Clostridia*	βgal15
CYC.contig-100_1151_7	GH2_5	823	93,95	*Bacteria*	βgal16
CYC.contig-100_156_7	GH2_5	835	94,96	*Bacteroidia*	βgal17
CYC.contig-100_20362_2	GH2_5	838	94,98	*Bacteroidia*	βgal18

Los 15 genes amplificados fueron incorporados al vector pGEM-T y clonados en *E. coli* DH5α. Para corroborar que el proceso de ensamblado y amplificación fue correcto, se seleccionaron 5 genes para secuenciar por capilares. Para cada β-galactosidasa, se tomaron 3 colonias y se chequeo la presencia del gen de interés mediante *colony PCR*. Para todos los clones positivos, los plásmidos fueron purificados y secuenciados. Todos los genes mostraron leves diferencias (Tabla S7), en promedio menos de 3 mutaciones y menos de 2 mutaciones no sinónimas, lo que confirma que los genes obtenidos mediante herramientas bioinformáticas eran correctos.

El siguiente paso fue incorporar los genes amplificados al vector pET-TEV, con el que se transformó *E. coli BL21* para su expresión. El análisis de las fracciones soluble e insoluble por SDS-PAGE mostró que 5 fueron expresados exitosamente produciendo proteínas del tamaño esperado (Figura 8.3a). Por su parte, βgal1 fue expresada en *E. coli* pero no de manera soluble, sino que se concentró en periplasma. Para lograr obtener una proteína soluble se optó por utilizar *S. cerevisiae BJ3505* como cepa de expresión, ya que para este sistema se cuentan con plásmidos que incluyen péptidos señal que permiten exportar los polipéptidos expresados al medio.

ENSAYOS DE ACTIVIDAD ENZIMÁTICA

Para realizar ensayos de actividad, las proteínas clonadas en *E. coli* se produjeron en medio líquido y fueron purificadas mediante cromatografía de afinidad. La enzima expresada en *S. cerevisiae*, fue utilizada directamente desde el sobrenadante. Las proteínas βgal1, βgal5 y βgal7 mostraron mayor nivel de expresión y, por ello, presentaron mayor cantidad de proteínas purificadas (tabla 8.5). Para las proteínas βgal11 y βgal12, una gran proporción de la expresión se acumuló en la fracción no soluble, impidiendo su purificación por afinidad.

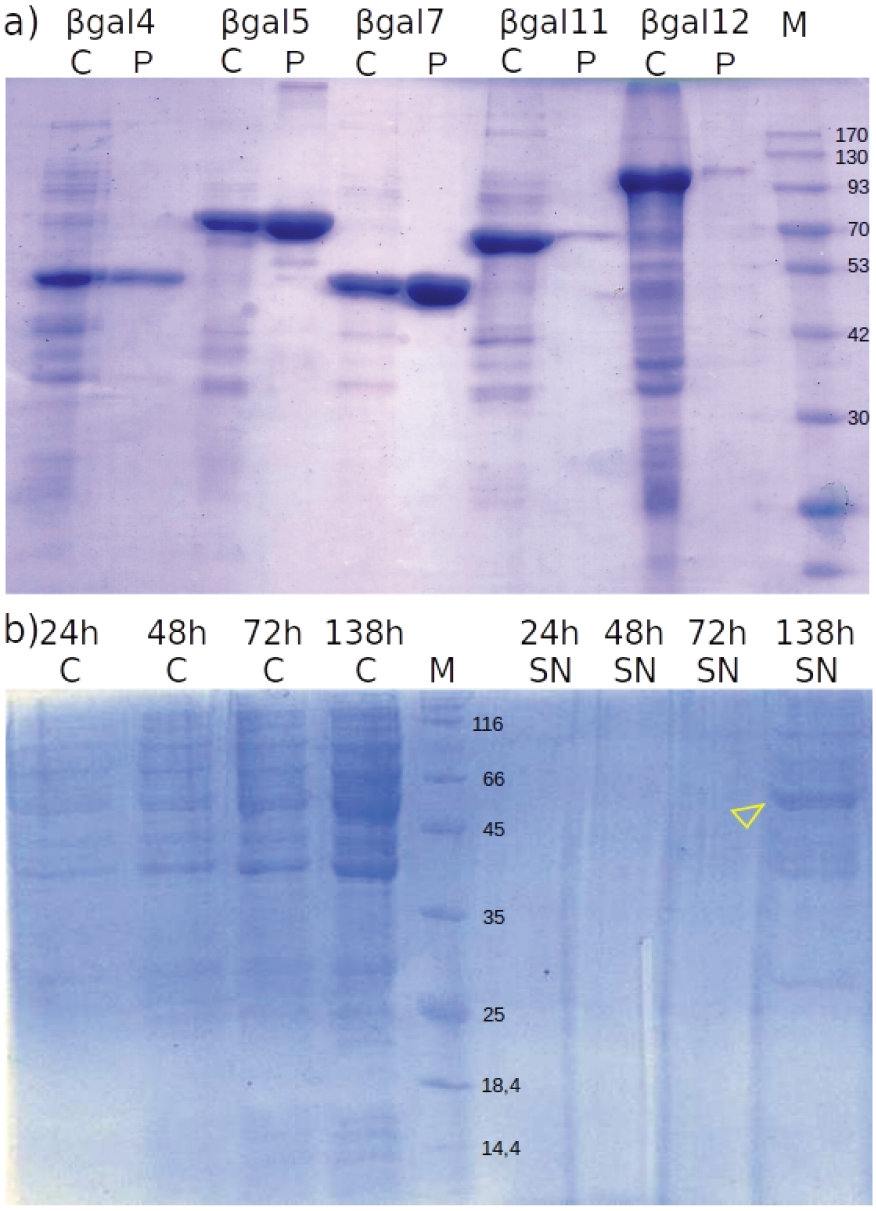

Figura 8.3: Expresión de β-galactosidasas de origen metagenómico. Gel de poliacrilamida con expresiones. a) Expresión en *E. coli* de 5 enzimas. C: extracto crudo; P: purificación por afinidad; M: Marcador de peso molecular. b) Expresión en *S. cerevisiae* de βgal1 a distintos tiempos. C: extracto crudo; SN: sobrenadante; M: marcador de peso molecular.

En 5 de las 6 enzimas expresadas se confirmó actividad β-galactosidasa utilizando ONPG como sustrato. La más destacada fue la βgal5, que mostró niveles de actividad dos órdenes de magnitud por encima del resto.

Para determinar el pH óptimo, se utilizó ONPG como sustrato y se midió actividad enzimática en el rango de pH 4 a 8. Se realizó un análisis similar para determinar la temperatura óptima, midiendo la actividad entre 30°C y 60°C. Todas las enzimas mostraron pH óptimos ligeramente ácidos, entre 6 y 6,5, siendo βgal1 la enzima con mayor rango de pH estable: su actividad se mantiene similar entre pH 5,5 y 7, con un nivel de actividad ligeramente mayor en 6,5. En cuanto a las temperaturas óptimas, βgal1 fue la única mesofílica, teniendo su pico de actividad a los 35°C. El resto mostraron actividad

ligeramente termofílica, con picos de actividad a partir de los 45°C, siendo βgal12 la que mostró mayor temperatura óptima (55°C).

Tabla 8.5: Actividad específica sobre ONPG

Enzima	Familia	Concentración de proteína pura (mg/mL)	Actividad enzimática (U/mg)	pH optimo	Temperatura optima (°C)
βgal1	GH1	1,55	4,65 ± 0,17	6,5	35
βgal4	GH1	0,22	4,10 ± 0,08	6	45
βgal5	GH42	1,47	294,90 ± 19,06	6,5	40
βgal7	GH1	2,03	4,49 ± 0,06	6,5	55
βgal11	GH42	0,18	ND	ND	ND
βgal12	GH2	0,13	5,71 ± 0,13	6	50

Una vez determinada la actividad con el sustrato colorimétrico, se hicieron pruebas con un sustrato natural, como es la lactosa. La mezcla de reacción estuvo compuesta por lactosa 160mM disuelta en buffer Z, y la reacción fue llevada a cabo a temperatura óptima. Nuevamente, la actividad de βgal5 fue órdenes de magnitud superior a la del resto (tabla 8.6). Es por ello que se optó por caracterizar sólo la actividad de esta enzima.

En primer lugar, se evaluó la termoestabilidad de la enzima. Como muestra la Figura 8.5, usando la temperatura óptima de 40°C, la enzima mantiene un 85% de su actividad luego de 1 hora de incubación, y se mantiene por encima del 55% de actividad luego de 8 horas. Sin embargo, cuando se mide la actividad a 45°C, luego de tan solo 1 hora de incubación, la actividad cae al 20%.

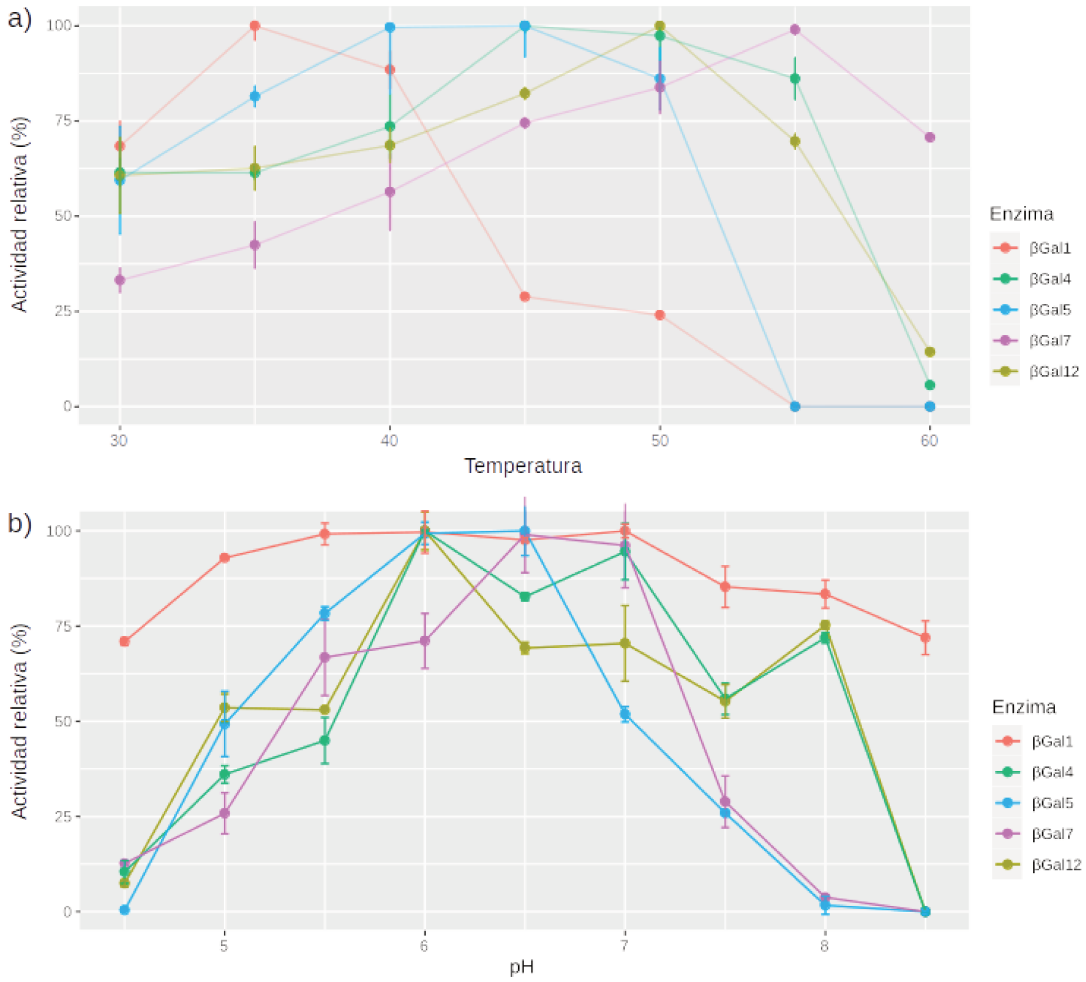

Figura 8.4: Determinación de pH y temperatura óptimos. La actividad enzimática fue medida utilizando ONPG como sustrato. El valor de actividad más alto para cada enzima fue considerado 100%. Los datos fueron tomados mediante réplicas biológicas (n=3).

Tabla 8.6: Actividad específica sobre lactosa

Enzima	Actividad con lactosa (U/mg)
βgal1	8,23 ± 0,08
βgal4	0,96 ± 0,02
βgal5	142,33 ± 18,67
βgal7	0,50 ± 0,01
βgal12	2,34 ± 0,22

También se testeó el efecto de 4 cationes en la actividad enzimática. La adición de Ca2+ incrementó la actividad un 26%, mientras que Zn2+ provocó un aumento del 12% y al agregar K+ se obtuvo un

incremento del 10% comparado con el control sin cationes (Tabla 8.7). En contraposición, al agregar Mg2+ la actividad fue inhibida en un 14%.

Tabla 8.7: Efecto de iones en la actividad de βgal5.

Reactivo añadido	Concentración (mM)	Actividad relativa (%)
MgCl2	10	86.4±7.7
	1	99.0±3.0
ZnSO4	10	108.1±4.4
	1	111.9±7.4
CaCl2	10	121.1±3.7
	1	126.6±3.5
KCl	10	102.1±3.2
	1	109.7±10.9

Para calcular los parámetros cinéticos de βgal5, se utilizaron distintas concentraciones de ONPG, abarcando de 2,5mM a 20mM. Se obtuvo una Vmax=93,46 µM/min, y Km= 14, 55 mM. Se realizó el mismo procedimiento utilizando lactosa y se obtuvo una Vmax=18,33 µM/min, y Km= 400 mM.

También se evaluó la especificidad de sustrato, utilizando 7 compuestos distintos. La mayor actividad se observó al utilizar 4-nitrofenil-β-D-glucopiranosa, mientras que también se encontró actividad, pero en menor medida, con 4-nitrofenil-β-D-fucopiranosa.

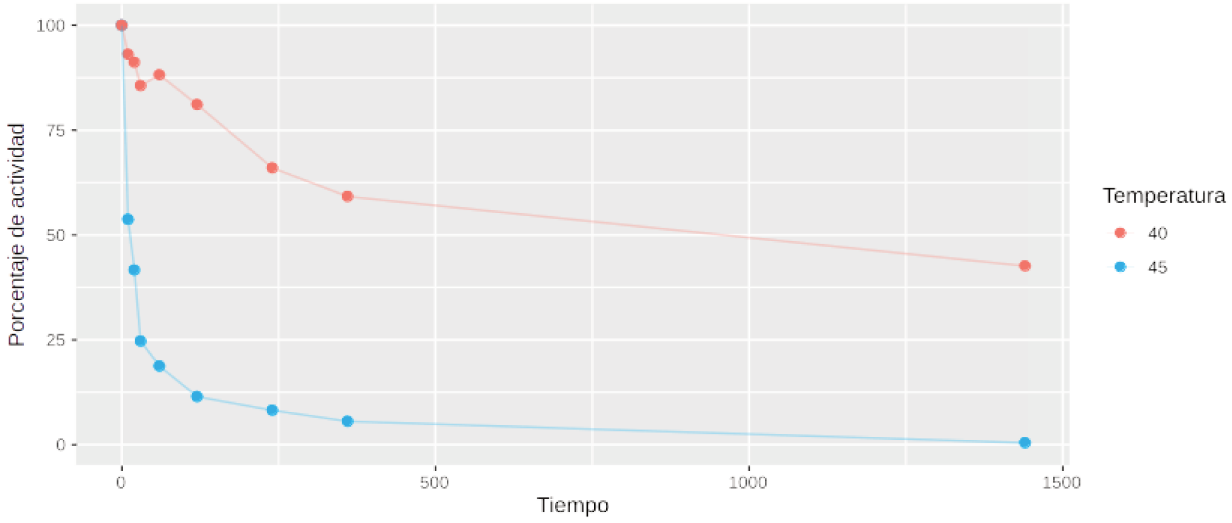

Figura 8.5: Termoestabilidad de la enzima βgal5.

Finalmente, se evaluó la actividad transgalactosidasa de la enzima, utilizando HPLC, con lactosa como sustrato de hidrólisis y aceptor de residuos de galactosa, para la generación de GOS. En condiciones de actividad óptimas, 40°C y pH 5,5, y una concentración inicial de lactosa del 40%, βgal5 fue capaz de transformar tan sólo el 2% de la lactosa en un trisacárido luego de 8 hs de incubación.

DISCUSIÓN

Debido a la composición de las aguas residuales que se vierten en las lagunas de pequeñas empresas lácteas, que contienen resto de leche o subproductos como el lactosuero, se podría esperar que los microorganismos en este tipo de ambiente cuenten con diferentes tipos de enzimas que actúen sobre carbohidratos. Los resultados obtenidos muestran que 1 de cada 100 genes predichos fueron clasificados como CAZymas, una proporción similar a la observada en otros ambientes, como rumen de vacas (Hess y col., 2011) e intestino de cerdos (Yang y col., 2016). También observamos que 1 de cada 65 CAZymas pertenece a alguna de las familias de interés: GH1, GH2, GH39, GH42 y GH59. Como el número de secuencias asignadas a las familias GH39 y GH59 es muy bajo (6 y 0 genes no parciales respectivamente), no fueron consideradas para los siguientes análisis ni para el clonado y expresión. En total, se identificaron 379 secuencias candidatas, distribuidas en 14 *phyla*, con 10 secuencias que no pudieron ser clasificadas taxonómicamente, ya que no contaban con ningún resultado cercano (al menos 50% identidad) en la base de datos de nr de la NCBI.

La identificación de las enzimas se basó en la presencia de dominios catalíticos, obtenidos de la base de datos dbCAN. A su vez, estos perfiles fueron construidos a partir de las secuencias disponibles en la base de datos CAZy. Esta última base de datos plantea una clasificación de enzimas en familias de acuerdo la similitud de las secuencias y la caracterización bioquímica de, al menos, un miembro de dicha familia. Sin embargo, ocurre que miembros de una misma familia pueden actuar sobre distintos sustratos, lo que dificulta la anotación funcional automática. Por ejemplo, la familia GH1 posee 20 actividades diferentes y la familia GH2, 10 actividades diferentes. Esta última familia ha sido estudiada en profundidad y se ha sugerido que las distintas actividades podrían estar relacionadas con la combinación de pequeños dominios (Talens-Pelares y col., 2016). Usando los criterios establecidos por estos autores, el número de genes candidatos para la familia GH2 que tendrían actividad β-galactosidasa pasó de 183 a 79 (43%). En la medida que incremente el número de enzimas caracterizadas en profundidad, procesos similares podrían llevarse a cabo para otras familias, lo cual daría lugar a la identificación de residuos o dominios claves y la mejora de la predicción funcional de CAZymas.

Utilizando la clasificación taxonómica y de familia, se seleccionaron 18 genes candidatos, de los cuales 6 se pudieron expresar y 5 mostraron actividad tanto con ONPG como con lactosa, lo cual representa, aproximadamente, un 28% de efectividad en el clonado y expresión. El pH óptimo de las distintas enzimas osciló entre 5 y 7, mientras que la temperatura óptima varió entre 30°C y 55°C. Dado

que el pH natural de la leche es de alrededor de 6,8, para utilizar una β-galactosidasa para hidrolizar lactosa en leche es necesario que su actividad ronde este valor. Por lo tanto, las enzimas más adecuadas para este eventual uso serían las βgal1, βgal5 y βgal7. La actividad específica con ONPG como sustrato de βgal5 fue de 280 U/mg, mientras que con lactosa bajó casi a la mitad: 142 U/mg. Este comportamiento se ha visto en otras enzimas de la familia GH42 (Park y Oh, 2010; Di Lauro y col, 2008).

