Universidad de Concepción
Dirección de postgrado
Facultad de Ciencias Biológicas - Programa de Bioquímica y Bioinformática

Sistema automatizado de anotación de genes asociados a características probióticas desde secuenciación de nueva generación

Tesis para optar al grado de Magíster en Bioquímica y Bioinformática

BRIAN FRANCISCO GATICA ROCHA
CONCEPCIÓN – CHILE
2020

Profesor Guía: Alexis Salas Burgos
Dpto. de Farmacología, Facultad de Ciencias Biológicas
Universidad de Concepción

Esta tesis ha sido realizada en el en el Laboratorio de Diseño de Fármacos del Departamento de Farmacología de la Facultad de Ciencias Biológicas, Universidad de Concepción.

Profesor tutor

Dr. Alexis Salas
Facultad de Ciencias Biológicas
Universidad de Concepción

Comisión Evaluadora

Dra. Apolinaria García
Facultad de Ciencias Biológicas
Universidad de Concepción

Dr. Julio Villena
Centro de Referencia para Lactobacilos
CONICET, Tucumán, Argentina

Director de Programa

Dra. Amparo Uribe
Facultad de Ciencias Biológicas
Universidad de Concepción

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

GO : Ontología de Genes
HM : Modelos escondidos de Marcov (acrónimo del inglés Hiden Marcov Model)
ORF : Marco de lectura abierto
MSA : Alineamiento múltiple de secuencias
NS : Puntaje normalizado
SCFA : ácidos grasos de cadena corta (acrónimo del inglés Short Chain Fat Acids)

RESUMEN

Los probióticos son microorganismos vivos que cuando se administran en cantidades adecuadas confieren un beneficio para la salud en el huésped, cumpliendo además criterios de seguridad que consideran infectividad, patogenicidad, factores de virulencia, toxicidad y actividades metabólicas. Estos microorganismos actúan en un amplio espectro de beneficios, siendo los más estudiados la modulación de la composición/actividad de la microbiota endógena, la modulación del sistema inmune, y la modulación de las respuestas metabólicas sistémicas. No obstante, a pesar de los grandes avances actuales en bioinformática y particularmente en genómica comparativa, las bases de datos existentes para microorganismos probióticos comprenden solo características a nivel de especie-beneficio y no existe un sistema unificado que permita la anotación funcional y caracterización de los genes asociados a la actividad probiótica. En esta investigación se construyó un sistema de identificación de características probióticas basado en anotaciones por ontología de genes asociados a características probióticas, estableciendo los marcos de identidad (≥70%), cobertura (>90%) y puntaje normalizado (≥1.5) adecuados para una precisa imputación de las secuencias mediante la evaluación de distintas estrategias de agrupamiento desde 561,911 secuencias. Esto permitió establecer una metodología robusta para la construcción de una base de datos de secuencias de proteínas y desde ésta, definir modelos de perfiles escondidos de Markov (HMM). En este trabajo se definieron 10 categorías y 19 subcategorías asociadas a características probióticas establecidas por ontología génica en los grupos incluyentes y excluyentes. Entre los primeros se encuentran: 1) Sobrevivencia al tracto intestinal, 2) Producción de compuestos bioactivos, 3) Competitividad bacteriana, 4) Regulación positiva del sistema inmune, 5) Regulación negativa del sistema inmune, 6) Asociados a receptores Toll-like, 7) No clasificados relacionados con el sistema inmune. Mientras que en los excluyentes tenemos: 1) Resistencia a antibióticos; 2) Patogenicidad; y 3) Producción de compuestos tóxicos. Finalmente, utilizamos este sistema de anotación funcional de genes llamado "*PBDBsearch*" para la evaluación de 4 cepas bacterianas de *Lactobacillus* (1 *L. rhamnosus*, 1 *L. plantarum*, y 2 *L. fermentum*) clasificando a *Lactobacillus fermentum* L26 como el potencial probiótico más inocuo

y a *Lactobacillus rhamnosus* L134 como el potencial probiótico con mayores facultades.

ABSTRACT

Probiotics are live microorganisms which administered in adequate amounts, confer a health benefit to the host, having in addition safety criteria that consider infectivity, pathogenicity, virulence factors, toxicity, and metabolic activities. These microorganisms act in a wide spectrum of benefits, the most studied being the modulation of the composition/activity of the endogenous microbiota, the modulation of the immune system, and the modulation of systemic metabolic responses. However, despite the great advances in bioinformatics and particularly in comparative genomics, the existing databases for probiotic microorganisms comprise only at species-benefit characteristics level, and there is no unified system that allows the functional annotation and characterization of genes associated with probiotic activity. In this research, a probiotic characteristics identification system was built based on genes associated with probiotic characteristics annotations by ontology, establishing the identity (≥70%), coverage (> 90%) and normalized score (≥1.5) suitable frames for precise imputation of sequences by evaluating different clustering strategies in 561,911 sequences. This allowed us to establish a robust methodology for the construction of a protein sequence database and from it to define hidden Markov profile models (HMM). In this work, 10 categories and 19 subcategories associated with probiotic characteristics established by gene ontology were defined in the inclusive and exclusive groups. In the first group were included: 1) Survival in the intestinal tract, 2) Production of bioactive compounds, 3) Bacterial competitiveness, 4) Positive regulation of the immune system, 5) Negative regulation of the immune system, 6) Association with Toll-like receptors, 7) Not classified related to the immune system. Meanwhile, in excluding groups were considered: 1) Resistance to antibiotics; 2) Pathogenicity; and 3) Production of toxic compounds. Finally, we used this functional gene annotation system called "*PBDBsearch*" for the evaluation of 4 *Lactobacillus* bacterial strains (1 *L. rhamnosus*, 1 *L. plantarum*, and 2 *L. fermentum*), classifying Lactobacillus fermentum L26 as the safest potential probiotic and Lactobacillus rhamnosus L134 as the potential probiotic with more facuties.