A la fecha, existen pocos reportes de otras β-galactosidasas de origen metagenómico (tabla 8.8). Se han reportado dos enzimas termófilas, ambas extraídas de muestras de suelo: Gal308 cuya temperatura óptima es de 78°C (Zhang y col., 2013) y ZD410, una enzima que, aunque su temperatura óptima sean 50°C, conserva actividad a 20°C y 0°C (Wang y col., 2010). M1 es una enzima aislada de suelo capaz de hidrolizar completamente lactosa de leche en 25 horas (Erich y col., 2015). Por otra parte, Lac161_ORF7 es una β-galactosidasa que no pudo ser encontrada por análisis bioinformático, pero si mediante pruebas funcionales (Cheng y col., 2017). La enzima BGalI17E2 fue identificada en muestras de ikaita submarina, y se reportó que es activa a bajas temperaturas y en condiciones alcalinas (Vester y col., 2014). Por último, BgaC es una enzima aislada de microbioma humano y activa en condiciones fisiológicas (Mulualem y col., 2021), mientras BGal_375 es la única β-galactosidasa identificada mediante metagenómica basada en secuencias a partir de muestras de suelo (Liu y col., 2019). La mayoría de estas enzimas mostraron actividad tanto con ONPG como con lactosa, pero ninguna mostró niveles de actividad mayores a βgal5.

Tabla 8.8: Comparación de βgal5 con otras enzimas de origen metagenómico. El pH y temperatura (Temp.) mostrados son los indicados como óptimos en cada reporte. PM: Peso molecular; ND: No disponible. Referencias (Ref.): [1] Eberhardt y col. (2021); [2] Zhang y col. (2013); [3] Wang y col. (2010); [4] Liu y col. (2019); [5] Erich y col. (2015); [6] Vester y col. (2014); [7] Cheng y col. (2017); [8] Mulualem y col. (2021)

Enzima	Ambiente	Tipo de ensayo	Familia	pH	Temp. (°C)	Km ONPG (mM)	Km Lactosa (mM)	Actividad ONPG (U/mg)	Actividad Lactosa (U/mg)	Ref.
βgal5	Lagunas de estabilización	Secuencia	GH42	6,5	40	14,55	403	294,9	142,3	[1]
Gal308	Suelo	Funcional	GH42	6,8	78	2,7	7,1	185	47,6	[2]
zd410	Suelo	Funcional	GH42	7	38	1,7	12,2	243	25,4	[3]
BGal_375	Suelo	Secuencia	GH2	8	50	1,65	ND	15,6	0,96	[4]
M1	Suelo	Funcional	GH2	7	37	ND	14,3	ND	ND	[5]
BGalI17E2	Ikaita submarina	Funcional	GH1	6	37	ND	ND	ND	ND	[6]
Lac161_ORF7	Suelo	Funcional	GH2	6	50	ND	1,8	ND	ND	[7]
BgaC	Microbioma humano	Funcional	GH2	7	37	2,5	3,7	107	22	[8]

La mayor limitante de la metagenómica basada en secuencia es la selección de candidatos. A partir de los análisis *in silico* es posible encontrar miles de candidatos, pero como el método de clonado y expresión heteróloga no es fácilmente escalable, es imposible probar todos las enzimas putativas. Por otra parte, la diversidad funcional dentro de una familia a CAZymas agrega otra capa más de dificultad a la hora de buscar una actividad en particular, como pueden ser las β-galactosidasas. Una alternativa es utilizar un enfoque basado en metagenómica funcional, donde el ADN metagenómico es clonado en vectores de expresión y se realiza una detección funcional con distintos sustratos de los clones obtenidos. Sin embargo, esta es una estrategia poco eficiente: a partir de una muestra ambiental se ha reportado un clon positivo para glucosidasas cada 31.190 clones estudiados (Ferrer y col, 2016). Son muchos los factores que pueden influir en la baja eficiencia, como el nivel de expresión del gen, ensayos de actividad ineficientes (Ngara y Zhang., 2018) o la selección del sustrato adecuado para una enzima (Nguyen y col. 2012; Ferrer y col. 2016). Si bien cada resultado positivo usando este enfoque es una enzima activa, la baja eficiencia y la cantidad de factores que pueden afectar a las pruebas de actividad hacen de la metagenómica basada en secuencia un enfoque más atractivo que la metagenómica funcional. En este capítulo se demostró que haciendo una selección correcta de enzimas candidatas, es factible la obtención de proteínas activas sin la necesidad de un gran número de candidatos.

CONCLUSIONES

Utilizando una estrategia de identificación, clonado y expresión de enzimas a partir de datos metagenómicos, fueron exitosamente expresadas 5 β-galactosidasas, alcanzando una tasa de éxito de alrededor del 28%. Todas las enzimas son activas tanto con ONPG, un sustrato artificial, como con lactosa, un sustrato natural. Entre las enzimas obtenidas, βgal5 mostró niveles de actividad significativamente mayores a los de las demás, βgal5 representa un punto de partida de interés para futuros ensayos, como por ejemplo de mutagénesis, para mejorar su actividad específica con lactosa, maximizar los niveles de hidrólisis o aumentar la actividad transgalactosidasa.

IDENTIFICACIÓN, CLONADO Y EXPRESIÓN DE PROTEASAS

9 IDENTIFICACIÓN, CLONADO Y EXPRESIÓN DE PROTEASAS

INTRODUCCIÓN

Las proteasas o peptidasas (EC: 3.4) son enzimas degradativas que hidrolizan enlaces peptídicos (Barrett y McDonald, 1986). El uso de proteasas se ha extendido por miles de años, desde la quimosina que forma parte del cuajo y es usada en la producción de quesos (Mishra y col., 2016), hasta la aplicación actual en distintas industrias, como la alimenticia, farmacéutica, producción de cueros y detergentes, entre otras tantas (Razzaq y col., 2019). Si bien es posible realizar hidrólisis química de proteínas, ya sea ácida o alcalina, ésta tiende a ser difícil de controlar y pueden producir residuos reducidos o aminoácidos inusuales, como la lisinoalanina (Provansal y col., 1975; Tavano, 2013). En cambio, el uso de proteasas permite la hidrólisis de enlaces peptídicos específicos, evitando exponer las proteínas a ambientes extremos que puedan modificar los aminoácidos y manteniendo, en el caso de la industria alimenticia, el valor nutricional de las proteínas a hidrolizar (Maldonado y col., 1998). Actualmente, se distinguen dos grandes tipos de enzimas con actividad proteolítica: las exopeptidasas, que hidrolizan enlaces terminales en una proteína, y las endopeptidasas, que clivan enlaces no terminales (McDonald, 1985). En la primer categoría pueden distinguirse las aminopeptidasas, que actúan sobre el extremo N-terminal, las carboxilpeptidasas, que actúan sobre el extremo C-terminal, las dipeptidil-peptidasas, que liberan dipéptidos, y las tripeptidil-peptidasas, que liberan tripéptidos (Tavano y col., 2018).

En 1993 se estableció una clasificación jerárquica para peptidasas, que consistió en agrupar enzimas homólogas en “especies”, luego clasificar las especies en “familias” y, por último, agrupar las familias en “clanes” (Rawlings y Barrett, 1993). Las proteasas se distinguen principalmente por su residuo catalítico: aspartil peptidasas, cisteín peptidasas, glutamil peptidasas, metalopeptidasas, asparagin peptidasas, serín peptidasas, treonin peptidasas, peptidasas mixtas (que poseen más de un tipo de residuo catalítico) y aquellas que se desconoce su residuo catalítico (Rawlings y col., 2016). MEROPS es una base de datos curada, que contiene la información de proteasas e inhibidores, incluyendo su familia, clan, secuencia e información sobre su actividad. Sin embargo, a diferencia de lo que ocurre para las β-galactosidasas discutidas en el capítulo anterior, no existe una base de datos con perfiles de HMM para todas las familias de proteasas. El uso de perfiles de HMM es recomendado dado que se construyen a partir de alineamientos de secuencias, lo que permite obtener información sitio-específica de las familias de proteínas, que los métodos basados en alineamientos de pares de secuencias no consideran, y la identificación de homólogos lejanos (Eddy, 2009). En cambio, MEROPS ofrece una herramienta de búsqueda basada en comparación de secuencias de a pares, utilizando BLAST.

Uno de los objetivos del presente trabajo es la identificación de proteasas en el metagenoma de lagunas de estabilización de PYMES lácteas para su utilización en la valoración de lactosuero. Diferentes enzimas, tanto comerciales como obtenidas a partir de aislamientos o cultivos celulares, han sido reportadas capaces de hidrolizar el WPC (Sinha y col., 2007; Madureira y col., 2010; Corrochano y col., 2018). Las más utilizadas son la tripsina (perteneciente a la familia S01), subtilisina (S08), pepsina (A01), quimotripsina (S01), proteinasa K (S08), papaína (C01). También se ha reportado el uso de cócteles, como Corolasa PP y Flavourzyme (Mers y col., 2015), pero su composición no ha sido detalladamente reportada.

En el siguiente capítulo se describe la identificación, selección, clonado y expresión exitosa de proteasas a partir de ADN metagenómico, y su caracterización funcional.

IDENTIFICACIÓN DE PROTEASAS

En primer lugar, se realizó una búsqueda general de todas las familias de proteasas en la totalidad de genes descritos previamente. Se identificaron en total 19.177 genes candidatos, 7.236 para AUR y 11.941 para CYC (Tabla 9.1). Se encontraron candidatos para los 9 tipos de proteasas listados en MEROPS, aunque para el grupo de las glutamil proteasas solo se identificaron 2 genes. El grupo más abundante y más diverso fue el de las metaloproteasas, con un total de 7010 secuencias, distribuidas en 82 familias distintas, seguido por las serin proteasas, con 6.246 secuencias y 47 familias.

Tabla 9.1 Predicción de proteasas en ambos metagenomas.

	AUR		CYC		Total	
Tipo	Genes	Familias	Genes	Familias	Genes	Familias
Aspartil proteasas (A)	182	10	338	13	520	13
Cistein proteasas (C)	1387	28	2222	30	3609	33
Glutamil proteasas (G)	0	0	2	1	2	1
Metaloproteasas (M)	2648	72	4362	76	7010	82
Asparagin proteasas (N)	48	2	73	3	121	3
Serin proteasas (S)	2362	42	3884	46	6246	47
Treonin proteasas (T)	185	7	238	6	423	7
Mixtas (P)	33	2	56	3	89	3
Desconocido (U)	391	6	766	7	1157	8

Para el presente trabajo, se optó por continuar con secuencias de familias con actividad reportada sobre proteínas de suero. De las 4 familias nombradas anteriormente, se encontraron candidatos para

todas, salvo para A01 (Tabla 9.2). Dado que las más abundantes fueron las dos serin proteasas, S01 y S08, se decidió continuar con ellas dos para la búsqueda de candidatos para el clonado y expresión.

Tabla 9.2 Número de secuencias para las familias de proteasas de interés. Las familias seleccionadas para continuar con el clonado y expresión se indican en negrita.

Familia	Cantidad de genes	Genes no parciales
A01	0	0
C01	122	67
S01	**505**	**229**
S08	**224**	**117**

La siguiente etapa consistió en la clasificación taxonómica de los miembros de las familias S01 y S08. Las proteasas se distribuyeron en 21 *phyla* y 49 clases distintas (Figura 9.1). Se encontró una sola proteasa de *Archaea*, una S08 perteneciente al *phylum Euryarchaeota*; 776 serían de origen bacteriano mientras que 2 no tuvieron ningún hit en la base de datos nr, por lo que permanecieron como no clasificadas. Tanto para la familia S01 como para S08, las tres clases más abundantes fueron *Clostridia*, *Gammaproteobacteria* y *Bacteroidetes*, representando en ambos casos, aproximadamente la mitad de los genes encontrados.

Figura 9.1: Distribución taxonómica de las proteasas predichas. I: Bacteroidetes; II: Verrucomicrobia; III: Firmicutes; IV: Patescibacteria; V: Proteobacteria; VI: Spirochaetes; VII: Synergystetes.

AMPLIFICACIÓN, CLONADO Y EXPRESIÓN DE GENES

Nuevamente, intentando representar ambas familias de interés y su diversidad taxonómica, se seleccionaron 10 genes candidatos para continuar con el clonado y expresión (Tabla 9.3). De ellos, 7 fueron exitosamente amplificadas del metagenoma: solo Pr02, Pr03 y Pr07 no pudieron ser amplificadas.

Tabla 9.3 Proteasas seleccionadas para el clonado y expresión.

Gen	Familia	Largo	PM (kDa)	Linaje	Nombre enzima
AUR.contig-100_17930_3	S08	564	58,5	*Synergistia*	Pr01
AUR.contig-100_23747_2	S08	759	80,95	*Chloroflexia*	Pr02
AUR.contig-100_4_168	S01	553	57,39	*Gammaproteobacteria*	Pr03
AUR.contig-100_591_6	S08	391	39,63	*Actinobacteria*	Pr04
AUR.contig-100_7246_4	S01	418	43,45	*Saccharibacteria*	Pr05
CYC.contig-100_1230_19	S01	550	61,31	*Bacteroidia*	Pr06
CYC.contig-100_2339_2	S08	1662	175,05	*Kiritimatiellae*	Pr07
CYC.contig-100_44516_2	S01	473	51,34	*Epsilonproteobacteria*	Pr08
CYC.contig-100_7627_4	S01	431	46,39	*Spirochaetia*	Pr09
CYC.contig-100_8118_5	S01	365	37,54	*Erysipelotrichia*	Pr10

Se intentó clonar los 7 genes amplificados en *E. coli BL21*, pero ninguno pudo ser expresado exitosamente. Luego, se intentó la expresión de los candidatos, usando *S. cerevisiae* como sistema alternativo (Tabla S6) y uno de los genes, Pr10, pudo ser expresado exitosamente (Figura 9.2). La proteína fue concentrada mediante filtrado del sobrenadante de cultivo.

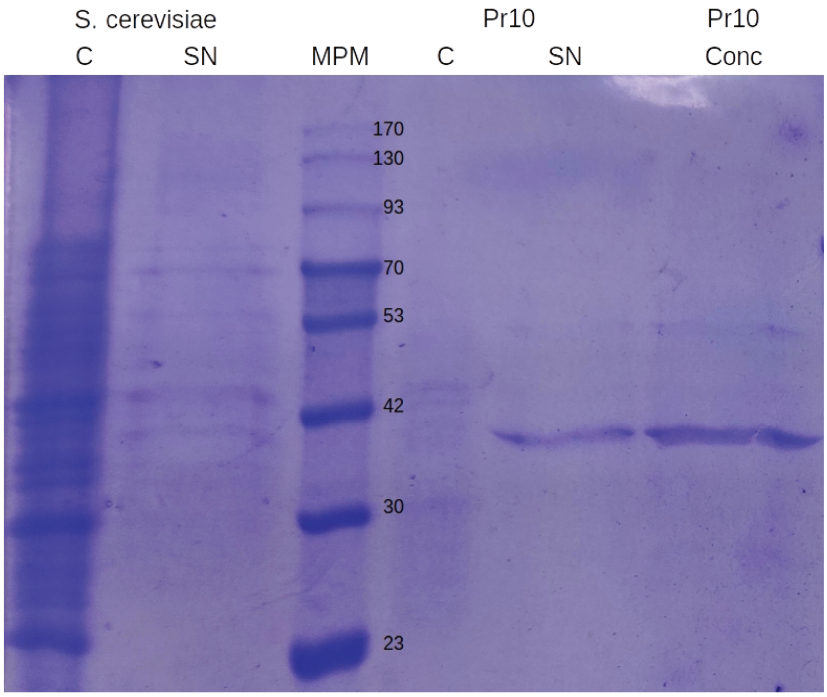

Figura 9.2: Expresión de las proteasas de origen metagenómico. Gel de poliacrilamida de la expresión de la proteasa Pr10. La concentración se hizo filtrando el sobrenadante de un cultivo de 184 hs de la cepa de expresión con filtros de 3kDa. M: marcador de peso molecular; P: pellet; SN: sobrenadante; Conc: enzima concentrada.

ENSAYOS DE ACTIVIDAD ENZIMÁTICA

Para comprobar la actividad proteasa, se realizó un ensayo con azocaseína. En primer lugar, se evaluó la actividad a distintas temperaturas (Figura 9.3a). La actividad máxima se observó a 50ºC, aunque a 25ºC y 30ºC la enzima conserva más de un 90% de su actividad. A temperaturas mayores a 50ºC se observó una caída fuerte en el nivel de actividad.

Luego, se evaluó la actividad de la enzima a diferentes niveles de pH. A diferencia de lo que se observa con la temperatura, la actividad se mantiene estable en todo el rango evaluado, alcanzando el valor máximo en 6 y 6,5, pero manteniendo más del 90% de su actividad en el rango de pH 5 a 9 (Figura 9.3b).

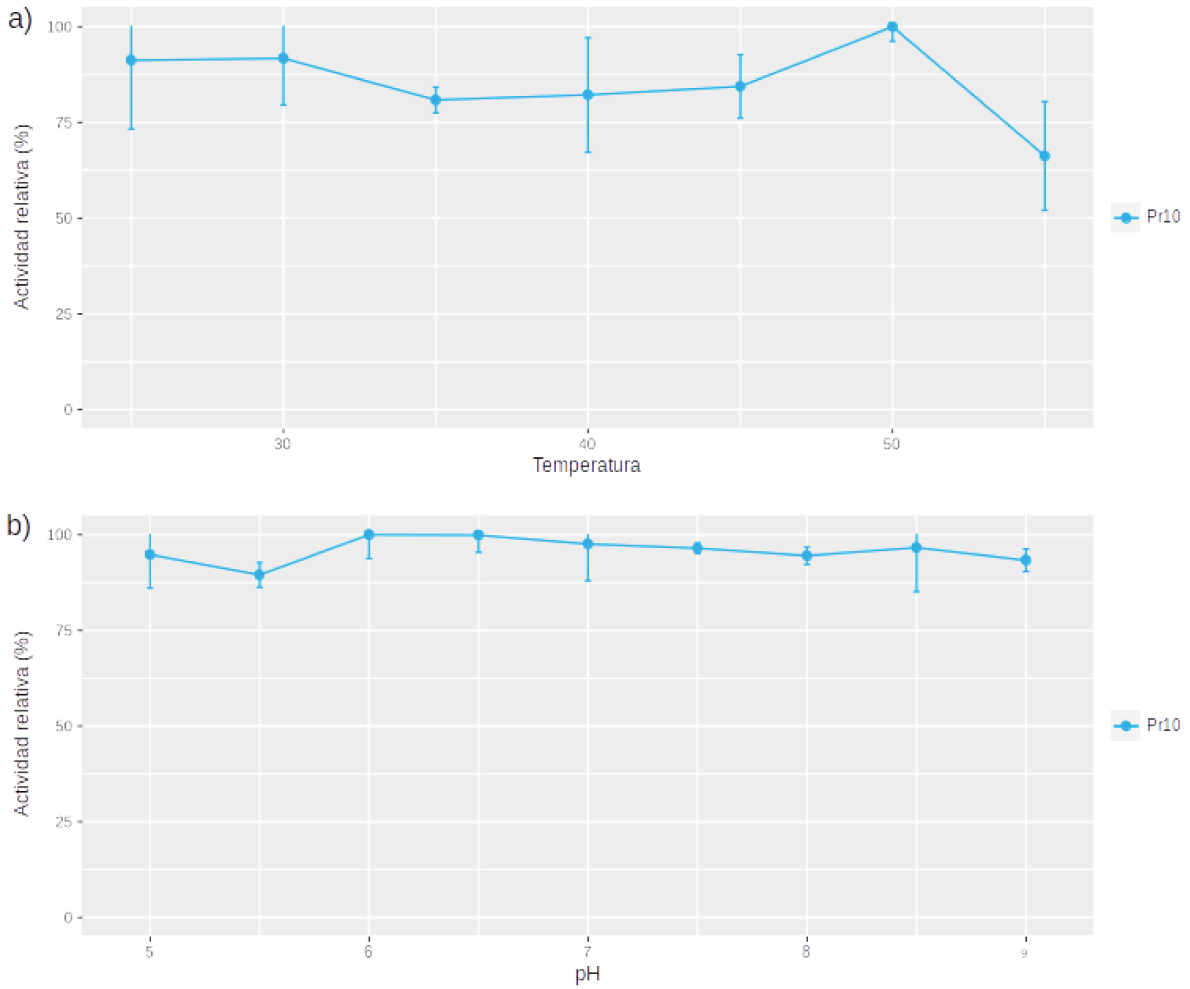

Figura 9.3: Determinación de temperatura (a) y pH (b) óptimos para Pr10. La actividad enzimática fue medida utilizando azocaseína como sustrato. El valor de actividad más alto para la enzima fue considerado 100%. Los datos fueron tomados mediante réplicas biológicas (n=3).

Por último, se probó la actividad con un sustrato natural, como la leche. La enzima fue sembrada en placas de agar-leche (1%) y se incubó durante 24 horas. El halo observado en la placa (Figura 9.4) indica que la enzima podría ser capaz de hidrolizar tanto caseína como proteínas del suero.

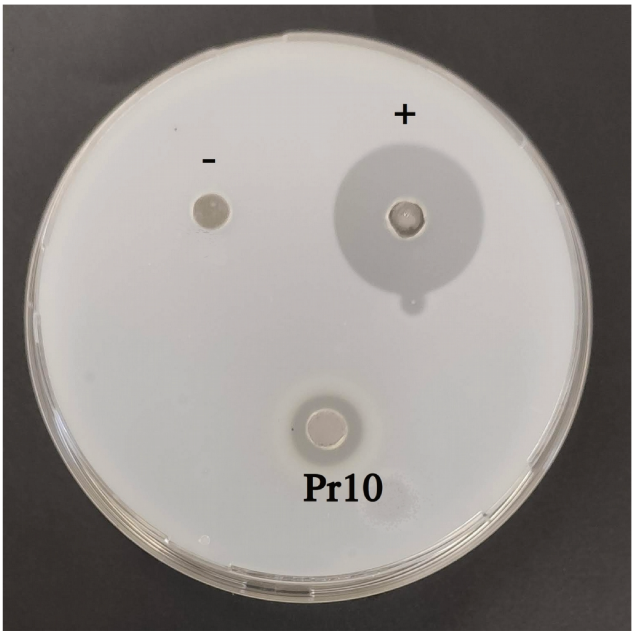

Figura 9.4: Determinación de actividad con sustratos naturales. La actividad de la enzima se evaluó para leche, utilizando placas de agar. Como control negativo (-) se utilizó el sobrenadante de la cepa de *S. cerevisiae* sin transformar y como control positivo (+) la enzima comercial Novozym (10028, Novozyme).

DISCUSIÓN

La secuenciación de metagenomas y la aplicación de técnicas de ADN recombinante ha permitido el descubrimiento de nuevas enzimas con características diferenciales (Ferrer y col., 2007). Un grupo de especial interés son las proteasas, que representan el 60% del mercado mundial de enzimas y se utilizan para la producción de alimentos, detergentes, fármacos y agroquímicos, entre otros (Sanchez y Demain, 2017).

De los más de 2 millones de genes predichos, en total se encontraron 19.177 proteasas putativas (~0,9%), un número similar a las *CAZymes* encontradas (21.296). Como se comentó en otros capítulos, la calidad de la anotación de una secuencia depende de la calidad de datos disponibles contra los cuales comparar. Las proteasas son enzimas con una gran diversidad estructural y funcional, lo que dificulta establecer un sistema general de clasificación (Gurumallesh y col., 2019). Han sido clasificadas de acuerdo a su forma de hidrolizar péptidos en “exopeptidasas” y “endopeptidasas” o, en base a su pH óptimo, en proteasas ácidas, neutras o alcalinas (Rao y col., 1998). El mayor esfuerzo para caracterizar estas enzimas en base a su secuencia y estructura es la base de datos MEROPS, que plantea una clasificación en 9 familias distintas, de acuerdo principalmente al residuo catalítico de la enzima. Sin embargo, no existe información bibliográfica que permita vincular dichas familias con actividades específicas. Así mismo, existen subfamilias propuestas basadas en una enzima caracterizada (por ejemplo, para S08 están propuestas las subfamilias S08A y S08B), pero no se indica claramente qué diferencia estas subfamilias. Otra limitante de este sistema de clasificación es que las familias y subfamilias propuestas agrupan secuencias muy diversas, lo cual dificulta la construcción de perfiles de HMM para la anotación

automática de genes. El interés actual en péptidos activos, capaces de llevar a cabo acciones específicas como antihipertensivos u opioides (Brandelli y col., 2015), hacen cada vez más necesario disponer del mayor conocimiento posible sobre la secuencia y estructuras de distintas proteasas, permitiendo una mejor predicción automática de enzimas mediante herramientas bioinformáticas.

Los 9 grupos de enzimas reportados por la base de datos MEROPS fueron encontrados, pero en cantidades muy disímiles: las metaloproteasas (7010 genes) y las serinproteasas (6246) fueron las más abundantes, mientras que las mixtas (89) y las glutamil proteasas (2) fueron las familias menos numerosas. En cuanto a las 4 familias de interés, las S01 (similares a tripsina) fueron las más abundantes, con 505 genes candidatos, seguidas de las S08 (similares a subtilisina), con 224 genes candidatos. No se encontró ninguna proteasa de la familia A01 (similares a pepsina), mientras que para las C01 (similares a papaína) se encontraron candidatos, pero más de la mitad (67 de 122) eran genes parciales. En este primer intento de expresión de proteasas se optó hacer foco en S01 y S08, dejando la posibilidad de clonar y expresar miembros de la familia C01 o algún miembro de otra familia en el futuro.

La expresión de las proteasas en *E. coli* no pudo llevarse a cabo de manera exitosa. Esta bacteria secreta muy pocas proteínas al medio extracelular (Hannig y Makrides, 1998); por el contrario, la mayoría de las enzimas expresadas permanecen en citoplasma o periplasma. Como expresar una proteasa recombinante en citoplasma podría ser perjudicial para la cepa de expresión, se decidió buscar un sistema alternativo y se optó por uno que secretase al medio externo las proteínas expresadas. Es por ello que se utilizó *S. cerevisiae* como cepa de expresión. Sin embargo, una de las desventajas de este sistema de expresión es que dificulta la obtención de la enzima pura, con métodos como el de afinidad. El plásmido YEp no posee un extremo de histidinas, como el plásmido pET-TEV de *E. coli*, y los cebadores utilizados son demasiado largos (más de 50 pares de bases) para incorporarles este extremo. Dado que la enzima se encuentra en el sobrenadante y en el gel de poliacrilamida no se observan otras proteínas, se planteó la alternativa de utilizar filtros para eliminar pequeñas partículas y tener una enzima lo más pura posible. Sin embargo, para futuras aplicaciones, es necesario buscar una alternativa que permita lograr una enzima que pueda ser purificada.

La mayor actividad con azocaseína fue de 501,73±31,24 U/ml y se observó a 50ºC y pH 6. La enzima mostró ser sensible a cambios en la temperatura, perdiendo más del 30% de su actividad a 55ºC. Por otro lado, la actividad en todos los pH medidos se mantuvo estable, mostrando una disminución de actividad siempre menor al 10%. La mayor actividad se observó a pH ligeramente ácidos (6 y 6,5), un resultado deseado para aplicaciones relacionadas a leche y productos lácteos.

Numerosas proteasas de origen metagenómico han sido reportadas (Tabla 9.4). Todos los estudios, salvo uno, se basan en búsquedas funcionales: la única enzima identificada mediante metagenómica basada en secuencia es Prt1SU, una enzima proveniente de residuos sólidos de curtiembre (Verma y

Sharma, 2021). Todas las enzimas reportadas poseen un pH óptimo alcalino, y casi en su totalidad son termófilas, destacándose pF1AL2, una proteasa proveniente de sedimento marino, cuya temperatura óptima es de 80ºC (Sun y col., 2020). La mayor limitación a la hora de comparar enzimas con actividad proteasa es que no existe un protocolo de determinación de actividad establecido que se utilice en todos los reportes. En aquellos reportes en los que se utiliza caseína como sustrato, la actividad está determinada por la cantidad de tirosina liberada. En cambio, aquellos reportes en los que se utiliza azocaseína como sustrato, el nivel de actividad es usualmente determinado por las variaciones en el nivel de absorbancia. Sin embargo, los tiempos de incubación y los valores de absorbancia usados varían. Existen estudios donde la unidad de actividad proteasa fue definida como la cantidad de enzima necesaria para aumentar la absorbancia 0.1 a A405 luego de una hora de incubación (Lee y col., 2007; Neveu y col., 2011), mientras que en otros se define como la cantidad de enzima necesaria para aumentar la absorbancia 0.01 a A405 luego de 30 minutos de incubación (Pushpam y col., 2011). Dado que el ensayo de actividad realizado en este estudio fue de 30 minutos, se optó por utilizar la definición de Pushpam y col. (2011). Esta enzima perteneciente a la familia S01 proviene de microorganismos presentes en la piel de cabra y posee una actividad de 100 U/ml, aunque es muy sensible a cambios de temperatura y pH. Por su parte, Pr10 posee una mayor estabilidad térmica y de pH. La comparación con otras enzimas se vuelve imposible dado que las condiciones de ensayo y la definición de las medidas de actividad no son las mismas. Es importante notar que tanto en Pushpam y col. como en el resto de los reportes consultados se define el valor de unidad enzimática (U), pero nunca se aclara si el volumen usado para definir

Tabla 9.4 Comparación de Pr10 con otras enzimas de origen metagenómico. El pH y temperatura (Temp.) mostrados son los indicados como óptimos en cada reporte. 1) Waschkowitz y col., 2009; 2) Lee y col., 2007; 3) Morris y Marchesi, 2016; 4) Neveu y col., 2011; 5) Pushpam y col., 2011; 6) Zhang y col., 2011; 7) Biver y col., 2013; 8) Purohit y Singh, 2013; 9)Sun y col., 2020; 10) Gong y col., 2017; 11) Verma y Sharma, 2021; 12) Devi y col., 2016; 13) Pessoa y col., 2017

Enzima	Ambiente	Tipo de ensayo	Familia	Sustrato	pH	Temp. (°C)	Referencia
Pr10	Lagunas de estabilización	Secuencia	S01	Azocaseína	6	50	Este estudio
mprA	Suelo	Funcional	M04	Caseina	8	65	[1]
mprB	Suelo	Funcional	M04	Caseina	8	65	[1]
pES63H9	Sedimento marino	Funcional	M12	Azocaseína	7	50	[2]
M1-2	Aguas de tratamiento	Funcional	M28	Azocaseína	8	42	[3]
M30	Arena	Funcional	S08	Azocaseína	12	40	[4]
DV1	Arena	Funcional	S08	Azocaseína	8	55	[4]
AS-protease	Piel de cabra	Funcional	S08	Azocaseína	10,5	42	[5]
ACPRO001	Sedimento de costa antarctica	Funcional	S08	Caseina	9	60	[6]
SBcas3.3	Suelo	Funcional	S08	Azocaseína	9	50	[7]
Alkaline protease from O.M.6.2	Suelo	Funcional	S08	Caseina	8	37	[8]
pF1AL2	Sedimento marino	Funcional	S08	Caseina	10	80	[9]
Pro1437	Sedimento petrolero	Funcional	M48	Caseina	8	52	[10]
Prt1SU	Residuos de curtiembre	Secuencia	S01	Caseina	9	37	[11]
Prt1A	Barros activados	Funcional	S08	Caseina	11	55	[12]
PR4A3	Sedimento de manglar	Funcional	M42	Azocaseína	8,5	60	[13]

U/ml es el volumen de reacción total o el volumen de enzima usado. Es por ello que es imprescindible el desarrollo de un estándar de técnicas y unidades que permitan la comparación entre distintas proteasas.

Finalmente, se hizo un análisis cualitativo, con placas de agar-leche (1%), para corroborar la capacidad de la enzima de hidrolizar un sustrato natural. Como se observa en la Figura 9.4, la enzima es activa, pero en el futuro es necesario realizar una caracterización más profunda sobre la actividad de Pr10. Además de parámetros cinéticos y de termoestabilidad, es necesario determinar su capacidad hidrolítica y eficiencia con distintas matrices, como leche y WPC, y determinar los perfiles de péptidos generados.

CONCLUSIÓN

Fue posible identificar y caracterizar funcionalmente una proteasa de origen metagenómico, siguiendo una estrategia basada en secuencias. Si bien es necesario profundizar más en su caracterización funcional, la enzima Pr10 mostró actividad óptima a 50°C y pH ligeramente ácido, lo cual la convierte en una candidata de interés para la hidrólisis de leche y subproductos lácteos. Al igual que con la enzima βGal5, estudios adicionales, como de mutagénesis, pueden ayudar a mejorar la eficiencia de esta enzima novedosa.

CONCLUSIONES

10 CONCLUSIONES

El lactosuero es el principal subproducto de la industria láctea. Los enormes volúmenes producidos por año y su alta demanda química y biológica de oxígeno del mismo, debido a su alta carga orgánica, lo convierten en un potencial problema ambiental. Debido a su composición, principalmente lactosa y proteínas, el suero puede considerarse como materia prima para la elaboración de productos de mayor valor, mediante el uso de enzimas. Las enzimas llamadas β-galactosidasas son capaces de transformar la lactosa en galactooligosacáridos, que poseen actividad prebiótica. Por su parte, las proteasas son enzimas capaces de hidrolizar las proteínas del suero y liberar péptidos, que presentan una gran variedad de actividades, como antimicrobiana y antihipertensiva, entre otras.

Las lagunas de estabilización facultativas son la principal tecnología usada por pequeñas industrias lácteas para el tratamiento de sus efluentes, entre los cuales se pueden encontrar restos de suero. En estas lagunas, la comunidad microbiana es la encargada de estabilizar la carga orgánica de los efluentes, combinando procesos aeróbicos y anaeróbicos. Dado que la principal fuente de materia orgánica volcada en las lagunas de estabilización de industrias lácteas es el lactosuero, es esperable que miembros de su comunidad microbiana presenten β-galactosidasas y proteasas.

En el presente trabajo se plantea el estudio de las comunidades microbianas de dos sistemas de lagunas facultativas de dos pequeñas industrias lácteas del centro de la provincia de Santa Fe. Utilizando metagenómica basada en secuenciación y un enfoque centrado en genomas, se reconstruyeron 110 genomas microbianos completos, los cuales fueron caracterizados taxonómica y funcionalmente. Entre ellos se identificaron 7 grupos taxonómicos que serían claves para los sistemas facultativos, ya que fueron encontrados en ambos sistemas y presentan rutas metabólicas de interés que no fueron identificadas en otros organismos.

En la base de datos específica CAZy existen 5 familias descritas con actividad β-galactosidasa (EC: 3.2.1.23). Comparando los genes predichos en el metagenoma contra esta base de datos, se identificaron 379 β-galactosidasas candidatas, en 3 de las familias de interés (GH1, GH2, y GH42). Estos genes fueron clasificados tanto taxonómicamente como en familia y se seleccionaron 18 candidatos para su clonado y expresión. De ellos, 5 pudieron ser expresados y mostraron actividad con orto-nitrofenil-β-galactósido y lactosa. Entre estas 5 enzimas, βgal5 mostró una actividad superior a la del resto de las enzimas expresadas y a la de otras enzimas similares de origen metagenómico.

Por su parte, la base de datos MEROPS describe 4 familias de proteasas que han sido reportadas que poseen actividad sobre las proteínas del suero. Dentro de los metagenomas se identificaron 851 proteasas candidatas, en 3 de las familias de interés (C01, S01 y S08). Estos genes fueron clasificados tanto taxonómicamente como en familia y se seleccionaron 10 candidatos para su clonado y expresión. De ellos, 1 pudo ser expresado y mostró actividad con azocaseína y proteínas de leche.

El estudio de los miembros de una comunidad microbiana permite conocer con mayor profundidad el funcionamiento de los procesos metabólicos dentro de un ambiente y la identificación de microorganismos de importancia para ellos. Por otra parte, las enzimas de origen metagenómico poseen el potencial de modificar subproductos de distintas industrias, como la láctea. Este trabajo describe la combinación de estas estrategias para contribuir al mejor entendimiento del funcionamiento de las lagunas de estabilización, así como a la identificación y producción de enzimas para el aprovechamiento de subproductos y su transformación para agregar valor.

11 REFERENCIAS

Adrio, J. L., & Demain, A. L. (2014). Microbial enzymes: tools for biotechnological processes. Biomolecules, 4(1), 117–139.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature biotechnology, 31(6), 533-538.

Anguita-Ruiz, A., Aguilera, C. M., & Gil, Á. (2020). Genetics of lactose intolerance: an updated review and online interactive world maps of phenotype and genotype frequencies. *Nutrients*, *12*(9), 2689.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics, 36(7), 2251-2252.

Arp, D. J., Chain, P. S., & Klotz, M. G. (2007). The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. Annu. Rev. Microbiol., 61, 503-528.

Ayling, Martin, Matthew D. Clark, and Richard M. Leggett. "New approaches for metagenome assembly with short reads." Briefings in bioinformatics 21.2 (2020): 584-594.

Bah, T. (2007). Inkscape: guide to a vector drawing program. prentice hall press.

Baldasso, C., Barros, T. C., & Tessaro, I. C. (2011). Concentration and purification of whey proteins by ultrafiltration. In Desalination (Vol. 278, Issues 1-3, pp. 381–386).

Barrett, A. J., and McDonald, J. K. (1986). Nomenclature: protease, proteinase and peptidase. Biochem. J. 237:935. doi: 10.1042/bj2370935

Becerra, M., Prado, S. D., Siso, M. G., & Cerdán, M. E. (2001). New secretory strategies for Kluyveromyces lactis β-galactosidase. Protein Engineering, 14(5), 379-386.

Belila, A., Abbas, B., Fazaa, I., Saidi, N., Snoussi, M., Hassen, A., & Muyzer, G. (2013). Sulfur bacteria in wastewater stabilization ponds periodically affected by the ‘red-water phenomenon’. Applied microbiology and biotechnology, 97(1), 379-394.

Beja, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P., ... & DeLong, E. F. (2000). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environmental Microbiology, 2(5), 516-529.

Berini, F., Casciello, C., Marcone, G. L., & Marinelli, F. (2017). Metagenomics: novel enzymes from non-culturable microbes. FEMS microbiology letters, 364(21), fnx211.

Biver, S., Portetelle, D., & Vandenbol, M. (2013). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. Springerplus, 2(1), 1-10.

Boyer, P. D., Lardy, H., & Myrbäck, K. (1963). The Enzymes.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

Božanić, R., Barukčić, I., & Lisak, K. (2014). Possibilities of whey utilisation. Austin Journal of Nutrition and Food Sciences, 2(7), 7.

Brandelli, A., Daroit, D. J., & Corrêa, A. P. F. (2015). Whey as a source of peptides with remarkable biological activities. Food Research International , 73, 149–161.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., ... & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature, 523(7559), 208-211.

Buan, N. R. (2018). Methanogens: pushing the boundaries of biology. Emerging Topics in Life Sciences, 2(4), 629-646.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2015). DADA2: High resolution sample inference from amplicon data. BioRxiv, 024034.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., ... & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the national academy of sciences, 108(Supplement 1), 4516-4522.

Cardman, Z., Arnosti, C., Durbin, A., Ziervogel, K., Cox, C., Steen, A. D., & Teske, A. (2014). Verrucomicrobia are candidates for polysaccharide-degrading bacterioplankton in an arctic fjord of Svalbard. Applied and environmental microbiology, 80(12), 3749-3756.

Carvalho, F., Prazeres, A. R., & Rivas, J. (2013). Cheese whey wastewater: Characterization and treatment. Science of the total environment, 445, 385-396.

Castelle, C. J., Brown, C. T., Thomas, B. C., Williams, K. H., & Banfield, J. F. (2017). Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation. Scientific reports, 7(1), 1-12.

Catanzaro, R., Sciuto, M., & Marotta, F. (2021). Lactose intolerance: An update on its pathogenesis, diagnosis, and treatment. *Nutrition Research*, *89*, 23-34.

Chatzipaschali, A. A., & Stamatis, A. G. (2012). Biotechnological Utilization with a Focus on Anaerobic Treatment of Cheese Whey: Current Status and Prospects. In Energies (Vol. 5, Issue 9, pp. 3492–3525).

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.

Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS computational biology, 1(2), e24.

Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., ... & Huang, L. (2021). Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. Nature communications, 12(1), 1-13.