1 INTRODUCCIÓN

1.1. Microorganismos Probióticos

Según la definición de la Organización de Alimentos y Agricultura (FAO) en conjunto con la Organización Mundial de la Salud, el término probióticos se atribuye a "microorganismos vivos que cuando se administran en cantidades adecuadas confieren un beneficio para la salud en el huésped" (Morelli y Capurso 2012), siendo corroborada el año 2013 por la asociación Científica Internacional de Probióticos y Prebióticos (ISAPP), quienes añadieron los límites y directrices para la clasificación de los microorganismos como probióticos (Hill et al. 2014a). En esta última definición se establece que, si bien la mayoría de los probióticos son extraídos desde la microbiota intestinal, para que un microorganismo sea considerado como probiótico, éste debe ser caracterizado y sus efectos deben ser corroborados mediante validación experimental, además de ser administrados en una dosis de un número mínimo de células viables (1×10^9 UFC al día) y cumplir con una serie de criterios de seguridad, considerando infectividad y patogenicidad, factores de virulencia, toxicidad y la actividad metabólica de los microorganismos (Ishibashi y Yamazaki 2001).

Los beneficios a la salud, al igual que los mecanismos moleculares mediante los que un probiótico puede ejercer su actividad son muy variados y dependen estrechamente del contexto biológico donde se desarrollan, considerando factores dependientes del huésped por sí solo (como hábitos alimenticios y estado de salud), la microbiota comensal presente, y la disponibilidad de factores abióticos que determinan la compatibilidad de coexistencia huésped-hospedador que finalmente determinan la manifestación de un efecto beneficioso para la salud. Estos efectos se han clasificado en tres categorías según su ocurrencia en microorganismos probióticos clasificándolos como raros, frecuentes y generalizados (Figura 1), donde podemos observar que un efecto raro se encuentra estrechamente relacionado con la cepa bacteriana (Hill et al. 2014b).

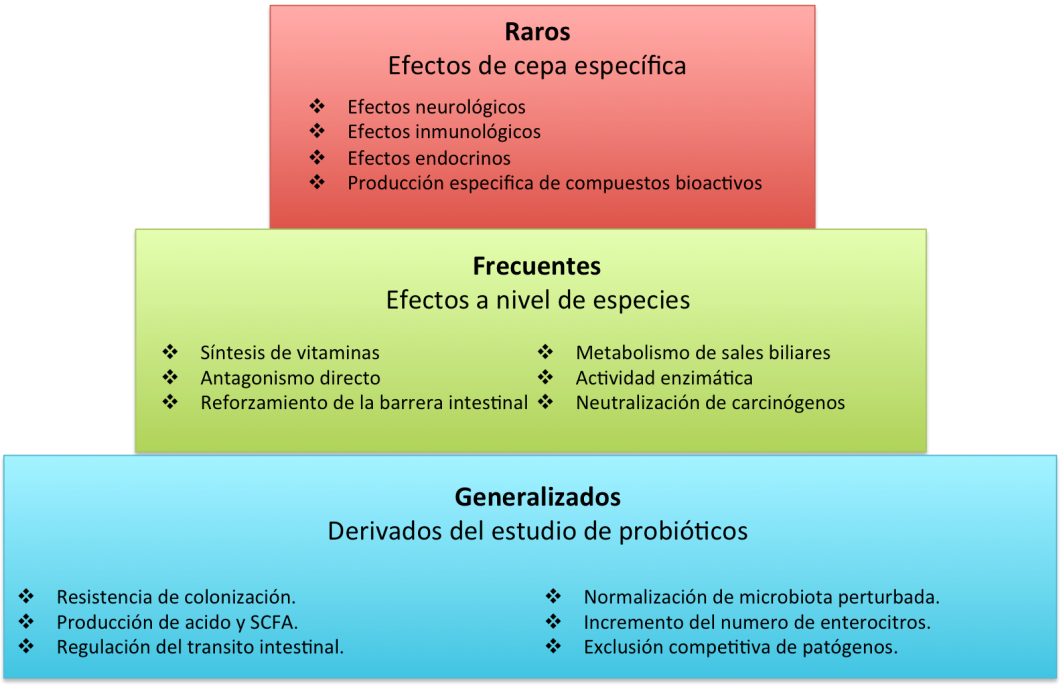


Figura 1: Clasificación de los efectos de microorganismos probióticos. Los efectos beneficiosos a la salud del huésped de microorganismos probióticos se encuentran distribuidos según su significancia y prevalencia en distintos niveles taxonómicos. Los mecanismos generalizados se encuentran entre géneros de especies probióticas, otros frecuentes se asocian a nivel de especie, y los más raros en sólo unas pocas cepas de una especie (modificado desde Hill et al. 2014).

En este contexto, los efectos beneficiosos para la salud del huésped son muy variados, donde los más estudiados son:

1.1.1. Modulación de la composición/actividad de la microbiota endógena:

La modulación de la composición/actividad de la microbiota endógena está mediada en gran parte gracias a las capacidades competitivas superiores de microorganismos frente a otros para asegurar su prevalencia en el tracto intestinal, entre los que se destacan la producción de bacteriocinas (péptidos antimicrobianos) (Dobson et al. 2012), sistemas toxina-antitoxina (Yamaguchi, Park, y Inouye 2011), mecanismos de adhesión (Ruas-Madiedo et al. 2006; Van Tassell y Miller 2011), formación de biopelículas, mecanismos de resistencia a pH extremo, degradación de sales biliares (como F1-F0

ATPasas, glutamato descarboxilasa, ureasas, arginina desaminasas, y mecanismos de reparación de ADN) (Cotter y Hill 2003).