Cheng, J., Romantsov, T., Engel, K., Doxey, A. C., Rose, D. R., Neufeld, J. D., & Charles, T. C. (2017). Functional metagenomics reveals novel β-galactosidases not predictable from gene sequences. PloS one, 12(3), e0172545.

Corrochano, A. R., Buckin, V., Kelly, P. M., & Giblin, L. (2018). Invited review: Whey proteins as antioxidants and promoters of cellular antioxidant pathways. Journal of dairy science, 101(6), 4747-4761.

Coulier, L., Timmermans, J., Bas, R., Van Den Dool, R., Haaksman, I., Klarenbeek, B., Slaghek, T., & Van Dongen, W. (2009). In-depth characterization of prebiotic galacto-oligosaccharides by a combination of analytical techniques. Journal of Agricultural and Food Chemistry, 57(18), 8488–8495.

Cydzik-Kwiatkowska, A., & Zielińska, M. (2016). Bacterial communities in full-scale wastewater treatment systems. World Journal of Microbiology and Biotechnology, 32(4), 66.\nDavies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. Structure , 3(9), 853–859.

de Roode, B. M., Franssen, M. C. R., van der Padt, A., & Boom, R. M. (2003). Perspectives for the industrial enzymatic production of glycosides. Biotechnology Progress, 19(5), 1391–1402.

Deeth, H. C., & Bansal, N. (2018). Whey Proteins: From Milk to Medicine. Academic Press.

Devi, S. G., Fathima, A. A., Sanitha, M., Iyappan, S., Curtis, W. R., & Ramya, M. (2016). Expression and characterization of alkaline protease from the metagenomic library of tannery activated sludge. Journal of bioscience and bioengineering, 122(6), 694-700.

Devries, M. C., & Phillips, S. M. (2015). Supplemental protein in support of muscle mass and health: advantage whey. Journal of Food Science, 80 Suppl 1, A8–A15.

Di Lauro, B., Strazzulli, A., Perugino, G., La Cara, F., Bedini, E., Corsaro, M. M., ... & Moracci, M. (2008). Isolation and characterization of a new family 42 β-galactosidase from the thermoacidophilic bacterium Alicyclobacillus acidocaldarius: Identification of the active site residues. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1784(2), 292-301.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. Genome biology, 10(8), 1-16.

Drula, E., Garron, M. L., Dogan, S., Lombard, V., Henrissat, B., & Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. Nucleic acids research, 50(D1), D571-D577.\nEdgar, R. (2010). Usearch. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. In Genome Informatics 2009: Genome Informatics Series Vol. 23 (pp. 205-211).

Elleuche, S., Schroeder, C., Sahm, K., & Antranikian, G. (2014). Extremozymes—biocatalysts with unique properties from extremophilic microorganisms. Current opinion in biotechnology, 29, 116-123.

Entcheva, P., Liebl, W., Johann, A., Hartsch, T., & Streit, W. R. (2001). Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Applied and Environmental Microbiology, 67(1), 89-99.

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., ... & Willis, A. D. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. Nature microbiology, 6(1), 3-6.

Erich, S., Kuschel, B., Schwarz, T., Ewert, J., Böhmer, N., Niehaus, F., ... & Fischer, L. (2015). Novel high-performance metagenome β-galactosidases for lactose hydrolysis in the dairy industry. Journal of biotechnology, 210, 27-37.\nFalkenby, L. G., Szymanska, M., Holkenbrink, C., Habicht, K. S., Andersen, J. S., Miller, M., & Frigaard, N. U. (2011). Quantitative proteomics

of Chlorobaculum tepidum: insights into the sulfur metabolism of a phototrophic green sulfur bacterium. FEMS microbiology letters, 323(2), 142-150.

Farizoglu, B., Keskinler, B., Yildiz, E., & Nuhoglu, A. (2007). Simultaneous removal of C, N, P from cheese whey by jet loop membrane bioreactor (JLMBR). In Journal of Hazardous Materials (Vol. 146, Issues 1-2, pp. 399–407).

Fenton-May, R. I., Hill JR, C. G., & Amundson, C. H. (1971). Use of ultrafiltration/reverse osmosis systems for the concentration and fractionation of whey. Journal of Food Science, 36(1), 14-21.

Ferrer, M., Golyshina, O., Beloqui, A., & Golyshin, P. N. (2007). Mining enzymes from extreme environments. Current opinion in microbiology, 10(3), 207-214.

Ferrer, M., Martínez-Martínez, M., Bargiela, R., Streit, W. R., Golyshina, O. V., & Golyshin, P. N. (2016). Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microbial biotechnology, 9(1), 22-34.

Fox, P. F. (2013). Advanced Dairy Chemistry Volume 3: Lactose, water, salts and vitamins. Springer Science & Business Media.\nGänzle, M. G. (2012). Enzymatic synthesis of galacto-oligosaccharides and other lactose derivatives (hetero-oligosaccharides) from lactose. International Dairy Journal / Published in Association with the International Dairy Federation, 22(2), 116–122.

Gibson, G. R., & Roberfroid, M. B. (1995). Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. The Journal of nutrition, 125(6), 1401-1412.

Gibson, G. R., Probert, H. M., Van Loo, J., Rastall, R. A., & Roberfroid, M. B. (2004). Dietary modulation of the human colonic microbiota: updating the concept of prebiotics. Nutrition research reviews, 17(2), 259-275.

Gong, B. L., Mao, R. Q., Xiao, Y., Jia, M. L., Zhong, X. L., Liu, Y., ... & Li, G. (2017). Improvement of enzyme activity and soluble expression of an alkaline protease isolated from oil-polluted mud flat metagenome by random mutagenesis. Enzyme and microbial technology, 106, 97-105.

González Siso, M. I. (1996). The biotechnological utilization of cheese whey: A review. In Bioresource Technology (Vol. 57, Issue 1, pp. 1–11).

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International journal of systematic and evolutionary microbiology, 57(1), 81-91.

Gosling, A., Stevens, G. W., Barber, A. R., Kentish, S. E., & Gras, S. L. (2010). Recent advances refining galactooligosaccharide production from lactose. Food Chemistry, 121(2), 307–318.

Grady Jr, C. L., Daigger, G. T., Love, N. G., & Filipe, C. D. (2011). Biological wastewater treatment. CRC press.

Grand View Research. (2020). Enzymes Market Size, Share & Trends Analysis Report By Application (Industrial Enzymes, Specialty Enzymes), By Product (Carbohydrase, Proteases, Lipases), By Source, By Region, And Segment Forecasts, 2020 - 2027.

Guo, J., Li, J., Chen, H., Bond, P. L., & Yuan, Z. (2017). Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. Water research, 123, 468-478.

Gurumallesh, P., Alagu, K., Ramakrishnan, B., & Muthusamy, S. (2019). A systematic reconsideration on proteases. International journal of biological macromolecules, 128, 254-267.\nHandelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews, 68(4), 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology, 5(10), R245-R249.

Hannig, G., & Makrides, S. C. (1998). Strategies for optimizing heterologous protein expression in Escherichia coli. Trends in biotechnology, 16(2), 54-60.

Henne, A., Daniel, R., Schmitz, R. A., & Gottschalk, G. (1999). Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Applied and Environmental Microbiology, 65(9), 3901-3907.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochemical Journal, 280 (Pt 2), 309–316.

Henrissat, B., & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. Biochemical Journal, 316 (Pt 2), 695–696.

Herbert, R. A. (1992). A perspective on the biotechnological potential of extremophiles. Trends in biotechnology, 10, 395-402.

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., ... & Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science, 331(6016), 463-467.

Holsinger, V. H., Posati, L. P., & DeVilbiss, E. D. (1974). Whey beverages: a review. Journal of Dairy Science, 57(8), 849-859.

Huang, Y., Zou, K., Qing, T., Feng, B., & Zhang, P. Metagenomics and Metatranscriptomics Insights into Antibiotic Synthesis in Activated Sludge. Available at SSRN 4083969.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... & Banfield, J. F. (2016). A new view of the tree of life. Nature microbiology, 1(5), 1-6.

Huh, K. T., Toba, T., & Adachi, S. (1991). Oligosaccharide structures formed during acid hydrolysis of lactose. Food Chemistry, 39(1), 39–49.

Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic biology, 61(6), 1061-1067.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 1-11.\nIsmail, B., & Gu, Z. (2010). Whey protein hydrolysates: Current knowledge and challenges. Midwest Dairy Foods Research Center, 55–79.\nJelen, P. (2009). Whey-based functional beverages. In Functional and speciality beverage technology (pp. 259-280). Woodhead Publishing.

Jankowski, P., Gan, J., Le, T., McKennitt, M., Garcia, A., Yanaç, K., ... & Uyaguari-Diaz, M. (2022). Metagenomic community composition and resistome analysis in a full-scale cold climate wastewater treatment plant. Environmental microbiome, 17(1), 1-20.

Jelen, P., & Buchheim, W. (1976). Norwegian whey cheese. Food Technology, 30(11), 62.

Jeličić, I., Božanić, R., & Tratnik, L. (2008). Whey-based beverages-a new generation of dairy products. Mljekarstvo, 58(3), 257-274.

Juliano, P., Muset, G. B., & Castells, M. L. (2017). Valorización del Lactosuero. Instituto Nacional de Tecnología Industrial. Compilado por Muset, GB 1a Ed. San Martín, Argentina ISBN, 978-950.\nKanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic acids research, 49(D1), D545-D551.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

Kirk, O., Borchert, T. V., & Fuglsang, C. C. (2002). Industrial enzyme applications. Current opinion in biotechnology, 13(4), 345-351.

Kitts, D. D., & Weiler, K. (2003). Bioactive proteins and peptides from food sources. Applications of bioprocesses used in isolation and recovery. Current Pharmaceutical Design, 9(16), 1309–1323.

Khorshid, M. A. (1974). Studies on the permeate from the ultrafiltration of whey. Australian Journal of Dairy Technology, 29(1), 37.

Korhonen, H. (2009). Milk-derived bioactive peptides: From science to applications. Journal of Functional Foods, 1(2), 177–187.

Korhonen, H., & Pihlanto, A. (2003). Food-derived bioactive peptides--opportunities for designing future foods. Current Pharmaceutical Design, 9(16), 1297–1308.

Kosikowski, F. V. (1968). Nutritional beverages from acid whey powder. Journal of dairy science, 51, 1299-1301.

Kosikowski, F. V. (1979). Whey utilization and whey products. Journal of Dairy Science, 62(7), 1149-1160.\nLaemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. nature, 227(5259), 680-685.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.

Lee, D. G., Jeon, J. H., Jang, M. K., Kim, N. Y., Lee, J. H., Lee, J. H., ... & Lee, S. H. (2007). Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. Biotechnology letters, 29(3), 465-472.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-2993.

Liu, X., & Kokare, C. (2017). Chapter 11 - Microbial Enzymes of Use in Industry. In G. Brahmachari (Ed.), Biotechnology of Microbial Enzyme

Liu, P., Wang, W., Zhao, J., & Wei, D. (2019). Screening novel β-galactosidases from a sequence-based metagenome and characterization of an alkaline β-galactosidase for the enzymatic synthesis of galactooligosaccharides. Protein expression and purification, 155, 104-111.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research, 42(D1), D490-D495.

Lorenz, P., & Eck, J. (2005). Metagenomics and industrial applications. Nature Reviews Microbiology, 3(6), 510-516.\nMacfarlane, G. T., Blackett, K. L., Nakayama, T., Steed, H., & Macfarlane, S. (2009). The gut microbiota in inflammatory bowel disease. Current pharmaceutical design, 15(13), 1528-1536.

Madureira, A. R., Tavares, T., Gomes, A. M. P., Pintado, M. E., & Malcata, F. X. (2010). Invited review: physiological properties of bioactive peptides obtained from whey proteins. Journal of dairy science, 93(2), 437-455.

Mahoney, R. R. (1998). Galactosyl-oligosaccharide formation during lactose hydrolysis: A review. Food Chemistry, 63(2), 147–154.

Malaspina, F., Cellamare, C. M., Stante, L., & Tilche, A. (1996). Anaerobic treatment of cheese whey with a downflow-upflow hybrid reactor. In Bioresource Technology (Vol. 55, Issue 2, pp. 131–139).

Maldonado, J., Gil, A., Narbona, E., & Molina, J. A. (1998). Special formulas in infant nutrition: a review. Early human development, 53, S23-S32.

Mawson, A. J. (1994). Bioconversions for whey utilization and waste abatement. In Bioresource Technology (Vol. 47, Issue 3, pp. 195–203).

McDonald, J. K. (1985). An overview of protease specificity and catalytic mechanisms: aspects related to nomenclature and classification. The Histochemical Journal, 17(7), 773-785.

McGarvey, J. A., Miller, W. G., Sanchez, S. U. S. A. N., Silva, C. J., & Whitehand, L. C. (2005). Comparison of bacterial populations and chemical composition of dairy wastewater held in circulated and stagnant lagoons. Journal of applied microbiology, 99(4), 867-877.

McGarvey, J. A., Miller, W. G., Zhang, R., Ma, Y., & Mitloehner, F. (2007). Bacterial population dynamics in dairy waste during aerobic and anaerobic treatment and subsequent storage. Applied and environmental microbiology, 73(1), 193-202.

McMahon, K. (2015). Metagenomics 2.0. Environmental microbiology reports, 7(1), 38-39.

Merz, M., Eisele, T., Berends, P., Appel, D., Rabe, S., Blank, I., ... & Fischer, L. (2015). Flavourzyme, an enzyme preparation with industrial relevance: automated nine-step purification and partial characterization of eight enzymes. Journal of Agricultural and Food Chemistry, 63(23), 5682-5693.

Militon, C., Hamdi, O., Michotey, V., Fardeau, M. L., Ollivier, B., Bouallagui, H., ... & Bonin, P. (2015). Ecological significance of Synergistetes in the biological treatment of tuna cooking wastewater by an anaerobic sequencing batch reactor. Environmental Science and Pollution Research, 22(22), 18230-18238.

Miller, J. (1978). H.(1972) Experiments in molecular genetics. ColdSpring Harbor Laboratory. Cold Spring Harbor, NY, 328-330.

Mishra, S. S., Ray, R. C., Rosell, C. M., & Panda, D. (2017). Microbial enzymes in food applications: history of progress. In Microbial enzyme technology in food applications (pp. 3-18). CRC Press.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., ... & Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic acids research, 49(D1), D412-D419.

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., ... & Finn, R. D. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic acids research, 46(D1), D726-D735.

Mockaitis, G., Ratusznei, S. M., Rodrigues, J. A. D., Zaiat, M., & Foresti, E. (2006). Anaerobic whey treatment by a stirred sequencing batch reactor (ASBR): effects of organic loading and

supplemented alkalinity. In Journal of Environmental Management (Vol. 79, Issue 2, pp. 198–206).

Mollea, C., Marmo, L., & Bosco, F. (2013). Valorisation of cheese whey, a by-product from the dairy industry. In Food industry. IntechOpen.

Montoya, L., Celis, L. B., Razo-Flores, E., & Alpuche-Solís, Á. G. (2012). Distribution of CO 2 fixation and acetate mineralization pathways in microorganisms from extremophilic anaerobic biotopes. Extremophiles, 16(6), 805-817.

Morris, L. S., & Marchesi, J. R. (2015). Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. Current microbiology, 70(1), 19-26.

Mulualem, D. M., Agbavwe, C., Ogilvie, L. A., Jones, B. V., Kilcoyne, M., O'Byrne, C., & Boyd, A. (2021). Metagenomic identification, purification and characterisation of the Bifidobacterium adolescentis BgaC β-galactosidase. Applied microbiology and biotechnology, 105(3), 1063-1078.\nNelson, W. C., & Stegen, J. C. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. Frontiers in microbiology, 6, 713.

Neveu, J., Regeard, C., & DuBow, M. S. (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. Applied Microbiology and Biotechnology, 91(3), 635-644.

Ngara, T. R., & Zhang, H. (2018). Recent advances in function-based metagenomic screening. Genomics, proteomics & bioinformatics, 16(6), 405-415.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution, 32(1), 268-274.

Nguyen N.H., Maruset L., Uengwetwanit T., Mhuantong W., Harnpicharnchai .P, Champreda V., Tanapongpipat S., Jirajaroenrat K., Rakshit S.K., Eurwilaichitr L., Pongpattanakitshote S. (2012) Identification and characterization of a cellulase-encoding gene from the buffalo rumen metagenomic library. Biosci Biotechnol Biochem 76:1075–1084

Nongonierma, A. B., & FitzGerald, R. J. (2015). Milk proteins as a source of tryptophan-containing bioactive peptides. Food & Function, 6(7), 2115–2127.\nO'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), D733-D745.

Oecd, OECD, & Food and Agriculture Organization of the United Nations. (2020). OECD-FAO Agricultural Outlook 2020-2029. In OECD-FAO Agricultural Outlook.

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986). Microbial ecology and evolution: a ribosomal RNA approach. Annual reviews in microbiology, 40(1), 337-365.\ nPaques, M., & Lindner, C. (2019). Lactose: Evolutionary Role, Health Effects, and Applications. Academic Press.

Park, A. R., & Oh, D. K. (2010). Effects of galactose and glucose on the hydrolysis reaction of a thermostable β-galactosidase from Caldicellulosiruptor saccharolyticus. Applied microbiology and biotechnology, 85(5), 1427-1435.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, 25(7), 1043-1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., ... & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology, 2(11), 1533-1542.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology, 36(10), 996-1004.

Parodi, P. W. (2007). A role for milk proteins and their peptides in cancer prevention. Current Pharmaceutical Design, 13(8), 813–828.

Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 28(11), 1420-1428.

Pessoa, T. B., Rezende, R. P., Marques, E. D. L. S., Pirovani, C. P., Dos Santos, T. F., dos Santos Gonçalves, A. C., ... & Dias, J. C. (2017). Metagenomic alkaline protease from mangrove sediment. Journal of basic microbiology, 57(11), 962-973.

Pintado, M. E., Macedo, A. C., & Malcata, F. X. (2001). Technology, chemistry and microbiology of whey cheeses. Food Science and Technology International, 7(2), 105-116.

Playne, M. J., & Crittenden, R. G. (2009). Galacto-oligosaccharides and Other Products Derived from Lactose. In P. McSweeney & P. F. Fox (Eds.), Advanced Dairy Chemistry: Volume 3: Lactose, Water, Salts and Minor Constituents (pp. 121–201). Springer New York.

Prayogo, F. A., Budiharjo, A., Kusumaningrum, H. P., Wijanarka, W., Suprihadi, A., & Nurhayati, N. (2020). Metagenomic applications in exploration and development of novel enzymes from nature: a review. Journal of Genetic Engineering and Biotechnology, 18(1), 1-10.

Prazeres, A. R., Carvalho, F., & Rivas, J. (2012). Cheese whey management: A review. Journal of environmental management, 110, 48-68.

Provansal, M. M., Cuq, J. L., & Cheftel, J. C. (1975). Chemical and nutritional modifications of sunflower proteins due to alkaline processing. Formation of amino acid crosslinks and isomerization of lysine residues. Journal of agricultural and food chemistry, 23(5), 938-943.

Purohit, M. K., & Singh, S. P. (2013). A metagenomic alkaline protease from saline habitat: cloning, over-expression and functional attributes. International journal of biological macromolecules, 53, 138-143.

Pushpam, P. L., Rajesh, T., & Gunasekaran, P. (2011). Identification and characterization of alkaline serine protease from goat skin surface metagenome. AMB express, 1(1), 1-10.\nR Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rao, M. B., Tanksale, A. M., Ghatge, M. S., & Deshpande, V. V. (1998). Molecular and biotechnological aspects of microbial proteases. Microbiology and molecular biology reviews, 62(3), 597-635.

Rastall, R. A. (2006). Galacto-oligosaccharides as prebiotic food ingredients. Prebiotics: Development and Applications, New York, John Wiley, 101–110.

Rawlings, N. D., & Barrett, A. J. (1993). Evolutionary families of peptidases. Biochemical Journal, 290(1), 205-218.

Rawlings, N. D., Barrett, A. J., & Finn, R. (2016). Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic acids research, 44(D1), D343-D350.