1.1.2. Modulación del sistema inmune:

Los microorganismos probióticos son capaces de modular respuestas inmunológicas a través de mecanismos de regulación epigenéticos (Azam et al. 2014; Takagi, Kano, y Kaga 2015; Iyer et al. 2008) desencadenando respuestas inmunes por medio de la interacción con monocitos, macrófagos y células dendríticas. También, se ha descrito que influencian el sistema inmune a través de la modulación de la activación de receptores Toll-like, mediante el reconocimiento de motivos CpG no metilados en el ADN (secuencias con un alto contenido de G-C conservado en genomas bacterianos) y otros patrones moleculares asociados a microorganismos como el peptidoglicano, ácidos lipoteicoicos y exopolisacáridos (Ménard et al. 2010). En este contexto se ha descrito que *L. jeseni* es capaz de modular la respuesta inflamatoria del sistema inmune intestinal mediada por la interacción con las vías de señalización de TLR4 (Villena y Kitazawa 2014), *L. rhamnosus* posee el potencial de adyuvante inmunitario frente a ciertas enfermedades respiratorias bacterianas y virales, ya que aumenta los niveles de interleuquinas (IL-6, IL-10) e interferón γ (IFN-γ) en el tracto respiratorio e induce el incremento del número de células T $CD3^{+}CD4^{+}IFN\text{-}\gamma^{+}$ en pulmón (Villena et al. 2012), mientras que *L. casei* mejora los mecanismos inmunes y reduce el daño al tejido pulmonar en infecciones contra *S. pneumoniae* (Racedo et al. 2006).

1.1.3. Modulación de las respuestas metabólicas sistémicas:

Se ha determinado que microorganismos probióticos son capaces de favorecer las respuestas metabólicas del huésped de distintas formas, siendo capaces de reducir significativamente los niveles de glucosa en la sangre, y la resistencia a la insulina (Yao et al. 2017). Por ejemplo, la actividad hidrolasa de la sales biliares, que afecta a las hormonas de saciedad y las modulaciones endocrinas (Lebeer et al. 2018), e incluso participando en mecanismos directos e indirectos de señalización con el sistema nervioso

central, como los cambios asociados con alteraciones en el comportamiento emocional y en los aspectos estructurales y funcionales de la actividad cerebral causada por una mayor producción del ácido γ-aminobutírico (GABA), N-acetil aspartato y glutamato luego de la administración de *L. rhamnosus* (Janik et al. 2016).

Se ha reportado que la coexistencia simbiótica de estos microorganismos pueden desembocar una amplia variedad efectos positivos en la protección y tratamiento de enfermedades en el huésped, siendo beneficiosos frente a patologías de origen intestinal, como el tratamiento de la diarrea infecciosa (por infección con *C. dificile*, *H. pylori*, y rotavirus), diarrea asociada al consumo de antibióticos, y síndrome de colon irritable, donde se han demostrado mejoras significativas en el tratamiento con microorganismos de los géneros *Lactobacillus* y *Bifidobacterium* (Tabla A-I). Además, el tratamiento combinado de *Lactobacillus acidopHilus*, *Butyrivibrio fibrisolvens*, *Bacillus polyfermenticus*, *Lactobacillus plantarum*, *Lactobacillus fermentum*, o una combinación de *Lactobacillus acidopHilus* y *Bifidobacterium bifidum* inhibe significativamente el desarrollo de cáncer de colon (Yu y Li 2016), contribuyendo de manera sustancial a la reducción en la formación de tumores y la metástasis en cáncer de colon, junto con la erradicación de *H. pylori* (Peek y Blaser 2002).

Entre otros beneficios, se han demostrado beneficios a la salud urogenital, proponiendo estos microorganismos como alternativa de tratamiento contra el cáncer cervical por favorecer la eliminación de anomalías citológicas relacionadas con el virus del papiloma humano (Cha et al. 2012; Verhoeven et al. 2013), poseer actividad antitumoral contra el cáncer de mama como consecuencia de mecanismos de regulación epigenéticos en células MDA-MB-231 (Azam et al. 2014), al igual que en células derivadas de la leucemia mieloide regulando la proliferación celular promoviendo la apoptosis (Iyer et al. 2008).

1.1.4. Síntesis de vitaminas, y degradación de carbohidratos.

Entre otros beneficios que se han reportado en microorganismos probióticos, se ha determinado que proporcionan una fuente alternativa de compuestos bioactivos como vitaminas del grupo B, aumentando su disponibilidad para el huésped (LeBlanc et al. 2011), donde la información genética para la biosíntesis de estos compuestos es específica de algunas cepas probióticas principalmente del grupo de las bacterias ácido lácticas (Capozzi et al. 2012). Se ha descrito el aumento de la biodisponibilidad, en alimentos fermentados, de riboflavina por parte de cepas de las especies *L. plantarum*, *L. acidophilus*, *L. fermentum*, y *L. muocase*, y el aumento de la biodisponibilidad de folatos por parte de cepas de las especies *L. delbrueckii*, *L. Lactis*, *L. sakei*, *S. gallolyticus*, y *S. thermopHilus* (Levit et al. s. f.).

Por otro lado, la fermentación microbiana de carbohidratos complejos no digeribles y glicanos derivados de la dieta del huésped en el intestino humano tiene importantes consecuencias para la salud favoreciendo su digestión, participando en la degradación de sustratos complejos como las paredes celulares de las plantas, las partículas de almidón y la mucina. Sin embargo, a pesar de que se han identificado un gran número de enzimas asociadas a la degradación de carbohidratos principalmente en el género *Bacteroidetes*, la compleja relación entre la composición de la dieta, la microbiota intestinal y los resultados metabólicos que favorecen a la salud humana requieren más estudios (Flint et al. 2012).

1.2. Criterios para la selección de probióticos

En el marco de la definición de probiótico, así como en las directrices para su clasificación, se describe indirectamente que además de los beneficios a la salud del huésped, estos microorganismos deben poseer la capacidad de sobrevivir y proliferar en el adverso ambiente del tracto digestivo, para lo cual deben ser capaces de enfrentar las condiciones extremas de pH causadas por la presencia de ácidos gástricos y altos contenidos de sales biliares. Deben igualmente, poseer una mayor eficiencia competitiva frente a otros microorganismos para no ser desplazados por los mismos que compiten por recursos similares, y utilizan mecanismos como la mayor adhesión a células intestinales, formación de biopelículas, producción de bacteriocinas y peróxidos (Lebeer, Vanderleyden, y De

Keersmaecker 2008) para la defensa contra otras bacterias. También se considera que estos microorganismos deben poseer la menor resistencia a compuestos antibióticos posible con el fin de no inducir la resistencia a antibióticos en otros microorganismos por transferencia horizontal de genes. Además, se considera que la producción de aminas biógenas debe ser mínima y sus concentraciones varían según el tipo de compuesto (EFSA; Barbieri et al. 2019) .

Estas características incluyentes y excluyentes para la selección de microorganismos probióticos actualmente son determinadas mediante ensayos experimentales, para lo que existe una batería de ensayos preliminares utilizados en la clasificación de una especie de microorganismo como probiótico, dentro de los que se encuentran: 1) ensayo de hidrofobicidad de la superficie celular por adhesión microbiana a solventes (MATS, que determina las propiedades hidrofóbicas y electrostáticas de la superficie celular, Pelletier, C. et al. 1997); 2) ensayo de actividad antagonista contra patógenos bacterianos (Sandes et al. 2017); 3) determinación de la producción de peróxido de hidrógeno (Rabe y Hillier 2003); 4) ensayos susceptibilidad a antibióticos (Sandes, S. et al. 2017); 5) ensayo de susceptibilidad a jugos gástricos (GJS); 6) ensayo de susceptibilidad a sales biliares (BSS) (Silva et al. 2013); y 7) ensayo de producción de bacteriocinas.

1.3. Anotación y ontología de genes

Un genoma consiste en el total del material genético de un organismo (ADN o ARN), el cual contiene bioquímicamente codificada toda la información necesaria para el desarrollo y mantenimiento de su existencia. En el nivel más básico, esta información se encuentra estructurada en un código de 4 dígitos que representan un nucleótido en la molécula, y contiene almacenada la información para llevar a cabo todos los procesos y respuestas químicas que puede desarrollar una célula a lo largo de su vida. En esta estructura, los segmentos (códigos) del genoma pueden diferenciarse entre regiones no codificantes y regiones codificantes llamadas genes (que codifican la información para la formación de productos celulares como ARN o proteínas), que finalmente son procesados para desembocar en un determinado fenotipo celular. Sin embargo, una secuencia codificante para un gen no posee ningún significado biológico por sí sola, y es necesario extraer su contexto funcional por medio del proceso llamado anotación de genes.

El proceso de anotación genómica, consiste en buscar y caracterizar las subsecuencias de genomas, entregándoles un significado y contexto funcional a través de la búsqueda de regiones codificantes y comparaciones por similitud con secuencias ya conocidas. Según el nivel de complejidad la anotación se clasifica en 3 tipos: anotación a nivel de nucleótidos, a nivel de proteínas, y a nivel de procesos. En primera instancia, la anotación a nivel de nucleótidos consiste en la determinación del código genético, estableciendo su codificación específica e identificando las regiones codificantes (genes) a partir de estas secuencias nucleotídicas, lo que se basa principalmente en la identificación de marcos de lectura abierto (ORFs) en ambas direcciones de la secuencia (3' a 5', y 5' a 3') por medio de la identificación de marcas de puntuación conocidas como marcadores genéticos, marcadores de radiación híbridos, secuencias de término y patrones de señal que indican la localización de genes, repeticiones y regiones no codificantes (Stein 2001).

Principio de secuencia-estructura-función: Un principio básico de la biología es que la secuencia de aminoácidos de las proteínas determina su estructura tridimensional y función bioquímica, ya que tanto el plegamiento como la función se configuran a partir de patrones de interacciones físicas entre los átomos que constituyen la macromolécula. Este principio, descrito por primera vez por Anfinsen (1973), dio nacimiento a la biología computacional, abriendo la posibilidad de usar la secuencia de aminoácidos de una proteína para predecir sus propiedades funcionales y estructurales.

Independiente de la naturaleza de las secuencias (codificantes o no), una vez realizada la anotación a nivel de nucleótidos es necesario evaluar la anotación a nivel de proteínas, estableciendo el valor funcional de las secuencias utilizando comparaciones de similitud que determinan el parentesco y por consecuencia la relación funcional de una secuencia, la que a su vez, es asociada a un proceso biológico (anotación a nivel de procesos).

1.3.1. Ontología de genes

La función de una proteína se puede categorizar en tres diferentes clases interdependientes: función molecular (describe la actividad molecular, como síntesis, degradación, catálisis, entre otros), proceso biológico (describe la función general que se asocia al conjunto de funciones moleculares, como una vía metabólica o un proceso biológico), y componente celular (que describe la localización celular donde una proteína desarrolla su función). Este sistema de clasificación fue establecido en el proyecto "ontología génica" (GO), que consiste en un recurso bioinformático desarrollado en conjunto con la comunidad que proporciona información sobre la función de los productos genéticos, utilizando ontologías para representar el conocimiento biológico ("Gene Ontology Consortium: Going Forward" 2015).