Razzaq, A., Shamsi, S., Ali, A., Ali, Q., Sajjad, M., Malik, A., & Ashraf, M. (2019). Microbial proteases applications. Frontiers in bioengineering and biotechnology, 7, 110.

Reddy, T. B., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., ... & Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification. Nucleic acids research, 43(D1), D1099-D1106.

Remmas, N., Melidis, P., Zerva, I., Kristoffersen, J. B., Nikolaki, S., Tsiamis, G., & Ntougias, S. (2017). Dominance of candidate Saccharibacteria in a membrane bioreactor treating medium age landfill leachate: Effects of organic load on microbial communities, hydrolytic potential and extracellular polymeric substances. Bioresource technology, 238, 48-56.

Research Nester. (2021). Galacto-oligosaccharide Market Segmentation By End User (Food & Beverage Industry, Pharmaceutical, Personal Care and Animal Feed) – Global Demand Analysis & Opportunity Outlook 2028 (No. 2814). https://www.researchnester.com/reports/galacto-oligosaccharide-market/2814

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., ... & Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature, 499(7459), 431-437.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014a). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics, 30(5), 629-635.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014b). Bypassing cultivation to identify bacterial species. Microbe, 9(3), 111-118.

Ryan, M. P., & Walsh, G. (2016). The biotechnological potential of whey. Reviews in Environmental Science and Biotechnology, 15(3), 479–498.\nSambrook, J., & Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol: chloroform. Cold Spring Harbor Protocols, 2006(1), pdb-prot4455.

Saranraj, P., & Naidu, M. A. (2014). Microbial pectinases: a review. Global J Trad Med Syst, 3(1), 1–9.

Sanchez, S., & Demain, A. L. (2017). Useful microbial enzymes—an introduction. In Biotechnology of microbial enzymes (pp. 1-11). Academic Press.

Schingoethe, D. J. (1976). Whey Utilization in Animal Feeding: A Summary and Evaluation. In Journal of Dairy Science (Vol. 59, Issue 3, pp. 556–570).

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., ... & McHardy, A. C. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nature methods, 14(11), 1063-1071.

Seviour, R. J., Kragelund, C., Kong, Y., Eales, K., Nielsen, J. L., & Nielsen, P. H. (2008). Ecophysiology of the Actinobacteria in activated sludge systems. Antonie Van Leeuwenhoek, 94(1), 21-33.

Seyedi, S., Venkiteshwaran, K., Benn, N., & Zitomer, D. (2020). Inhibition during anaerobic co-digestion of aqueous pyrolysis liquid from wastewater solids and synthetic primary sludge. Sustainability, 12(8), 3441.

Sienkiewicz, T., & Riedel, C. L. (1990). Whey and whey utilization: possibilities for utilization in agriculture and foodstuffs production, Verlag Th. Mann, Gelsenkirchen-Buer, Germany.

Sinha, R., Radha, C., Prakash, J., & Kaul, P. (2007). Whey protein hydrolysate: Functional properties, nutritional quality and utilization in beverage formulation. Food Chemistry, 101(4), 1484-1491.

Smithers, G. W. (2008). Whey and whey proteins—From “gutter-to-gold.” In International Dairy Journal (Vol. 18, Issue 7, pp. 695–704).

Staden, R., Beal, K. F., & Bonfield, J. K. (2000). The staden package, 1998. In Bioinformatics methods and protocols (pp. 115-130). Humana Press, Totowa, NJ.

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual review of microbiology, 39(1), 321-346.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313.

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics–the key to the uncultured microbes. Current opinion in microbiology, 7(5), 492-498.

Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers in microbiology, 3, 410.

Sun, L., Pope, P. B., Eijsink, V. G., & Schnürer, A. (2015). Characterization of microbial community structure during continuous anaerobic digestion of straw and cow manure. Microbial biotechnology, 8(5), 815-827.

Sun, J., Li, P., Liu, Z., Huang, W., & Mao, X. (2020). A novel thermostable serine protease from a metagenomic library derived from marine sediments in the East China Sea. Applied Microbiology and Biotechnology, 104(21), 9229-9238.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., ... & Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of the National Academy of Sciences, 110(28), 11463-11468.

Sysoev, M., Grötzinger, S. W., Renn, D., Eppinger, J., Rueping, M., & Karan, R. (2021). Bioprospecting of Novel Extremozymes From Prokaryotes—The Advent of Culture-Independent Methods. Frontiers in microbiology, 12.\nTalens-Perales, D., Górska, A., Huson, D. H., Polaina, J., & Marín-Navarro, J. (2016). Analysis of Domain Architecture and Phylogenetics of Family 2 Glycoside Hydrolases (GH2). PloS One, 11(12), e0168035.

Tavano, O. L. (2013). Protein hydrolysis using proteases: An important tool for food biotechnology. Journal of Molecular Catalysis B: Enzymatic, 90, 1-11

Tavano, O. L., Berenguer-Murcia, A., Secundo, F., & Fernandez-Lafuente, R. (2018). Biotechnological applications of proteases in food technology. Comprehensive Reviews in Food Science and Food Safety, 17(2), 412-436.

Torres, D. P. M., Gonçalves, M. do P. F., Teixeira, J. A., & Rodrigues, L. R. (2010). Galacto-oligosaccharides: Production, properties, applications, and significance as prebiotics. Comprehensive Reviews in Food Science and Food Safety, 9(5), 438–454.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. Current opinion in microbiology, 11(5), 442-446.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., ... & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 428(6978), 37-43.

Tzortzis, G., & Vulevic, J. (2009). Galacto-Oligosaccharide Prebiotics (p. 207).\nUrashima, T., Suyama, K., & Adachi, S. (1988). The condensation of 5-(hydroxymethyl)-2-furfuraldehyde with some aldoses on heating. In Food Chemistry (Vol. 29, Issue 1, pp. 7–17).\nVan Loo, J., Cummings, J., Delzenne, N., Englyst, H., Franck, A., Hopkins, M., Kok, N., Macfarlane, G., Newton, D., Quigley, M., Roberfroid, M., van Vliet, T., & van den Heuvel, E. (1999). Functional food properties of non-digestible oligosaccharides: a consensus report from the ENDO project (DGXII AIRII-CT94-1095). The British Journal of Nutrition, 81(2), 121–132.

Vester, J. K., Glaring, M. A., & Stougaard, P. (2014). Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. Microbial cell factories, 13(1), 1-14.

Verma, S. K., & Sharma, P. C. (2021). Isolation and biochemical characterization of a novel serine protease identified from solid tannery waste metagenome. Biologia, 76(10), 3163-3174.

Vocadlo, D. J., & Davies, G. J. (2008). Mechanistic insights into glycosidase chemistry. Current Opinion in Chemical Biology, 12(5), 539–555.

von Sperling, M. (2007). Waste Stabilisation Ponds. IWA Publishing.\nWalstra, P., Walstra, P., Wouters, J. T. M., & Geurts, T. J. (2005). Dairy Science and Technology. CRC Press.

Wang, K., Li, G., Yu, S. Q., Zhang, C. T., & Liu, Y. H. (2010). A novel metagenome-derived β-galactosidase: gene cloning, overexpression, purification and characterization. Applied microbiology and biotechnology, 88(1), 155-165.

Waschkowitz, T., Rockstroh, S., & Daniel, R. (2009). Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Applied and environmental microbiology, 75(8), 2506-2516.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.

Woese, C. R. (1987). Bacterial evolution. Microbiological reviews, 51(2), 221-271.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome biology, 20(1), 1-13.

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., ... & Banfield, J. F. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science, 337(6102), 1661-1665.

Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics, 32(4), 605-607.\nYang, H., Huang, X., Fang, S., Xin, W., Huang, L., & Chen, C. (2016). Uncovering the composition of microbial community structure and metagenomics among three gut locations in pigs with distinct fatness. Scientific reports, 6(1), 1-11.

Yang, S. T., & Silva, E. M. (1995). Novel products and new technologies for use of a familiar carbohydrate, milk lactose. Journal of Dairy Science, 78(11), 2541-2562.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic acids research, 40(W1), W445-W451.

Youngblut, N. D., De la Cuesta-Zuluaga, J., Reischer, G. H., Dauser, S., Schuster, N., Walzer, C., ... & Ley, R. E. (2020). Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. Msystems, 5(6), e01045-20.\nZehr, J. P., Jenkins, B. D., Short, S. M., & Steward, G. F. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environmental microbiology, 5(7), 539-554.

Zhang, X., Li, H., Li, C. J., Ma, T., Li, G., & Liu, Y. H. (2013). Metagenomic approach for the isolation of a thermostable β-galactosidase with high tolerance of galactose and glucose from soil samples of Turpan Basin. BMC microbiology, 13(1), 1-10.

Zhang, Y., Zhao, J., & Zeng, R. (2011). Expression and characterization of a novel mesophilic protease from metagenomic library derived from Antarctic coastal sediment. Extremophiles, 15(1), 23-29.

Zhou, L., Wang, S., Zou, Y., Xia, C., & Zhu, G. (2015). Species, abundance and function of ammonia-oxidizing archaea in inland waters across China. Scientific reports, 5(1), 1-13.

12 TABLAS SUPLEMENTARIAS

Tabla S1: Parámetros fisicoquímicos obtenidos de las 6 lagunas muestreadas

Muestra	pH	Sólidos totales (mg/l)	Sulfuro (mg/l)	Fósforo total (mg/l)	DQO (mg/l)	DBO (mg/l)	Subst. Sol. en eter (mg/l)	Conductividad (mS/cm)	Temp. (ºC)	Oxígeno disuelto (mg/l)	Fecha (dd/mm/aa)
AUR1	8,96	12616	254,2	3,6	4938	2524	229,2	10,4	22,9	4,24	05/09/13
AUR2	9.23	11574	75,6	2,6	2928	1827	66,4	11,3	20,5	2,69	05/09/13
CYC1	6,59	4726	1,75	27,9	1714	1192	147,2	6,01	20,08	4,36	07/04/13
CYC2	7	5074	2,1	14,3	1448	1093	39,2	6,8	19,25	4,88	07/04/13
CYC3	7,34	5240	17,5	23,3	1181	926	27,6	7,08	19,67	3,75	07/04/13
CYC4	7,55	5242	13,5	17,8	1290	775	62,4	7,04	19,39	6,01	07/04/13

Table S2. Módulos y familias de genes de KEGG buscados

KEGG ID	Metabolic Process	Description
K01026	Acetogenesis	Acetil-CoA -> acetato
K01067	Acetogenesis	Acetil-CoA -> acetato
K01905+K22224	Acetogenesis	Acetil-CoA -> acetato
K18118	Acetogenesis	Acetil-CoA -> acetato
K24012	Acetogenesis	Acetil-CoA -> acetato

M00307	Acetogenesis	Oxidación de piruvato a acetil-CoA
(K00248,K00249)+K01692+(K00020,K01782)+K00140	Acidogenesis	Val → propanoil-CoA
(K00248,K00249)+K01692+K01782+K00632	Acidogenesis	Ile → propanoil-CoA+Acetil-CoA
K00814+K14260+K14272	Acidogenesis	Ala → piruvato
K00826+(K00166+K00167,K11381)+K00382+K09699	Acidogenesis	Degradación de aminoácidos ramificados
K00827+K00830+K14272	Acidogenesis	Ala → piruvato
K01752	Acidogenesis	Ser → piruvato + NH_3
K01754	Acidogenesis	Ser → piruvato + NH_3
K17217+K01758	Acidogenesis	Cys → piruvato
K17989	Acidogenesis	Ser → piruvato + NH_3
M00001	Acidogenesis	Glicólisis
M00032	Acidogenesis	Degradación de lisina
M00087	Acidogenesis	Beta-oxidación
K00260+K00261+K00262	Amonificación	Glu→ amoníaco
K00281+K00282+K00283+K00605	Amonificación	Gly→ amoníaco
K005597	Amonificación	Gln→ amoníaco
K01425	Amonificación	Gln→ amoníaco
K01955+K01956	Amonificación	Gln→ amoníaco

K23265	Amonificación	Gln→ amoníaco
M00045	Amonificación	His → Glu
M00165	Fijación de carbono	Ciclo reductivo de pentosa fosfato (ciclo de Calvin)
M00166	Fijación de carbono	Ciclo reductivo de pentosa fosfato, ribulosa-5P => gliceraldehido-3P
M00167	Fijación de carbono	Ciclo reductivo de pentosa fosfato, gliceraldehido-3P => ribulosa-5P
M00173	Fijación de carbono	Ciclo reductivo del citrato (Arnon-Buchanan cycle)
M00374	Fijación de carbono	Ciclo del dicarboxilato-hidroxibutirato
M00375	Fijación de carbono	Ciclo de dihidroxipropionato-hidroxibutirato
M00376	Fijación de carbono	Bi-ciclo de 3-Hidroxipropionato
M00377	Fijación de carbono	Ruta de reducción de Acetil-CoA (Wood-Ljungdahl pathway)
M00579	Fijación de carbono	Ruta de fosfato Acetiltransferase-acetato kinasa, Acetil-CoA => acetato
M00529	Denitrificación	Denitrificación, nitrato => nitrógeno
M00356	Metanogénesis	Metanogénesis a partir de metanol
M00357	Metanogénesis	Metanogénesis acetoclastica
M00563	Metanogénesis	Metanogénesis a partir de metilamina
M00567	Metanogénesis	Metanogénesis hidrogenotrófica

M00530	Reducción de nitrato	Reducción disimilatoria de nitrato, nitrato => amoníaco
M00531	Reducción de nitrato	Reducción asimilatoria de nitrato, nitrato => amoníaco
M00528	Nitrificación	Nitrificación, amoníaco => nitrito
M00175	Fijación de nitrógeno	Fijación de nitrógeno, nitrógeno => amoníaco
K00023	Sintesis de PHB	Aceto-acetil-CoA reductasa
K00626	Sintesis de PHB	Acetil-CoA acetiltransferasa
K03821	Sintesis de PHB	Polihidroxialcanoato sintetasa
K22881	Sintesis de PHB	Polihidroxialcanoato sintetasa
M00161	Fotosintesis	Fotosistema II
M00597	Fotosintesis	Fotosistema II anoxigénico [BR:ko00194]
M00176	Metabolismo de sulfuro	Reducción asimilatoria de sulfato, sulfato => H_2S
M00595	Metabolismo de sulfuro	Oxidación de tiosulfato por complejo SOX, tiosulfato => sulfato
M00596	Metabolismo de sulfuro	Reducción disimilatoria de sulfato, sulfato => H_2S

Tabla S3: Resultados del proceso de *binning*. Estadísticas genómicas para cada bin reconstruido. La cobertura (Cob) de cada *bin* fue calculada como la suma de las coberturas de los *contigs* en dicho *bin*, dividido por el número de *contigs* en el *bin*. Los valores fueron normalizados para cada muestra, dividiendo la cobertura obtenida por el número de lecturas en cada set de datos, dividido por 1 millón. Comp: completitud; Cont: contaminación; HC: heterogeneidad de cepa; CML: Cobertura cada millón de lecturas

Bin	Tamaño	Comp	Cont	HC	Status	Genes	CML muestra 1	CML muestra 2	CML muestra 3	CML muestra 4
INTA.AUR.001	173,744	0	0	0	Incompleto	231	2.01E-01	8.51E-01		
INTA.AUR.002	3,204,701	95.54	0.79	50	Alta calidad	3234	1.07E-01	4.25E-02		
INTA.AUR.003	4,773,275	90.71	6.34	0	En borrador	4046	9.85E-02	3.99E-02		
INTA.AUR.004	3,466,087	90.73	3.94	11.11	Alta calidad	3478	8.99E-02	3.09E-02		
INTA.AUR.005	2,940,775	91.18	2.17	37.5	Alta calidad	3241	8.00E-02	4.57E-01		
INTA.AUR.006	928,382	57.32	0.14	100	En borrador	1063	6.36E-02	3.95E-02		
INTA.AUR.007	3,104,077	93.41	2.73	28.57	Alta calidad	2843	5.17E-02	3.19E-02		
INTA.AUR.008	712,339	64.5	0.85	0	En borrador	762	4.34E-02	7.47E-03		
INTA.AUR.009	2,697,931	90.42	0.92	100	Alta calidad	2927	4.80E-02	1.48E-01		
INTA.AUR.010	2,245,734	84.1	2.46	20	En borrador	2245	3.36E-02	1.50E-02		
INTA.AUR.011	2,497,937	93.43	4.13	26.67	Alta calidad	2855	3.57E-02	9.89E-03		
INTA.AUR.012	2,535,658	89.34	35.11	9.76	Incompleto	2840	2.78E-02	1.30E-02		
INTA.AUR.013	950,343	87.27	14.55	0	En borrador	932	3.21E-02	1.32E-02		

INTA.AUR.014	2,685,231	100	3.23	0	Alta calidad	2632	2.43E-02	7.63E-03
INTA.AUR.015	1,756,667	86.21	41.58	3.53	En borrador	2043	2.53E-02	7.06E-03
INTA.AUR.016	1,323,050	14.12	5.42	90.91	Incompleto	1573	2.01E-02	1.56E-02
INTA.AUR.017	3,512,283	80.19	1.89	0	En borrador	3588	1.73E-02	8.17E-03
INTA.AUR.018	1,410,640	69.44	8.79	23.08	Incompleto	1784	1.82E-02	8.99E-03
INTA.AUR.019	967,597	63.85	26.27	89.47	Incompleto	1199	1.77E-02	1.38E-02
INTA.AUR.020	2,294,909	60.28	25.61	69.51	Incompleto	2700	1.65E-02	1.37E-02
INTA.AUR.021	3,212,545	89.76	7.05	6.25	En borrador	3362	1.50E-02	6.81E-03
INTA.AUR.022	1,114,822	28.34	17.93	86.96	Incompleto	1370	1.36E-02	3.03E-03
INTA.AUR.023	637,478	67.04	8.42	0	En borrador	655	1.33E-02	7.92E-03
INTA.AUR.024	827,426	69.27	5.62	0	En borrador	915	1.36E-02	8.09E-03
INTA.AUR.025	1,911,535	4.17	0	0	Incompleto	2380	8.49E-03	1.78E-02
INTA.AUR.026	3,280,597	95.47	3.36	0	Alta calidad	3251	1.23E-02	3.30E-03
INTA.AUR.027	2,940,801	96.38	1.7	0	Alta calidad	3145	1.08E-02	4.05E-03
INTA.AUR.028	2,644,959	100	0.31	0	Alta calidad	3141	9.88E-03	3.47E-03
INTA.AUR.029	875,167	16.7	0	0	Incompleto	1114	1.06E-02	8.03E-03
INTA.AUR.030	4,234,589	94.05	16.31	3.12	En borrador	4217	1.05E-02	5.22E-03
INTA.AUR.031	1,792,188	81.04	1.74	11.11	En borrador	1922	9.85E-03	6.58E-03

INTA.AUR.032	2,727,084	93.1	6.74	0	En borrador	2779	6.93E-03	1.43E-03
INTA.AUR.033	1,512,753	55.48	3.27	75	Incompleto	1993	8.01E-03	1.28E-03
INTA.AUR.034	3,778,965	96.77	7.28	4.55	En borrador	3348	7.51E-03	2.86E-03
INTA.AUR.035	2,611,648	82.02	3.32	62.5	En borrador	2875	7.21E-03	2.94E-03
INTA.AUR.036	3,252,935	99.41	10.28	68.42	En borrador	3490	5.78E-03	2.93E-03
INTA.AUR.037	4,650,668	87.71	23.49	4.92	Incompleto	4872	5.53E-03	1.65E-03
INTA.AUR.038	1,602,964	80.56	11.86	0	En borrador	2228	3.77E-03	3.74E-03
INTA.AUR.039	3,830,360	87.67	19.51	6.67	En borrador	4070	4.44E-03	3.45E-03
INTA.AUR.040	2,072,076	75.44	7.02	20	En borrador	2321	5.45E-03	2.13E-03
INTA.AUR.041	831,985	3.76	0	0	Incompleto	928	5.28E-03	2.24E-03
INTA.AUR.042	4,095,300	84.48	13.98	0	En borrador	4298	6.01E-03	8.24E-03
INTA.AUR.043	2,185,781	28.65	5.81	0	Incompleto	2925	4.02E-03	2.31E-03
INTA.AUR.044	2,178,504	88.76	3.31	0	En borrador	2316	4.86E-03	1.82E-03
INTA.AUR.045	3,069,974	97.49	52.99	15.25	En borrador	3747	4.13E-03	1.20E-03
INTA.AUR.046	2,223,342	7.97	0	0	Incompleto	2662	3.88E-03	2.74E-03
INTA.AUR.047	3,171,438	81.68	47.37	44.23	En borrador	3895	4.33E-03	2.11E-03
INTA.AUR.048	2,969,165	59.18	18.6	48.78	Incompleto	3667	4.08E-03	1.72E-03
INTA.AUR.049	3,878,925	28.9	11.23	17.54	Incompleto	5191	3.92E-03	2.17E-03