Actualmente, la forma más utilizada y accesible para predecir la función de una proteína sin utilizar métodos experimentales es a través de la herencia por homología. Se dice que dos objetos son homólogos cuando realizan una función idéntica o equivalente en un mismo ámbito, y en el contexto bioquímico se define como la relación de correspondencia que ofrecen entre sí distintas moléculas o alguna de sus partes, que tienen origen y función semejantes (ASALE y RAE s.f.). De esta manera se puede entender que dos proteínas son homólogas cuando éstas desarrollan la misma función y presentan características comunes provenientes de un origen común. Desde el punto de vista genético esta caracterización se divide en dos términos según el origen de proveniencia de una secuencia, donde se llaman genes ortólogos aquellos homólogos funcionales presentes en dos especies distintas producto de un evento de especiación desde su ancestro común, y se llaman parálogos aquellos homólogos funcionales que se pueden encontrar dentro de una misma especie producto de un evento de duplicación y sus variantes son consecuencia de eventos de selección y mutación (Fitch 1970).

Existe un alto grado de conservación de las características que pueda exhibir un grupo de especies cuando se desarrollan bajo condiciones similares, y su diferenciación comprende eventos de evolución basados en selección natural guiados por la disponibilidad de factores abióticos. Por ejemplo, si examinamos el metabolismo de organismos superiores muy distintos, como los del reino animal y

vegetal, encontraremos que las diferencias son evidentes, sin embargo podemos observar que ambos confluyen en la utilización de la molécula de oxígeno para la producción de energía, lo cual es consecuencia de mayores requerimientos energético para su desarrollo y la conjunción de los factores abióticos disponibles en el transcurso de su evolución (como el clima y la disponibilidad de oxígeno en el ambiente) que favorecieron la prevalencia de estos organismos sobre otros (Holland 2006). Si bien, estos eventos de especiación son guiados selectivamente por los factores inherentes de la ecología de un ecosistema y la interacción de un organismo con éste, la ocurrencia o eliminación de una determinada característica en un organismo es un evento que ocurre a gran escala, y se manifiesta forma aleatoria como consecuencia de la acumulación de mutaciones selectivas en una o más secuencias específicas de su genoma que sean coherentes con su existencia, estableciendo al proceso de selección natural como motor de principal de la evolución fenotípica y evolución molecular por medio de la conservación selectiva (Ellegren 2008).

A nivel de funciones celulares, este grado de conservación se manifiesta significativamente y se mantiene en un amplio espectro de especies, donde mientras más cercano sea su parentesco, mayor similitud existe en los niveles inferiores (proteínas, y secuencias del genoma). Como se mencionó, el potencial de una célula de interactuar con su entorno está definido por la batería de genes y secuencias que contiene su genoma, la cual es expresada por medio del desarrollo de las funciones de sus componentes bioquímicos.

En este contexto, el nivel de conservación disminuye gradualmente a medida que se desciende en la escala biológica, pero es capaz de mantener el grado de concordancia necesario que permite el desempeño de una determinada característica en una célula. Dos proteínas que desarrollan una función homóloga no necesariamente poseen una estructura idéntica, y al descender un nivel organizacional, analizando su secuencia de aminoácidos y nucleótidos, se evidenciará con mayor énfasis las diferencias entre ambas. Por ejemplo, los motivos de hélice alfa son un componente recurrente de estructura secundaria en dominios de las proteínas de membrana debido su estabilidad con los componentes hidrofóbicos y polares de la membrana, no obstante, no todas las proteínas de

membrana que poseen este tipo de estructura tienen la misma secuencia, ni siquiera dentro de los segmentos que corresponden a estas estructuras, sin embargo conservan aspectos similares (sinónimos o idénticos) que son capaces de efectuar la misma función obteniendo un resultado similar, y es por medio de la comparación y análisis de estos aspectos similares conservados en las secuencias biológicas que se pueden establecer relaciones de parentesco entre proteínas, funciones celulares, y organismos.

1.3.2. Comparaciones de la similitud e identificación de genes

El método más certero que permite establecer relaciones de parentesco a través de la similitud entre secuencias es mediante el alineamiento de secuencias. En un alineamiento de secuencias dos o más secuencias son superpuestas de tal forma que se obtenga el mayor número de coincidencias entre los componentes de ambas, donde se pueden introducir espacios o *gaps* que indican la posible ganancia o pérdida de un componente (residuos en el caso de proteínas, y nucleótidos en el caso de ADN/ARN). Dependiendo del número de secuencias con las que se realiza un alineamiento, estos se pueden clasificar en alineamiento de pares (dos secuencias son alineadas) y alineamientos múltiples (más de dos secuencias son alineadas), los que pueden ser globales (que consideran toda la secuencia) o locales (que consideran parte de la secuencia). El alineamiento de secuencias múltiple entrega mayor información biológica al considerar un mayor número de secuencias, permitiendo establecer con mayor precisión sitios conservados entre las secuencias en un alineamiento. Este tipo de alineamiento es uno de los métodos más utilizados en análisis genómicos comparativos, identificación y cuantificación de regiones conservadas o motivos funcionales en familias de proteínas, búsqueda de homólogos funcionales, estimación de divergencia evolutiva entre secuencias y estimación de ancestros de secuencia (Kumar y Filipski 2007; 2007; Do y Katoh 2008a).

De manera general, la obtención del mejor alineamiento entre secuencias se realiza a través de la construcción de matrices de puntaje que muestran el grado de similitud entre cada componente de una secuencia y otra en el alineamiento, asignando un puntaje positivo a elementos concordantes, y penalizando con

puntajes bajos o negativos, dependiendo de las propiedades fisicoquímicas de los componentes diferentes en errores de coincidencia, o un salto en la secuencia (*gap*) (Figura 2), siendo las matrices de sustitución más utilizadas PAM (Point Accepted Mutation) y BLOSUM (Blocked Substitution Matrix)(Henikoff y Henikoff 1992).

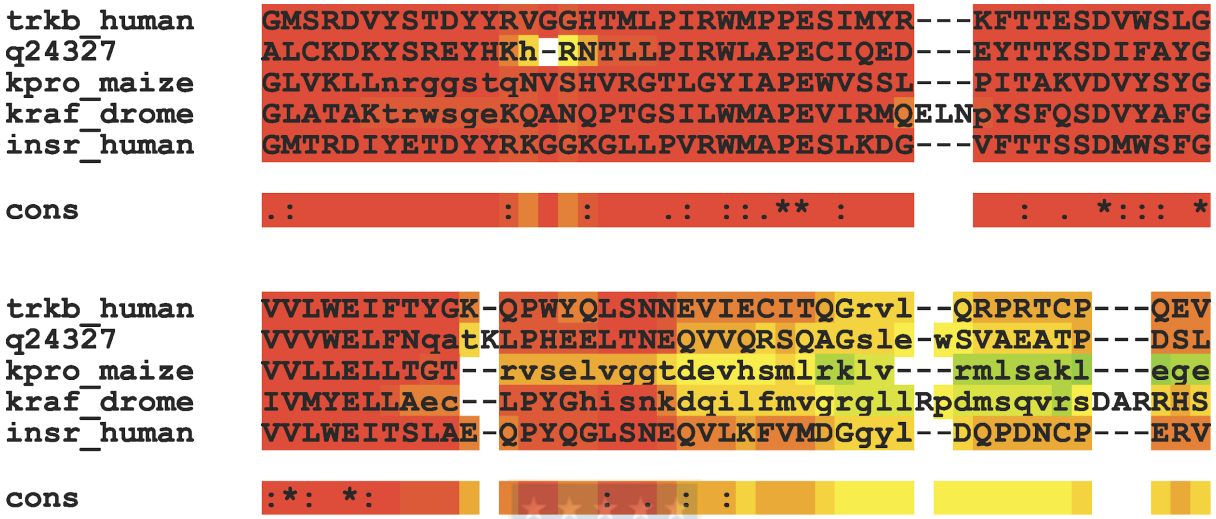


Figura 2: Alineamiento múltiple de secuencias de los datos kinase1_ref5 de BaliBase, combinando los software MUSCLE, MAFFT, POA, Dialign-T, T-Coffee, ClustalW, PCMA y ProbCons con M-Coffee. Los residuos alineados correctamente están en mayúsculas; los incorrectos están en minúsculas. El color de cada residuo indica el nivel de concordancia en cada posición, variado desde rojo azul. Rojo indica residuos alineados de manera similar entre todos los MSA individuales; Azul indica una concordancia muy baja entre los MSA. Se considera que los residuos de color amarillo oscuro, naranja y rojo están alineados de manera confiable. Extraído de (Notredame 2007).

Un alineamiento múltiple puede desarrollarse bajo tres modelos principales: 1) Modelos evolutivos complejos de inserción, deleción y mutación en múltiples secuencias; 2) Modelos de perfil de dimensionalidad fija para representar proteínas de familias específicas; 3) Modelos híbridos que combinan modelos probabilísticos con métodos tradicionales técnicas de alineamiento *ad hoc* (Do y Katoh 2008b). De estas tres aproximaciones, los modelos evolutivos alineamientos estático son capaces de entregar una representación más explícita de cambios en las secuencias biológicas como un proceso estocástico, derivado de la sustitución de aminoácidos siguiendo un proceso tiempo-reversible de Markov y la

inserción/deleción se trata como una adición/eliminación de vínculos imaginarios que unen las letras de una secuencia (Metzler 2003; Miklós, Lunter, y Holmes 2004). En modelos de perfiles escondidos de Markov (HMM) se construye una matriz frecuencias de probabilidad por posición específicas y luego se optimiza el modelo de acuerdo a criterios de verosimilitud, generando un perfil probabilístico de caracteres al que las secuencias son alineadas usando el algoritmo de Viterbi para buscar la mejor correspondencia entre cada secuencia individual y el perfil (Sippl 1999; Krogh et al. 1994). Esto permite que la utilización de perfiles HMM y sus variantes sean la base de muchas técnicas de identificación de familias de proteínas, presentando resultados más precisos que técnicas de identificación basadas en modelos evolutivos de inserción, deleción y mutación, para la detección de homólogos distantes.

1.3.3. Software de anotación de genes y bases de datos.

El actual desarrollo y masificación de las distintas técnicas de secuenciación de material genético ha generado que hoy en día sea una de las técnicas más utilizadas en investigación biológica, siendo aplicable como método de validación o determinación para casi cualquier línea de investigación que involucre organismos vivos. Este alto nivel de acceso ha incrementado de manera sustancial la información disponible de secuencias biológicas como genomas, transcriptomas, mutaciones, deleciones, polimorfismos de nucleótido simple, proteínas, y otros productos genéticos, lo que a su vez ha enriquecido cada vez más nuestro conocimiento sobre la funcionalidad y complejidad de los mecanismos epigenéticos a través de la clasificación empírica y teórica de los distintos componentes del genoma, progresando en el entendimiento de la conservación evolutiva y funcional del material genético, y como éste se manifiesta en el fenotipo de un organismo.