INTA.AUR.050	3,538,557	87.93	51.81	1.32	Incompleto	4667	3.96E-03	1.42E-03
INTA.AUR.051	2,001,640	87.78	3.69	18.18	En borrador	2219	4.06E-03	2.60E-03
INTA.AUR.052	230,781	25.17	10.84	14.81	Incompleto	321	4.03E-03	1.17E-03
INTA.AUR.053	151,985	3.92	0	0	Incompleto	194	3.76E-03	1.20E-03
INTA.AUR.054	2,353,015	58.86	17.24	0	Incompleto	3115	3.88E-03	1.42E-03
INTA.AUR.055	3,107,757	88.98	15.21	0	En borrador	3780	3.53E-03	2.67E-03
INTA.AUR.056	486,137	18.73	0	0	Incompleto	669	3.71E-03	8.69E-04
INTA.AUR.057	3,674,302	59.75	18.13	0	Incompleto	4693	3.43E-03	1.81E-03
INTA.AUR.058	3,079,956	67.45	40.75	0	Incompleto	4459	3.20E-03	1.39E-03
INTA.AUR.059	2,349,078	31.55	4.13	0	Incompleto	2974	3.21E-03	1.93E-03
INTA.AUR.060	5,101,967	55.96	37.04	0	Incompleto	6727	2.99E-03	2.76E-03
INTA.AUR.061	1,952,744	61.85	16.38	0	Incompleto	2460	3.30E-03	2.25E-03
INTA.AUR.062	3,549,440	59.17	16.15	17.65	Incompleto	4757	3.23E-03	2.11E-03
INTA.AUR.063	3,747,220	93.09	3.32	0	Alta calidad	3766	3.16E-03	4.66E-03
INTA.AUR.064	1,614,971	84.33	54	71.11	En borrador	1823	2.30E-03	1.47E-02
INTA.AUR.065	1,262,175	75.78	58.44	6.04	Incompleto	1778	3.01E-03	1.98E-03
INTA.AUR.066	1,053,370	13.48	0.24	0	Incompleto	1490	2.74E-03	2.60E-03
INTA.AUR.067	1,985,931	63.21	27.04	2.38	Incompleto	2777	3.07E-03	1.98E-03

INTA.AUR.068	5,644,885	92.06	3.68	6.25	Alta calidad	4647	3.03E-03	1.76E-02
INTA.AUR.069	4,011,941	70	51.72	0	Incompleto	5710	2.85E-03	2.82E-03
INTA.AUR.070	4,813,963	99.44	2.59	0	Alta calidad	4546	6.19E-03	4.87E-02
INTA.AUR.071	1,535,452	52.74	20.3	0	Incompleto	2323	2.67E-03	2.25E-03
INTA.AUR.072	3,562,310	63.27	21.05	4.05	Incompleto	4360	2.34E-03	2.90E-03
INTA.AUR.073	4,157,485	69.39	39.11	4.11	Incompleto	5829	2.47E-03	1.98E-03
INTA.AUR.074	1,680,487	61.82	0	0	Incompleto	2199	2.38E-03	2.80E-03
INTA.AUR.077	4,348,638	85.07	41.18	0	Incompleto	5808	2.09E-03	2.81E-03
INTA.AUR.078	1,160,122	56.16	2.3	7.14	Incompleto	1561	2.20E-03	3.16E-03
INTA.AUR.079	3,114,545	88.73	62.76	6.45	Incompleto	4233	2.10E-03	3.53E-03
INTA.AUR.080	1,920,005	65.97	47.16	5.41	Incompleto	2555	1.97E-03	3.25E-03
INTA.AUR.081	1,869,582	64.99	37.37	0	Incompleto	2708	1.79E-03	3.32E-03
INTA.AUR.082	3,745,249	72.98	43.5	81.82	En borrador	4185	1.76E-03	1.38E-02
INTA.AUR.083	1,922,422	42.59	29.47	0	Incompleto	2986	1.43E-03	3.72E-03
INTA.AUR.084	7,364,994	74.41	19.09	90.62	En borrador	9058	1.41E-03	7.69E-03
INTA.AUR.085	5,272,200	80.54	25.26	2.78	Incompleto	6258	1.39E-03	5.81E-03
INTA.AUR.086	1,244,915	67.02	3.26	12.5	Incompleto	1521	1.40E-03	5.55E-03
INTA.AUR.087	5,915,492	48.9	42.24	69.44	Incompleto	6620	1.27E-03	9.94E-03

INTA.AUR.088	4,218,825	87.93	37.99	0	Incompleto	5302	1.24E-03	5.85E-03		
INTA.AUR.089	1,955,385	66.19	3.66	0	Incompleto	2693	1.18E-03	4.15E-03		
INTA.AUR.090	3,765,987	46.97	10.22	60.47	Incompleto	4789	1.03E-03	7.78E-03		
INTA.AUR.091	3,422,847	80.58	25.37	1.49	Incompleto	4144	9.89E-04	5.17E-03		
INTA.AUR.092	2,190,143	54.01	5.48	0	Incompleto	3031	9.58E-04	4.57E-03		
INTA.AUR.093	516,154	36.53	2.73	0	Incompleto	703	9.69E-04	4.87E-03		
INTA.AUR.094	3,038,463	60.48	18	1.79	Incompleto	4328	7.95E-04	4.77E-03		
INTA.AUR.095	1,807,093	51.33	14.18	2.22	Incompleto	2355	7.02E-04	4.65E-03		
INTA.AUR.1003	3,174,928	95.17	2.08	50	Alta calidad	3171	2.28E-03	9.50E-03		
INTA.AUR.1005	2,988,185	93.92	5.74	0	En borrador	2903	2.22E-03	1.52E-02		
INTA.CYC.001	693,413	64.59	0.85	100	En borrador	725	5.30E-01	2.76E-01	1.01E-01	5.94E-02
INTA.CYC.002	887,751	12.1	0.3	100	Incompleto	1066	2.83E-01	2.14E-01	1.30E-01	9.98E-02
INTA.CYC.003	2,023,513	47.41	0	0	Incompleto	2505	2.61E-01	1.98E-01	1.16E-01	8.89E-02
INTA.CYC.004	654,041	43.19	0.38	100	Incompleto	830	2.14E-01	1.94E-01	1.11E-01	8.65E-02
INTA.CYC.005	629,433	15.52	1.72	100	Incompleto	909	1.57E-01	1.49E-01	8.68E-02	7.05E-02
INTA.CYC.006	187,641	0	0	0	Incompleto	257	1.97E-01	2.73E-01	2.10E-01	2.11E-01
INTA.CYC.007	2,798,169	96.64	0	0	Alta calidad	2684	1.16E-01	7.19E-02	2.42E-02	1.70E-02
INTA.CYC.008	267,793	0	0	0	Incompleto	364	8.77E-02	8.18E-02	4.70E-02	3.64E-02

INTA.CYC.009	2,405,603	94.35	1.34	0	Alta calidad	2291	7.96E-02	1.14E-01	1.01E-01	9.06E-02
INTA.CYC.010	651,137	2.04	0	0	Incompleto	782	7.73E-02	3.00E-02	1.12E-02	6.99E-03
INTA.CYC.011	802,148	0	0	0	Incompleto	959	6.30E-02	6.10E-02	3.54E-02	2.82E-02
INTA.CYC.012	2,678,597	95.58	0	0	Alta calidad	2595	5.50E-02	2.47E-02	1.39E-02	9.19E-03
INTA.CYC.013	1,439,619	64.47	4.55	20	Incompleto	1805	5.19E-02	7.00E-02	3.67E-02	3.21E-02
INTA.CYC.014	812,129	56.16	3.73	0	Incompleto	1030	5.01E-02	6.00E-02	5.38E-02	6.16E-02
INTA.CYC.015	3,412,212	94.24	7.61	0	En borrador	3097	3.80E-02	3.37E-02	2.46E-02	2.26E-02
INTA.CYC.016	1,664,513	62.64	2	16.67	Incompleto	1721	3.80E-02	2.25E-02	9.65E-03	6.80E-03
INTA.CYC.017	2,950,326	96.87	2.27	50	Alta calidad	3153	2.76E-02	2.97E-02	1.23E-02	9.34E-03
INTA.CYC.018	1,305,518	58.89	1.88	33.33	En borrador	1601	2.70E-02	1.56E-02	7.31E-03	5.30E-03
INTA.CYC.019	1,295,529	64.93	1.59	20	Incompleto	1550	2.75E-02	5.13E-02	4.79E-02	5.43E-02
INTA.CYC.020	3,514,115	99.33	0.34	0	Alta calidad	3514	2.84E-02	6.54E-03	2.34E-03	1.96E-03
INTA.CYC.021	2,013,321	97.9	0.4	0	Alta calidad	2030	2.32E-02	1.09E-02	6.06E-03	3.44E-03
INTA.CYC.022	1,455,176	38.49	2.63	5.88	Incompleto	1525	2.00E-02	1.88E-02	1.23E-02	9.59E-03
INTA.CYC.023	2,328,171	73.34	3.35	55.56	En borrador	2643	2.28E-02	2.54E-02	1.82E-02	1.53E-02
INTA.CYC.024	2,047,803	21.51	1.61	0	Incompleto	2343	2.27E-02	2.61E-02	2.17E-02	2.11E-02
INTA.CYC.025	4,186,274	89.07	8.99	0	En borrador	4452	1.57E-02	2.37E-02	1.79E-02	1.70E-02
INTA.CYC.026	2,109,840	95.09	3.84	0	Alta calidad	2186	1.86E-02	8.93E-03	5.44E-03	3.74E-03

INTA.CYC.027	1,533,755	70.06	1.28	0	En borrador	1615	1.90E-02	3.52E-02	4.18E-02	3.28E-02
INTA.CYC.028	1,555,880	2.07	0	0	Incompleto	1979	1.68E-02	1.23E-02	7.64E-03	5.32E-03
INTA.CYC.029	4,543,380	98.73	4.11	0	Alta calidad	4116	1.46E-02	1.66E-02	1.47E-02	1.81E-02
INTA.CYC.030	2,421,830	99.25	2.06	14.29	Alta calidad	2592	1.28E-02	2.26E-03	6.34E-04	3.32E-04
INTA.CYC.031	2,660,887	56.81	7.02	0	Incompleto	3341	1.37E-02	2.27E-02	2.32E-02	3.17E-02
INTA.CYC.032	3,216,546	100	2.17	0	Alta calidad	3048	1.32E-02	5.28E-03	1.71E-03	1.32E-03
INTA.CYC.033	4,310,811	74.73	54.44	79.59	En borrador	5249	1.32E-02	3.03E-03	6.50E-04	3.29E-04
INTA.CYC.034	3,831,921	100	1.52	16.67	Alta calidad	3231	1.05E-02	1.45E-02	1.88E-02	2.24E-02
INTA.CYC.035	2,958,270	76.15	5.22	6.67	En borrador	3241	1.07E-02	8.42E-03	5.80E-03	5.10E-03
INTA.CYC.036	869,567	23.41	6.18	93.33	Incompleto	1075	1.22E-02	1.69E-02	1.18E-02	5.83E-03
INTA.CYC.037	808,098	31.5	8.62	100	Incompleto	930	1.01E-02	1.36E-02	9.05E-03	4.22E-03
INTA.CYC.038	5,701,692	95.84	48.23	69.86	En borrador	6472	6.31E-03	1.73E-03	4.06E-04	2.13E-04
INTA.CYC.039	3,868,148	99.46	1.08	0	Alta calidad	3122	7.31E-03	7.21E-03	4.49E-03	3.52E-03
INTA.CYC.040	7,095,645	35.34	18.1	83.33	Incompleto	7842	5.75E-03	1.56E-03	4.32E-04	2.55E-04
INTA.CYC.041	2,755,591	93.19	4.19	23.53	Alta calidad	2931	6.94E-03	7.82E-03	3.10E-03	2.16E-03
INTA.CYC.042	4,279,771	23.88	0.56	0	Incompleto	4749	6.42E-03	6.11E-03	4.29E-03	3.85E-03
INTA.CYC.043	2,653,079	92.24	40.44	8.33	Incompleto	2822	5.63E-03	7.85E-03	4.51E-03	3.69E-03
INTA.CYC.044	1,518,806	48.61	28.56	77.78	Incompleto	1552	5.52E-03	1.31E-02	1.28E-02	1.29E-02

INTA.CYC.045	3,941,754	81.36	31.39	18.92	Incompleto	4380	5.48E-03	1.13E-02	1.03E-02	1.16E-02
INTA.CYC.046	955,564	35.82	3.72	42.86	Incompleto	894	6.39E-03	1.39E-02	1.44E-02	1.46E-02
INTA.CYC.047	151,198	1.72	0	0	Incompleto	188	6.11E-03	1.15E-02	1.36E-02	2.10E-02
INTA.CYC.048	871,437	26.49	0	0	Incompleto	838	3.86E-03	1.07E-02	1.18E-02	1.22E-02
INTA.CYC.049	6,190,208	85.89	20.29	2.4	Incompleto	7448	4.16E-03	1.82E-03	6.31E-04	3.71E-04
INTA.CYC.050	2,374,794	88.74	44.95	0	Incompleto	2483	6.06E-03	9.86E-03	7.47E-03	1.17E-02
INTA.CYC.051	2,192,915	90.21	6.23	50	En borrador	2778	4.19E-03	1.29E-03	7.11E-04	5.83E-04
INTA.CYC.052	4,563,645	90.09	7.43	10	En borrador	5507	4.34E-03	1.74E-03	9.53E-04	7.63E-04
INTA.CYC.053	5,391,134	49.06	25.86	45.45	Incompleto	7256	3.88E-03	1.08E-03	3.14E-04	2.15E-04
INTA.CYC.054	338,085	13.39	5.45	100	Incompleto	349	3.98E-03	8.13E-03	7.78E-03	9.32E-03
INTA.CYC.055	1,158,709	17.5	1.15	0	Incompleto	1378	4.19E-03	9.43E-03	9.27E-03	1.00E-02
INTA.CYC.056	2,385,348	37.05	8.62	80	Incompleto	2411	4.03E-03	7.72E-03	7.26E-03	8.34E-03
INTA.CYC.057	2,548,382	98.32	2.68	0	Alta calidad	2482	3.54E-03	6.37E-03	5.11E-03	5.50E-03
INTA.CYC.058	2,719,691	84.48	18.97	18.18	En borrador	2780	3.15E-03	1.78E-03	8.37E-04	5.51E-04
INTA.CYC.059	402,358	28.37	9.25	80	Incompleto	411	3.80E-03	7.69E-03	7.55E-03	8.64E-03
INTA.CYC.060	869,317	18.61	5.17	66.67	Incompleto	859	3.88E-03	7.84E-03	7.73E-03	9.09E-03
INTA.CYC.061	1,976,299	16.61	5.96	88.89	Incompleto	1955	3.54E-03	6.94E-03	6.39E-03	7.22E-03
INTA.CYC.062	2,498,561	42.03	4.31	0	Incompleto	3225	3.99E-03	5.68E-03	5.68E-03	6.46E-03

INTA.CYC.063	1,239,603	95.7	2.15	0	Alta calidad	1248	4.14E-03	3.26E-03	2.33E-03	1.84E-03
INTA.CYC.064	3,176,966	97.65	11.33	3.7	En borrador	3239	3.13E-03	3.54E-03	2.12E-03	1.91E-03
INTA.CYC.065	1,183,635	91.57	3.93	0	Alta calidad	1134	3.29E-03	2.66E-03	1.62E-03	1.35E-03
INTA.CYC.066	3,111,400	100	0.59	0	Alta calidad	2782	2.60E-03	2.28E-03	1.84E-03	1.80E-03
INTA.CYC.067	1,169,140	58.69	0.28	0	En borrador	1315	2.82E-03	5.39E-03	4.76E-03	5.22E-03
INTA.CYC.068	2,428,518	77.02	10.35	15	En borrador	3071	2.96E-03	1.71E-03	1.10E-03	1.33E-03
INTA.CYC.069	4,430,114	100	39.49	2.17	En borrador	4855	2.22E-03	5.62E-03	6.55E-03	7.32E-03
INTA.CYC.070	1,210,671	65.53	3.42	20	En borrador	1278	2.39E-03	3.97E-03	3.47E-03	3.58E-03
INTA.CYC.071	2,588,940	94.44	7.3	31.82	En borrador	2455	2.97E-03	3.69E-03	2.90E-03	3.03E-03
INTA.CYC.072	2,108,884	69.69	10.28	0	Incompleto	2330	1.66E-03	1.55E-03	9.74E-04	8.71E-04
INTA.CYC.073	3,354,135	29.15	14.64	11.63	Incompleto	4391	1.99E-03	1.56E-03	6.92E-04	6.32E-04
INTA.CYC.074	4,768,026	93.76	12.15	5.56	En borrador	4926	2.20E-03	2.76E-03	2.13E-03	2.18E-03
INTA.CYC.075	1,633,249	75.85	8.89	0	En borrador	1739	1.66E-03	1.53E-03	1.03E-03	1.05E-03
INTA.CYC.076	753,375	49.92	6.79	0	Incompleto	810	1.55E-03	2.21E-03	1.41E-03	1.38E-03
INTA.CYC.077	2,135,824	86.75	0.79	0	En borrador	2351	1.81E-03	1.50E-03	8.56E-04	8.32E-04
INTA.CYC.078	3,429,049	88.6	7.4	45.45	En borrador	3715	1.32E-03	1.65E-03	1.20E-03	1.31E-03
INTA.CYC.079	6,365,639	60.24	30.35	0	Incompleto	8249	1.56E-03	1.39E-03	9.15E-04	8.91E-04
INTA.CYC.080	2,279,884	61.24	26.01	0	Incompleto	3039	1.67E-03	1.49E-03	1.16E-03	1.15E-03

INTA.CYC.081	2,261,101	86.2	6.44	6.25	En borrador	2740	1.59E-03	1.47E-03	8.37E-04	7.47E-04
INTA.CYC.082	2,469,134	74.69	21.61	50	Incompleto	3001	1.71E-03	5.30E-03	6.72E-03	7.63E-03
INTA.CYC.083	3,994,979	28.07	10.4	0	Incompleto	4791	2.47E-03	3.69E-03	3.39E-03	4.15E-03
INTA.CYC.084	3,315,188	96.52	8.9	0	En borrador	3785	1.44E-03	1.93E-03	1.27E-03	1.25E-03
INTA.CYC.086	6,038,627	95.52	52.93	15.49	En borrador	7215	1.49E-03	2.13E-03	1.18E-03	1.06E-03
INTA.CYC.087	2,653,424	92.24	56.77	0	Incompleto	3276	1.54E-03	1.44E-03	9.47E-04	7.76E-04
INTA.CYC.088	2,332,532	71.59	19.52	1.92	En borrador	2992	1.35E-03	1.26E-03	8.49E-04	7.49E-04
INTA.CYC.089	3,059,777	83.46	20.95	4.17	Incompleto	3628	1.04E-03	1.41E-03	9.02E-04	9.19E-04
INTA.CYC.090	3,934,006	83.98	46.14	1.04	Incompleto	5225	1.02E-03	1.24E-03	9.54E-04	9.56E-04
INTA.CYC.091	3,005,459	96.1	9.08	76.19	En borrador	2956	1.15E-03	3.47E-03	2.32E-03	2.06E-03
INTA.CYC.092	5,482,358	21.39	6.84	7.14	Incompleto	7678	1.05E-03	1.06E-03	1.12E-03	1.23E-03
INTA.CYC.093	729,346	55.6	15.86	60	En borrador	830	1.44E-03	3.67E-03	3.22E-03	3.29E-03
INTA.CYC.094	3,423,791	89.02	4.92	26.67	En borrador	3042	1.52E-03	4.00E-03	4.81E-03	4.99E-03
INTA.CYC.095	3,364,953	76.66	11.93	6.45	En borrador	4013	7.46E-04	1.23E-03	1.20E-03	1.04E-03
INTA.CYC.096	2,918,789	2.9	0.86	100	Incompleto	3679	1.24E-03	2.21E-03	1.82E-03	2.08E-03
INTA.CYC.097	1,392,233	68.17	4.27	0	Incompleto	1510	1.16E-03	1.71E-03	1.51E-03	1.27E-03
INTA.CYC.098	4,930,265	44.43	12.89	0	Incompleto	6547	1.02E-03	1.54E-03	1.13E-03	1.13E-03
INTA.CYC.099	6,636,099	73.68	61.58	2.28	Incompleto	8690	9.37E-04	1.23E-03	1.06E-03	1.10E-03

INTA.CYC.100	2,282,586	90.35	34.85	5.26	Incompleto	2719	9.77E-04	1.60E-03	1.88E-03	1.54E-03
INTA.CYC.1002	1,422,893	95.56	2.82	25	Alta calidad	1568	2.25E-03	1.09E-02	4.03E-03	7.50E-03
INTA.CYC.101	913,770	18.13	12.39	3.85	Incompleto	1355	9.60E-04	1.12E-03	7.43E-04	8.03E-04
INTA.CYC.102	1,332,250	25.79	4.06	0	Incompleto	1793	6.81E-04	1.01E-03	9.60E-04	1.14E-03
INTA.CYC.103	1,108,589	83.27	54.24	6.25	En borrador	1408	9.85E-04	1.42E-03	1.20E-03	1.50E-03
INTA.CYC.104	3,869,238	93.53	1.88	25	Alta calidad	3335	9.22E-04	4.63E-03	8.84E-03	8.90E-03
INTA.CYC.105	4,009,592	61.15	22.4	4.86	Incompleto	5411	8.38E-04	1.43E-03	1.13E-03	1.21E-03
INTA.CYC.106	1,060,895	28.4	0	0	Incompleto	1622	8.18E-04	1.02E-03	1.13E-03	1.08E-03
INTA.CYC.107	3,090,672	64.69	8.77	40	Incompleto	3880	9.08E-04	2.10E-03	2.58E-03	3.43E-03
INTA.CYC.108	2,642,188	89.25	14.62	5	En borrador	2846	7.80E-04	1.56E-03	1.52E-03	1.17E-03
INTA.CYC.109	1,178,912	92.98	22.33	50	En borrador	1260	7.99E-04	2.31E-03	2.27E-03	2.51E-03
INTA.CYC.110	2,600,572	54.28	20.28	0	Incompleto	3304	7.66E-04	1.48E-03	1.63E-03	1.66E-03
INTA.CYC.111	5,672,747	74.16	34.8	1.43	Incompleto	7180	6.37E-04	1.19E-03	1.05E-03	1.08E-03
INTA.CYC.112	4,676,396	59.55	38.43	2.78	Incompleto	5599	7.72E-04	1.24E-03	9.88E-04	1.04E-03
INTA.CYC.113	4,275,639	84.89	39.51	7	Incompleto	4885	7.08E-04	1.29E-03	1.60E-03	1.76E-03
INTA.CYC.114	2,099,011	67	19.83	0	Incompleto	2780	7.08E-04	1.13E-03	1.10E-03	1.14E-03
INTA.CYC.115	3,888,427	85.89	45.02	13.64	Incompleto	5288	6.46E-04	1.45E-03	1.17E-03	1.24E-03
INTA.CYC.116	3,967,674	5.95	0.16	0	Incompleto	5369	6.85E-04	1.21E-03	1.13E-03	1.28E-03

INTA.CYC.117	1,323,556	68.18	42.33	9.68	Incompleto	1768	6.55E-04	1.45E-03	1.65E-03	1.60E-03
INTA.CYC.118	5,496,463	52.41	27.52	0	Incompleto	8006	5.22E-04	8.31E-04	8.63E-04	1.12E-03
INTA.CYC.119	751,762	41.4	16.14	0	Incompleto	1021	6.09E-04	1.22E-03	1.50E-03	1.51E-03
INTA.CYC.120	497,812	38.26	29.17	33.33	Incompleto	599	6.32E-04	1.34E-03	1.48E-03	1.49E-03
INTA.CYC.121	3,132,108	41.7	7.05	0	Incompleto	4724	6.33E-04	8.27E-04	1.31E-03	1.64E-03
INTA.CYC.122	2,624,195	4.66	0	0	Incompleto	3609	6.26E-04	1.18E-03	1.24E-03	1.67E-03
INTA.CYC.123	6,735,399	67.05	39.12	10.94	Incompleto	8976	5.44E-04	1.16E-03	1.27E-03	1.55E-03
INTA.CYC.124	573,636	52.8	31.5	4	Incompleto	954	6.44E-04	1.00E-03	1.03E-03	1.10E-03
INTA.CYC.125	856,682	11.43	0.41	0	Incompleto	1156	5.72E-04	1.05E-03	9.15E-04	1.03E-03
INTA.CYC.126	3,731,960	99.53	75.24	36.23	En borrador	4955	5.55E-04	7.78E-04	2.16E-03	3.01E-03
INTA.CYC.127	6,143,445	87.29	54.78	1.32	Incompleto	7414	3.91E-04	1.18E-03	1.22E-03	1.50E-03
INTA.CYC.128	1,012,128	58.1	37.72	5.36	Incompleto	1429	5.83E-04	1.15E-03	1.31E-03	1.63E-03
INTA.CYC.129	452,298	61.74	19.4	27.03	Incompleto	601	5.49E-04	1.37E-03	1.61E-03	1.59E-03
INTA.CYC.130	3,058,959	53.97	26.03	2.86	Incompleto	3932	5.59E-04	9.73E-04	9.47E-04	9.66E-04
INTA.CYC.131	840,370	45.73	20.98	2.27	Incompleto	1199	6.28E-04	1.04E-03	1.10E-03	1.09E-03
INTA.CYC.132	3,221,322	11.15	0.33	0	Incompleto	3267	5.20E-04	1.19E-03	1.19E-03	1.31E-03
INTA.CYC.133	271,074	6.9	0	0	Incompleto	376	5.13E-04	1.16E-03	1.04E-03	9.49E-04
INTA.CYC.134	821,792	80.9	6.18	57.14	En borrador	970	4.90E-04	2.43E-03	2.44E-03	3.17E-03

INTA.CYC.135	875,530	22.41	6.9	75	Incompleto	946	4.85E-04	2.34E-03	3.34E-03	3.56E-03
INTA.CYC.136	525,176	56.03	1.72	0	Incompleto	604	4.84E-04	1.71E-03	2.12E-03	2.43E-03
INTA.CYC.137	807,677	12.15	0.16	0	Incompleto	764	4.77E-04	2.27E-03	3.22E-03	3.38E-03
INTA.CYC.138	4,362,828	69.36	50.45	3.39	Incompleto	6187	4.54E-04	9.45E-04	9.71E-04	9.78E-04
INTA.CYC.139	592,715	38.84	19.41	0	Incompleto	845	4.16E-04	1.06E-03	1.13E-03	1.46E-03
INTA.CYC.140	3,086,775	73.9	37.61	15.87	Incompleto	4478	4.10E-04	7.94E-04	9.90E-04	1.91E-03
INTA.CYC.141	1,194,560	44.66	8.24	75	Incompleto	1123	5.07E-04	2.55E-03	3.76E-03	4.17E-03
INTA.CYC.142	1,293,947	38.43	1.2	100	Incompleto	1326	4.22E-04	1.36E-03	1.87E-03	1.81E-03
INTA.CYC.143	4,468,022	97.33	5.33	0	En borrador	4248	8.79E-05	2.86E-03	3.70E-03	3.82E-03
INTA.CYC.144	2,721,628	70.93	8.26	0	En borrador	2924	3.74E-04	1.67E-03	1.92E-03	2.20E-03
INTA.CYC.145	4,731,251	74.35	48.58	0	Incompleto	6884	3.15E-04	8.94E-04	9.19E-04	1.04E-03
INTA.CYC.146	424,499	22.86	3.45	0	Incompleto	636	3.18E-04	1.51E-03	2.04E-03	2.65E-03
INTA.CYC.147	2,416,737	58.18	25.97	4.44	Incompleto	2880	3.21E-04	1.48E-03	1.46E-03	1.43E-03
INTA.CYC.148	1,932,649	51.46	12.25	0	Incompleto	2588	3.10E-04	6.81E-04	8.83E-04	1.41E-03
INTA.CYC.149	952,896	40.82	18.67	50	Incompleto	1026	3.15E-04	1.48E-03	1.52E-03	1.71E-03
INTA.CYC.150	1,986,971	84.24	7.33	4.35	En borrador	2123	2.84E-04	1.90E-03	1.46E-03	1.48E-03
INTA.CYC.151	452,964	36.68	15.81	78.26	Incompleto	613	2.93E-04	1.22E-03	1.35E-03	1.42E-03
INTA.CYC.152	2,774,430	67.01	43.97	10	Incompleto	3582	2.93E-04	8.96E-04	1.35E-03	1.21E-03

INTA.CYC.153	4,473,501	96.24	3.76	66.67	Alta calidad	3684	3.34E-04	2.50E-03	5.26E-03	5.28E-03
INTA.CYC.154	4,328,753	72.34	33.83	5.08	Incompleto	5145	2.82E-04	1.10E-03	1.28E-03	1.43E-03
INTA.CYC.155	880,273	58.45	2.43	66.67	En borrador	1063	2.63E-04	7.88E-03	1.37E-02	1.42E-02
INTA.CYC.156	2,599,207	86.21	36.11	3.12	Incompleto	3326	2.37E-04	1.74E-03	1.58E-03	1.13E-03
INTA.CYC.157	919,082	76.3	19.95	35	En borrador	1021	2.75E-04	1.89E-03	2.47E-03	3.32E-03
INTA.CYC.158	3,047,626	53.45	25.86	0	Incompleto	3811	1.88E-04	1.04E-03	1.52E-03	1.28E-03
INTA.CYC.159	2,708,520	85.1	17.31	0	En borrador	3418	1.94E-04	8.16E-04	1.24E-03	1.81E-03
INTA.CYC.160	1,635,345	47.2	4.07	9.52	Incompleto	2382	1.47E-04	3.03E-04	9.33E-04	1.90E-03
INTA.CYC.161	2,204,351	73.07	3.91	9.09	En borrador	2353	9.68E-05	1.69E-03	1.45E-03	1.34E-03
INTA.CYC.162	3,683,627	20.06	10.84	6.25	Incompleto	4747	6.42E-05	4.58E-04	1.84E-03	4.06E-03
INTA.CYC.163	975,988	69.8	4.14	0	En borrador	1010	7.90E-05	2.08E-03	4.50E-03	9.53E-03
INTA.CYC.164	2,396,994	73.75	6.21	6.67	En borrador	3104	3.05E-05	4.50E-04	1.45E-03	2.25E-03
INTA.CYC.165	4,771,663	98.37	6.18	13.33	En borrador	4175	4.69E-05	2.52E-03	4.14E-03	5.17E-03
INTA.CYC.166	5,244,048	98.92	7.08	20.83	En borrador	5271	5.16E-05	4.54E-04	2.27E-03	6.49E-03
INTA.CYC.167	3,469,606	56.41	3.36	73.33	Incompleto	4277	1.19E-05	3.76E-03	4.24E-02	1.27E-01
INTA.CYC.168	3,893,883	32.12	2.25	32	Incompleto	4718	6.58E-06	8.36E-04	2.56E-02	6.56E-02
INTA.CYC.169	4,831,734	92.69	4.42	26.92	Alta calidad	4807	1.15E-06	5.57E-04	4.08E-03	1.11E-02

Tabla S4: Clasificación taxonómica obtenida con GTDBtk de los MAGs de alta calidad y en borrador.

BinID	Linaje de GTDB	Metodo de asignacion
INTA.AUR.002	d__Bacteria|p__Desulfobacterota|c__Desulfovibrionia|o__Desulfovibrionales	ANI
INTA.AUR.003	d__Bacteria|p__Verrucomicrobiota|c__Kiritimatiellae	Novedad determinada mediante RED
INTA.AUR.004	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.AUR.005	d__Bacteria|p__Actinobacteriota|c__Actinomycetia|o__Actinomycetales	Topologia y ANI
INTA.AUR.006	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Topologia y ANI
INTA.AUR.007	d__Bacteria|p__Verrucomicrobiota|c__Kiritimatiellae|o__LD1-PB3	Topologia y ANI
INTA.AUR.008	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Novedad determinada mediante RED
INTA.AUR.009	d__Bacteria|p__Actinobacteriota|c__Actinomycetia|o__Actinomycetales	Topologia y ANI
INTA.AUR.010	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.AUR.011	d__Bacteria|p__Campylobacterota|c__Campylobacteria|o__Campylobacterales	Topologia y ANI
INTA.AUR.013	d__Bacteria|p__Patescibacteria|c__ABY1|o__BM507	Novedad determinada mediante RED
INTA.AUR.014	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.AUR.015	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Topologia y ANI
INTA.AUR.017	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	Topology
INTA.AUR.021	d__Bacteria|p__Firmicutes_C|c__Negativicutes|o__Acidaminococcales	Novedad determinada mediante RED

INTA.AUR.023	d__Bacteria|p__Patescibacteria|c__ABY1|o__BM507	Topology
INTA.AUR.024	d__Bacteria|p__Patescibacteria|c__ABY1|o__BM507	Novedad determinada mediante RED
INTA.AUR.026	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Lachnospirales	Novedad determinada mediante RED
INTA.AUR.027	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	Topologia y ANI
INTA.AUR.028	d__Bacteria|p__Synergistota|c__Synergistia|o__Synergistales	Topologia y ANI
INTA.AUR.030	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Saccharofermentanales	Topologia y ANI
INTA.AUR.031	d__Bacteria|p__Firmicutes_B|c__Peptococcia|o__Peptococcales	Topologia y ANI
INTA.AUR.032	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topology
INTA.AUR.034	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.AUR.035	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Clostridiales	Novedad determinada mediante RED
INTA.AUR.036	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Saccharofermentanales	Topologia y ANI
INTA.AUR.038	d__Bacteria|p__Patescibacteria|c__ABY1|o__UBA11705	Novedad determinada mediante RED
INTA.AUR.039	d__Bacteria|p__Verrucomicrobiota|c__Kiritimatiellae|o__LD1-PB3	ANI
INTA.AUR.040	d__Bacteria|p__Firmicutes|c__Bacilli|o__Erysipelotrichales	Novedad determinada mediante RED
INTA.AUR.042	d__Bacteria|p__Verrucomicrobiota|c__Verrucomicrobiae|o__Pedosphaerales	Topologia y ANI
INTA.AUR.044	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Saccharofermentanales	Novedad determinada mediante RED
INTA.AUR.045	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Novedad determinada mediante RED
INTA.AUR.047	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	

INTA.AUR.051	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	ANI
INTA.AUR.055	d__Bacteria|p__Synergistota|c__Synergistia|o__Synergistales	ANI
INTA.AUR.063	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Lachnospirales	ANI
INTA.AUR.064	d__Bacteria|p__Patescibacteria|c__Paceibacteria_A|o__Moranbacterales	Topologia y ANI
INTA.AUR.068	d__Bacteria|p__Verrucomicrobiota|c__Kiritimatiellae|o__SLAD01	Novedad determinada mediante RED
INTA.AUR.070	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Chromatiales	Novedad determinada mediante RED
INTA.AUR.082	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Chromatiales	Topologia y ANI
INTA.AUR.084	d__Bacteria|p__Chloroflexota|c__Chloroflexia|o__Chloroflexales	Topologia y ANI
INTA.AUR.1003	d__Bacteria|p__Desulfobacterota|c__Desulfovibrionia|o__Desulfovibrionales	Topology
INTA.AUR.1005	d__Bacteria|p__Proteobacteria|c__Alphaproteobacteria|o__Micavibrionales	Topology
INTA.CYC.001	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Novedad determinada mediante RED
INTA.CYC.007	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Lachnospirales	ANI
INTA.CYC.009	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.012	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI
INTA.CYC.015	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	ANI
INTA.CYC.017	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	Topologia y ANI
INTA.CYC.018	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Topologia y ANI
INTA.CYC.020	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Lachnospirales	Novedad determinada mediante RED

INTA.CYC.021	d__Bacteria|p__Firmicutes_C|c__Negativicutes|o__Veillonellales	ANI
INTA.CYC.023	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.025	d__Bacteria|p__Proteobacteria|c__Alphaproteobacteria|o__Micavibrionales_A	Topologia y ANI
INTA.CYC.026	d__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales	Topologia y ANI
INTA.CYC.027	d__Bacteria|p__Patescibacteria|c__Gracilibacteria|o__BD1-5	Topologia y ANI
INTA.CYC.029	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Lachnospirales	ANI
INTA.CYC.030	d__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales	Topologia y ANI
INTA.CYC.032	d__Bacteria|p__Firmicutes_C|c__Negativicutes|o__Anaeromusales	Topologia y ANI
INTA.CYC.033	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales	Topologia y ANI
INTA.CYC.034	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.CYC.035	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Peptostreptococcales	Topologia y ANI
INTA.CYC.038	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales	Topologia y ANI
INTA.CYC.039	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.041	d__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales	Topologia y ANI
INTA.CYC.051	d__Bacteria|p__Campylobacterota|c__Campylobacteria|o__Campylobacterales	Topologia y ANI
INTA.CYC.052	d__Bacteria|p__Firmicutes|c__Bacilli|o__Erysipelotrichales	Novedad determinada mediante RED
INTA.CYC.057	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI
INTA.CYC.058	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI

INTA.CYC.063	d__Bacteria|p__Proteobacteria|c__Alphaproteobacteria|o__RF32	Novedad determinada mediante RED
INTA.CYC.064	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI
INTA.CYC.065	d__Bacteria|p__Firmicutes_A|c__Clostridia_A|o__Christensenellales	Topologia y ANI
INTA.CYC.066	d__Bacteria|p__Desulfobacterota|c__Desulfovibrionia|o__Desulfovibrionales	Topologia y ANI
INTA.CYC.067	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Topologia y ANI
INTA.CYC.068	d__Bacteria|p__Firmicutes_A|c__Clostridia_A|o__Christensenellales	Topologia y ANI
INTA.CYC.069	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Peptostreptococcales	Topologia y ANI
INTA.CYC.070	d__Bacteria|p__Patescibacteria|c__Saccharimonadia|o__Saccharimonadales	Topologia y ANI
INTA.CYC.071	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	ANI
INTA.CYC.074	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.075	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI
INTA.CYC.077	d__Bacteria|p__Firmicutes|c__Bacilli|o__Erysipelotrichales	ANI
INTA.CYC.078	d__Bacteria|p__Synergistota|c__Synergistia|o__Synergistales	Topologia y ANI
INTA.CYC.081	d__Bacteria|p__Firmicutes_A|c__Clostridia_A|o__Christensenellales	Topologia y ANI
INTA.CYC.084	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	ANI
INTA.CYC.086	d__Bacteria|p__Spirochaetota|c__Spirochaetia|o__Sphaerochaetales	Topologia y ANI
INTA.CYC.088	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Oscillospirales	Topologia y ANI
INTA.CYC.091	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Saccharofermentanales	Topologia y ANI