De forma general, la caracterización de genes utilizando anotación genómica automatizada es un procedimiento bastante sencillo, que comprende dos etapas esenciales: 1) Predicción de genes utilizando métodos "*ab initio*" y de comparaciones por homología, donde se determinan y caracterizan estructuralmente las secuencias; 2) Clasificación por comparación con Bases de datos, donde se clasifican funcionalmente los productos de la predicción utilizando

comparaciones de similitud basadas en alineamiento de secuencia. Para este procedimiento, se han desarrollado una gran cantidad de software que realizan la anotación funcional automatizada de secuencias biológicas, cómo MARKER2 (Holt y Yandell 2011), NCBI Eukaryotic Annotation Pipeline (Thibaud-Nissen et al. 2013), CAT (Fiddes et al. 2018), BRAKER1 (Hoff et al. 2016), y para la caracterización funcional en microorganismos los más utilizados son PROKKA (Seemann 2014) y el servidor RAST (Aziz et al. 2008).

Si bien, el procedimiento de anotación no presenta mayores dificultades, para ealizar una anotación precisa y de calidad es esencial escoger una base de datos con un alto grado de confiabilidad y coherente con las líneas de investigación que se desarrollan. Actualmente, los bancos de datos más extensos y utilizados son los de NCBI (Pruitt K. et al. 2006) y EMBL (McWilliam H. et al. 2013), que proporcionan e integran de forma transversal distintos niveles de información almacenados en diferentes bases de datos (organismos, genomas, secuencias nucleotídicas, secuencias proteicas, secuencias de ARN, y productos de secuenciación entre otros).

Las bases de datos biológicas constituyen una poderosa herramienta para entregar significancia biológica de nuevas secuencias y situarlas en el contexto de nuestra comprensión, en función de los procesos y composición biológica a través de la anotación (Stein L. et al. 2001), y se pueden clasificar en dos tipos según la información que proporcionan: 1) Bases de datos universales, que cubren información de todas las especies; 2) Bases de datos específicas, que contienen información específica para un cierto grupo de proteínas o especie (Wilkins et al. 2013). Existen numerosas bases de datos específicas que albergan la información de genes y sus productos, enfocadas en determinadas características biológicas (C. Chen, Huang, y Wu 2017), entre las que se pueden destacar para proteínas: Bases de datos orientadas a ontología de genes, GO (“Gene Ontology Consortium: Going Forward” 2015), PRO (Natale et al. 2014); Caracterización de familias y dominios de proteínas, PFAM (Baetman A. et al. 2004), HAMAP (Lima et al. 2009), PROSITE (Hulo et al. 2006); Vías y actividad enzimática, ENZIME (Bairoch 2000), Reactome (Croft et al. 2011), BRENDA (Schomburg et al. 2004); Interacciones proteína-proteína, BioGRID (Stark et al. 2006); y otras más específicas como

GenDB especializada en patógenos (Meyer, F. et al. 2003), y BATIBASE específica de bacteriocinas (Hammami, R. et al. 2010).

En el marco de microorganismos probióticos las bases de datos existentes como ProBio database (Tao, L. et al. 2017), y OptiBac (Sui, j. et al. 2002), ofrecen información sólo a nivel descriptivo de organismos y sus facultades, con 918 cepas, y 145 especies bacterianas, asociándolas con los efectos benéficos para la salud humana y las enfermedades que son capaces de prevenir, sin describir cuales son los mecanismos moleculares que desembocan en dichas actividades, lo que tampoco resuelve el gran número de protocolos y partidores patentados que se han desarrollado para la identificación de microorganismos probióticos patentados principalmente de los géneros *Lactobacillus* y *Bifidobacterium* (Magalhães, J. et al. 2008, Reque, E.et al. 2000). Esta poca profundización en el contexto de los mecanismos moleculares involucrados en los distintos beneficios a la salud del huésped, hace que esta información sólo sea capaz de describir y caracterizar los microorganismos a nivel de especie-beneficio, y no proporciona ninguna herramienta que permita evaluar a través de sistemas de anotación genética las características específicas de los microorganismos en función de sus facultades probióticas desde un punto de vista biológico basado en ontología de genes.

2 HIPÓTESIS

Las facultades de microorganismos probióticos son identificables a través de la caracterización específica de secuencias genómicas mediante mecanismos bioinformáticos de anotación utilizando una base de datos especializada para microorganismos probióticos basada en ontología de genes.

3 OBJETIVOS

Para responder a esta hipótesis, se plantearon los siguientes objetivos, general y específicos:

3.1 Objetivo general

Construir un sistema de categorías probióticas, basado en perfiles ocultos de Markov para la anotación de características probióticas desde secuencias obtenidas por métodos masivos de secuenciación genómica bacteriana.

3.2 Objetivo específicos

3.2.1. Definir categorías y construir una base de datos robusta y precisa de secuencias de proteínas y perfiles HMM asociados a características probióticas.

3.2.2. Desarrollar una herramienta automatizada de anotación funcional de las características probióticas a través de identificación y categorización de las secuencias de proteína involucradas en actividad probiótica desde datos de secuenciación.

3.2.3. Evaluar las facultades probióticas de 4 cepas bacterianas del género *Lactobacillus* utilizando el sistema de anotación desarrollado.

4 MATERIALES Y MÉTODOS

4.1. Construcción de base de datos

4.1.1. Búsqueda de genes y clasificación de secuencias proteicas

Basándose en los efectos y en las características probióticas más estudiados (modulación de la composición/actividad de la microbiota endógena, modulación del sistema inmune, modulación de las respuestas metabólicas sistémicas) como criterios de selección de organismos probióticos, se definieron tres niveles de clasificación para las secuencias de proteínas involucradas en procesos funcionales asociados a actividades probióticas. Establecimos 3 campos de clasificación (primer nivel) general de las características (excluyente, incluyente intrínseca, e incluyente extrínseca) que poseen 10 categorías (segundo nivel) basadas en los efectos generales de características probióticas y sus criterios de selección, las que a su vez poseen subcategorías (19 en total) específicas de funcionalidad (tercer nivel). La descripción de los niveles organizacionales de la base de datos se muestra en la sección de resultados.