INTA.CYC.093	d__Bacteria|p__Patescibacteria|c__ABY1|o__BM507	Topologia y ANI
INTA.CYC.094	d__Bacteria|p__Verrucomicrobiota|c__Kiritimatiellae|o__LD1-PB3	Topologia y ANI
INTA.CYC.095	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.1002	d__Archaea|p__Thermoplasmatota|c__Thermoplasmata|o__Methanomassiliicoccales	Topologia y ANI
INTA.CYC.103	d__Bacteria|p__Firmicutes|c__Bacilli|o__RFN20	Topologia y ANI
INTA.CYC.104	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.CYC.108	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.109	d__Bacteria|p__Firmicutes|c__Bacilli|o__RFN20	Topologia y ANI
INTA.CYC.126	d__Bacteria|p__Campylobacterota|c__Campylobacteria|o__Campylobacterales	Topologia y ANI
INTA.CYC.134	d__Bacteria|p__Firmicutes|c__Bacilli|o__RFN20	Topologia y ANI
INTA.CYC.143	d__Bacteria|p__Riflebacteria|c__Ozemobacteria|o__Ozemobacterales	Novedad determinada mediante RED
INTA.CYC.144	d__Bacteria|p__Riflebacteria|c__Ozemobacteria|o__Ozemobacterales	Topologia y ANI
INTA.CYC.150	d__Bacteria|p__Firmicutes_A|c__Clostridia_A|o__Christensenellales	ANI
INTA.CYC.153	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Novedad determinada mediante RED
INTA.CYC.155	d__Bacteria|p__Patescibacteria|c__Paceibacteria_A|o__Moranbacterales	ANI
INTA.CYC.157	d__Bacteria|p__Patescibacteria|c__Paceibacteria|o__UBA9983	Novedad determinada mediante RED
INTA.CYC.159	d__Bacteria|p__Firmicutes_A|c__Clostridia|o__Saccharofermentanales	Novedad determinada mediante RED
INTA.CYC.161	d__Bacteria|p__Firmicutes_A|c__Clostridia_A|o__Christensenellales	Novedad determinada mediante RED

INTA.CYC.163	d__Bacteria|p__Patescibacteria|c__ABY1|o__BM507	Topologia y ANI
INTA.CYC.164	d__Bacteria|p__Proteobacteria|c__Alphaproteobacteria|o__Rhodobacterales	Topologia y ANI
INTA.CYC.165	d__Bacteria|p__Bacteroidota|c__Bacteroidia|o__Bacteroidales	Topologia y ANI
INTA.CYC.166	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Chromatiales	Novedad determinada mediante RED
INTA.CYC.169	d__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Chromatiales	Topologia y ANI

Tabla S5: Oligonucleótidos diseñados para la amplificación por PCR de los genes candidatos

Gen	ID	Dirección del cebador	Secuencia	Temp. Fusion (ºC)	Sitio Restricción
AUR.contig-100_513_12	Bgal1	Directo	ATGCATTCATATCAAATACC	43	NdeI
		Inverso	TTAGCCATTGTTTTTCTCC	44	XbaI
AUR.contig-100_17147_1	Bgal2	Directo	ATCATGACACTCTATTTTCCA	46	
		Inverso	TCAACGCATCATCTATTCAAG	48	
AUR.contig-100_3427_6	Bgal3	Directo	TGTCAAGGGATGAGCACCGTTC	58	
		Inverso	CATCGCGTCAGGGCAGTTC	55	
AUR.contig-100_2450_8	Bgal4	Directo	ATAATGAGAAAACTACCAGAAG	47	NdeI
		Inverso	TTATTTAAGCTCCTTTGTTTC	45	SacI
AUR.contig-100_1620_7	Bgal5	Directo	GGAAGCTATATGAGTCAAGAAGTG	54	BamHI
		Inverso	CATTTTCAGTTCAGTGCTCC	50	SacI
AUR.contig-100_1130_5	Bgal6	Directo	ATATAATGATTAAATGCGAGTTC	46	
		Inverso	GGAAACAGAATTACTTGCCT	48	
CYC.contig-100_2881_9	Bgal7	Directo	ATGACTGGTAAGTTTCCGA	47	NdeI
		Inverso	TCAGAGTTCATTTCCATTCGT	48	SacI
CYC.contig-100_4237_11	Bgal8	Directo	ATGAAACATTTTTTATTGTCATC	45	
		Inverso	AGTACGCAGGTGAA	48	
CYC.contig-100_9789_2	Bgal9	Directo	AGAACAGGAGGAGAGGCT	50	NdeI
		Inverso	TTATAGCTCTTCTCCGTTGC	50	SalI

CYC.contig-100_1083_6	Bgal10	Directo	GTGCTGAAGCCTGTAGAGAATC	55	NdeI
		Inverso	GTTCATTTTTAACCTCCGAAGA	49	SalI
CYC.contig-100_2840_6	Bgal11	Directo	ATGTCTGATTTATTTTATGGCGTG	51	NdeI
		Inverso	CCGCTTAACTTTCGATGATAGT	51	XbaI
AUR.contig-100_220_22	Bgal12	Directo	ATGTTACGCGATGCAAAGAATC	51	
		Inverso	TTTGCTGAACAGGTTATTTCATAAA	49	
AUR.contig-100_5793_4	Bgal13	Directo	ATGATCGTCCCGAACCACCACG	59	
		Inverso	GATGGGGGTTACTCACTGAACACC	59	
CYC.contig-100_1928_8	Bgal14	Directo	AAGGAGGAAGAAAACCAGATGAAAACA	55	
		Inverso	TGTATTTTTCAATTATGTTTATCAGATTCC	52	
AUR.contig-100_1432_6	Bgal15	Directo	AACAAAATGGACAGTAGAGACTGGCAG	58	
		Inverso	TTATGACCTTTCAGTCGTTTTATAGTTATATGG	57	
CYC.contig-100_1151_7	Bgal16	Directo	TGGATTGTCTAAAATAAGAAATAAAATATG	51	BamHI
		Inverso	TTATTTTGTTTTTACTGTAAGTTTCGTG	51	HindII
CYC.contig-100_156_7	Bgal17	Directo	ACAGAACAAATGATGAAGAAACTAATCTTG	55	
		Inverso	AAGGCTTATTTACTGCGCTGTAATTTC	55	
CYC.contig-100_20362_2	Bgal18	Directo	TCCGGAATGAAAAAGATTTGTCTC	52	
		Inverso	GTGGTTTCGTAATTTCTATTTCATGTG	54	
AUR.contig-100_17930_3	Pr1	Directo	ATGGGGAAAAAATCGTTG	43	NdeI
		Inverso	TTACCTACGAAGGAACAG	46	SacI
AUR.contig-100_23747_2	Pr2	Directo	ATGGGCCTAATGGTGGGC	53	

		Inverso	TCACCATACCAACGGGAAGAT	52	
AUR.contig-100_4_168	Pr3	Directo	ATGGCGGCGGGTGCCCTC	60	
		Inverso	GCCGCGCGATCAATCCATGAT	56	
AUR.contig-100_591_6	Pr4	Directo	ACCATGAAGGACAAGTACGTCTCA	56	NdeI
		Inverso	TCACTCCTGCGGGGCCTG	57	EcoRI
AUR.contig-100_7246_4	Pr5	Directo	ATGGAAGAAAAAGAGCAACAA	47	NdeI
		Inverso	TTATTCGTAACTGCCGAG	46	EcoRI
CYC.contig-100_1230_19	Pr6	Directo	attataattATGAAAAGAATACTT	42	NdeI
		Inverso	aagCTATTCTACAATTAGTTTATT	45	SacI
CYC.contig-100_2339_2	Pr7	Directo	ccgATGAAAGCCCGCTCCTTC	59	
		Inverso	TCATGGCTTGCGGAGTCCGAA	56	
CYC.contig-100_44516_2	Pr8	Directo	acgATGCAAAAAACGATTTTA	45	NdeI
		Inverso	accTTAAGGCATAACCAAAAT	46	EcoRI
CYC.contig-100_7627_4	Pr9	Directo	ATGAAAGATACTACGCAATTC	46	BamHI
		Inverso	TTAACGAACCATCCATTCCAT	48	EcoRI
CYC.contig-100_8118_5	Pr10	Directo	ATGATTGGAGATGATCTT	41	NdeI
		Inverso	TTAAACTTCTATCTGAATTGT	43	EcoRI

Tabla S6: Oligonucleótidos diseñados para el clonado en *S. cerevisiae.* En negrita se indica la región de homología con el vector YEP, mientras que la secuencia restante corresponde al gen a clonar.

ID	Cebador directo	Cebador inverso
Pr01	**AAAGAGACTACAAGGATGACGATGACAAGG**ATGGGGAAAAAATCGTTGTT	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**TTACCTACGAAGGAACAGAA
Pr04	**AAAGAGACTACAAGGATGACGATGACAAGG**ACCATGAAGGACAAGTACGTCTCA	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**TCACTCCTGCGGGGCCTG
Pr05	**AAAGAGACTACAAGGATGACGATGACAAGG**ATGGAAGAAAAAGAGCAACAA	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**TTATTCGTAACTGCCGAG
Pr06	**AAAGAGACTACAAGGATGACGATGACAAGG**ATGAAAAGAATACTTTTATC	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**CTATTCTACAATTAGTTTAT
Pr08	**AAAGAGACTACAAGGATGACGATGACAAGG**acgATGCAAAAAACGATTTTA	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**accTTAAGGCATAACCAAAAT
Pr09	**AAAGAGACTACAAGGATGACGATGACAAGG**ATGAAAGATACTACGCAATTC	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**TTAACGAACCATCCATTCCAT
Pr10	**AAAGAGACTACAAGGATGACGATGACAAGG**ATGATTGGAGATGATCTTAT	**CTCGACGGATCAGCGGCCGCTTACCGCGGG**TTAAACTTCTATCTGAATTG

Tabla S7: Resultados de secuenciación por capilares.

Gen	Familia	Largo (b)	Clon	¿Lecturas unidas?	Largo secuenciado (b)	Diferencias con gen predicho	Mutaciones no sinónimas	Ns
βgal1	GH1	1359	1.1	SI	1402	1	0	0
			1.2	SI	1380	1	0	1
			1.3	NO	1229	3	2	6
βgal4	GH1	1404	4.1	SI	1445	3	1	0
			4.2	SI	1442	0	0	0
βgal5	GH42	2031	5.1	NO	1293	4	1	4
			5.3	NO	1800	4	1	5
			5.5	NO	1849	4	2	6
βgal6	GH2	1857	6.1	NO	1811	2	4	9
			6.9	NO	1736	4	4	9
			6.10	NO	1756	0	0	6
βgal7	GH1	1407	7.1	SI	1451	3	2	1
			7.3	SI	1452	1	0	0

Tabla S8. Arquitecturas de dominio en las GH2 encontradas

		GH2N		GH2		GH2C		DUF4981		DUF4982		Bgal_Small_N	
Id	Largo	Inicio	Fin	Inicio	Fin	Inicio	Fin	Inicio	Fin	Inicio	Fin	Inicio	Fin
AUR.contig-100_1051_3	792	35	177	191	299					612	679		
AUR.contig-100_1072_8	735	5	150	161	249	254	549			565	617		
AUR.contig-100_1130_5	619	70	224	226	321								
AUR.contig-100_1222_7	994	42	209	211	316	318	602	609	697				
AUR.contig-100_12476_3	825			189	293								
AUR.contig-100_12671_2	829	25	173	175	286	290	546			646	707		
AUR.contig-100_1432_6	1025	43	203	205	314	316	607	616	705				
AUR.contig-100_1479_7	963			229	351			592	681				
AUR.contig-100_1620_8	1035	66	239	241	339	341	637	644	729				
AUR.contig-100_178_18	1273	68	239	241	354	356	640	650	739				
AUR.contig-100_178_9	847	49	204	208	314								
AUR.contig-100_1893_9	1005	31	203	205	324	326	615						
AUR.contig-100_19072_2	727			162	270								
AUR.contig-100_220_22	1037	42	210	212	324	326	619	629	718				
AUR.contig-100_2309_7	1053	56	223	224	324	326	639	649	743				

AUR.contig-100_2475_8	805	26	220	215	332						
AUR.contig-100_262_1	1006	56	219	221	340			597	686		
AUR.contig-100_3001_8	824	49	178	193	286	288	563				
AUR.contig-100_3475_2	947	28	186	196	293	297	549	563	643		
AUR.contig-100_364_19	629	6	146	162	261	263	500				
AUR.contig-100_404_15	1039	47	215	206	311	313	628				
AUR.contig-100_4055_2	834	28	180	181	293	299	555			655	717
AUR.contig-100_4836_5	826	7	157	188	296						
AUR.contig-100_4938_2	1047	64	237	237	341	343	628	638	728		
AUR.contig-100_5694_3	789	5	145	146	240					617	676
AUR.contig-100_5793_4	1022	44	210	212	309	311	616	626	714		
AUR.contig-100_6402_2	874	22	212	213	318						
AUR.contig-100_6414_2	1048	58	226	228	318	320	632	642	728		
AUR.contig-100_7564_2	1122	55	224	226	315	317	629	641	740		
AUR.contig-100_7607_3	1072	68	248	250	359	361	649	659	747		
AUR.contig-100_7904_2	1023	44	212	214	311	313	618	628	716		
AUR.contig-100_808_10	1051	57	223	226	322	324	637	647	741		
AUR.contig-100_8169_3	976			432	544						

AUR.contig-100_931_11	718			161	271						
AUR.contig-100_9742_2	780	7	152	152	245	248	526				
CYC.contig-100_10_118	1051	47	220	222	332	334	641	651	740		
CYC.contig-100_10089_3	724			603	696						
CYC.contig-100_10618_2	972	52	214			314	549	584	662		
CYC.contig-100_1074_5	957	26	191	196	301	303	598				
CYC.contig-100_1083_11	601	12	182	184	276	278	594				
CYC.contig-100_1083_6	605	12	180	182	273	275	595				
CYC.contig-100_1114_2	957	24	217	221	325						
CYC.contig-100_1151_14	835	35	187	189	298	303	534			654	714
CYC.contig-100_1151_15	841	27	179	181	291					664	724
CYC.contig-100_1151_7	823	28	180	181	292	294	583			627	705
CYC.contig-100_12240_1	971	87	243	287	401	474	718				
CYC.contig-100_12577_2	761	56	191	194	301	317	533				
CYC.contig-100_12884_2	776	21	152	157	242	244	521				
CYC.contig-100_1327_21	799	9	157	159	264						
CYC.contig-100_1343_11	849	101	230	232	344						
CYC.contig-100_1343_4	879	63	215	217	328					701	761

CYC.contig-100_1361_21	863	23	172	222	332						
CYC.contig-100_13669_4	861	29	168	220	317						
CYC.contig-100_13965_3	745	31	155	158	263	265	544				
CYC.contig-100_14113_3	806	30	162	163	263	265	542				
CYC.contig-100_15330_2	798	36	183	196	305					618	685
CYC.contig-100_15512_1	768			161	266						
CYC.contig-100_156_7	835	27	182	183	294	298	606			656	717
CYC.contig-100_1565_13	1054	73	241	243	348	350	635	645	733		
CYC.contig-100_1630_5	839	8	147	187	294						
CYC.contig-100_165_17	677	24	169	178	286	289	594			612	667
CYC.contig-100_1677_5	850	25	195	197	324	326	601			651	732
CYC.contig-100_17173_1	1022	44	210	212	309	311	616	626	714		
CYC.contig-100_1738_3	1025	43	203	205	314	316	607	616	705		
CYC.contig-100_18329_2	784	9	151	148	242						
CYC.contig-100_1850_9	840	34	203	205	313	315	593			641	722
CYC.contig-100_188_6	603	56	208	202	305						
CYC.contig-100_1905_7	1036	63	236	238	336	338	621	631	720		
CYC.contig-100_1928_8	1035	66	239	241	339	341	637	644	729		

CYC.contig-100_1936_6	829			186	298						
CYC.contig-100_1938_5	839	32	184	186	298	303	556			660	721
CYC.contig-100_1970_10	840			175	284						
CYC.contig-100_1975_15	834	27	196	198	307	309	587			635	716
CYC.contig-100_20362_2	838	28	181	183	294	299	608			659	720
CYC.contig-100_20419_2	746	17	155	157	263	266	543				
CYC.contig-100_20715_4	586	11	183	179	273	275	582				
CYC.contig-100_2115_16	835	26	180	181	292	297	602			654	715
CYC.contig-100_2129_5	870	30	157	222	326						
CYC.contig-100_21962_2	583	36	195			283	579				
CYC.contig-100_2224_4	1036	58	225	227	341	343	637	647	735		
CYC.contig-100_2230_10	1032	61	234	236	334	336	619	631	720		
CYC.contig-100_2406_14	1007	56	228	220	340			598	687		
CYC.contig-100_245_34	760	3	153	164	260	264	562			577	631
CYC.contig-100_24805_2	832	6	160	190	296						
CYC.contig-100_25263_1	866	30	160	222	326						
CYC.contig-100_265_15	1343			272	392	407	690	700	789		
CYC.contig-100_2725_13	962	23	218	220	327						

CYC.contig-100_276_45	1352	82	275	277	402	416	699	709	798				
CYC.contig-100_28_66	1107	91	264	310	389	391	702	712	801				
CYC.contig-100_34_113	832	35	219	265	349					652	711		
CYC.contig-100_349_46	995	37	212	209	308	310	604	611	697			1209	1337
CYC.contig-100_35_94	1503	64	251	253	357	359	642	651	739				
CYC.contig-100_3538_5	776	15	152	162	242	244	519						
CYC.contig-100_35534_1	770	43	189	211	298								
CYC.contig-100_358_33	1130	74	245	247	357	373	659	669	757			1139	1266
CYC.contig-100_38509_2	594	15	184			286	589					1139	1266
CYC.contig-100_3943_10	955	23	177	178	288	294	517			629	688		
CYC.contig-100_3943_10	955	23	177	178	288	294	517			768	837		
CYC.contig-100_415_14	1016	24	187	202	311					643	703		
CYC.contig-100_415_20	779	33	189	190	301	306	537			601	660		
CYC.contig-100_4533_1	707	26	180	184	301	305	603					672	791
CYC.contig-100_464_5	1076	68	248	250	361	363	651	661	749			672	791
CYC.contig-100_4686_5	961	32	202	204	291	293	580	587	669			865	981
CYC.contig-100_473_17	629	6	146	162	261	263	500					865	981
CYC.contig-100_5141_5	834			204	301								

CYC.contig-100_5166_4	875	41	197	199	319					674	726	
CYC.contig-100_5166_4	875	41	197	199	319					620	700	
CYC.contig-100_518_24	1025	60	233	235	333	335	618	628	717			
CYC.contig-100_5212_2	1035	45	213	215	305	307	619	629	715			
CYC.contig-100_523_14	1130	72	250	252	363	374	659	669	757			
CYC.contig-100_523_14	1130	72	250	252	363	374	659	848	932			
CYC.contig-100_5347_6	1132	194	346	348	462							
CYC.contig-100_56_5	1036	61	238	240	346	348	633	643	736			
CYC.contig-100_5628_8	1051	57	223	226	322	324	637	647	741			
CYC.contig-100_5702_9	776	6	152			244	521					
CYC.contig-100_617_8	1274	67	238	240	353	355	639	649	738			
CYC.contig-100_656_21	789	5	145	146	240					617	676	
CYC.contig-100_66_15	857			215	323							
CYC.contig-100_6820_3	935	30	191	187	293			565	664			
CYC.contig-100_6823_2	923	31	182	185	283			555	637			
CYC.contig-100_6881_4	856	43	163	215	312							
CYC.contig-100_6927_1	1012			368	478							
CYC.contig-100_7007_5	907	5	158	160	260	262	557					

CYC.contig-100_704_34	1031	41	208	209	318	320	604	614	700		
CYC.contig-100_730_25	1012	48	217	219	315	310	620	629	716		
CYC.contig-100_75_44	963	24	213	220	331						
CYC.contig-100_76_46	1821	742	912	914	1027	1029	1314	1324	1413		
CYC.contig-100_76_69	942	27	187	189	290	295	523	560	656		
CYC.contig-100_815_27	873	44	207	209	323						
CYC.contig-100_8213_5	768	29	158	160	266	268	569				
CYC.contig-100_898_11	911	108	260	261	382					731	792
CYC.contig-100_928_29	1026	47	214	216	328	330	624	634	722		
CYC.contig-100_958_10	734	4	147	158	246	249	545			71	123
CYC.contig-100_958_14	861	7	153	164	248					603	661
CYC.contig-100_96_79	856	63	219	224	326						