Para la búsqueda de genes asociados a actividad probiótica, se llevó a cabo una investigación bibliográfica basada en la selección de procesos funcionales según clasificaciones de ontología génica. La selección de estos procesos funcionales, y criterios de búsqueda se determinó realizando un screening de todas las categorías de ontología de genes de la base de datos de GO (“Gene Ontology Consortium: Going Forward” 2015), seleccionando aquellos procesos involucrados en las 19 subcategorías definidas. La búsqueda de secuencias relacionadas con cada proceso funcional de ontología de genes fue realizada utilizando la base de datos de secuencias de proteínas UNIPROT (The UniProt Consortium 2015), seleccionando aquellas proteínas de referencia anotadas manualmente, aplicando los criterios de filtro para procesos funcionales y texto libre (cuando correspondía) utilizando los términos coherentes con cada subcategoría. Los criterios de búsqueda seleccionados para cada subcategoría, además del número de genes encontrados se muestran en las tablas de la sección de anexos (Tabla A-II – A-IV). Además, recopilamos la información complementaria de identificadores y publicaciones

asociadas utilizando la base de datos de NCBI con ENTREZ (Maglodtt, D. et al. 2005).

4.1.2. Determinación de parámetros para la construcción de la base de datos.

La definición de la metodología óptima de construcción de la base de datos de perfiles se realizó evaluando diferentes estrategias de agrupamiento por similitud en las que se variaron los valores de identidad y cobertura utilizando las proteínas seleccionadas desde Swiss-Prot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz) correspondiente a 561,911 secuencias (obtenidas el 15/04/2020) y la información asociada (nombre proteína, nombre del gen), que fueron agrupadas y analizadas en tres fases de agrupamiento.

4.1.2.1. Fase 1 de agrupamiento: Exclusión por redundancia.

Con el objetivo de establecer un marco de referencia y reducir los niveles de redundancia a nivel secuencia se realizó el agrupamiento de las secuencias con CD-HIT (Fu et al. 2012), estableciendo un marco de tolerancia en identidad/cobertura de 100%/100%, y determinando la heterogeneidad de los grupos resultantes mediante conteo de ocurrencias diferentes para los campos de nombre proteína, nombre del gen en los grupos resultantes. De la misma forma se evaluó la heterogeneidad en la clasificación en estos campos utilizando un marco de 95%/95% para identidad/cobertura, la cual se estableció como referencia para las siguientes estrategias de agrupamiento. Luego, evaluamos la clasificación (positiva - negativa) de los grupos resultantes en distintos agrupamientos con CD-HIT en marcos de identidad de 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, y 50%, y cobertura de 95%, 90%, 85%, 80%, 70%, 75%, 65%, y 60%, utilizando como referencia los valores asignados para cada campo en el grupo de referencia, es decir, para cada secuencia asignada dentro de un grupo en una estrategia de agrupamiento se tomaron como entradas válidas para cada campo (clasificación positiva), todas las entradas que exhibía el grupo al que pertenece en la de referencia 95%/95% I-C (identidad/cobertura), obteniendo una clasificación negativa cuando se encontraron ocurrencias nuevas dentro del grupo evaluado.

4.1.2.2. Fase 2 de agrupamiento: Exclusión por representatividad

Se seleccionaron las estrategias de agrupamiento obtenidos de la fase 1 con una porción mayor o igual de grupos clasificados positivamente que la de la referencia 95/95% I-C evaluada contra a la estrategia de agrupamiento 100/100% I-C (valores bajo 80% de cobertura y 65% de identidad fueron descartados). Posteriormente, para cada estrategia de agrupamiento (incluyendo la referencia 95/95% I-C) se construyeron bases de datos con los perfiles HMM de sus grupos y se evaluó la representatividad de cada grupo realizando el alineamiento de las secuencias con HMMER (Zhang y Wood 2003) contra cada base de datos correspondiente observando el puntaje normalizado (SN abreviado del inglés "Normalized Score", equivalente al puntaje obtenido dividido por el largo de la secuencia objetivo) del mejor alineamiento para cada secuencia y comparándolo con el rango de variabilidad que exhiben los puntajes obtenidos en la referencia 95%/95% I-C. Los alineamientos de secuencia para la construcción de los perfiles fueron realizados con MUSCLE (Edgar 2004). Las estrategias de agrupamiento que presentaban 3 cuartiles de valores máximos de NS no superaron el valor mínimo de la distribución teórica de la referencia 95/95% I-C fueron excluidas.

4.1.2.3. Fase 3 de agrupamiento: re-distribución de los grupos

Seleccionando la mejor estrategia resultante de la fase 2 se realizó una reasignación de las secuencias de cada grupo y se reconstruyó la base de datos de perfiles HMM con esta nueva distribución. Aquellas secuencias que presentaban valores mayores de NS en grupos distintos al original fueron reasignadas, y aquellas con valores de NS<1.5 dentro del grupo fueron excluidas.

4.2. Base de datos para anotación funcional de genes

Para el proceso de anotación funcional mediante análisis bioinformáticos se construyeron 2 tipos de grupos de datos, uno de secuencias de proteína (desde las secuencias de proteínas encontradas en la etapa de búsqueda), y otro de perfiles HMM. La base de datos de perfiles HMM para las secuencias seleccionadas en la etapa de caracterización y búsqueda de genes, fue construida utilizando los parámetros estimados en la sección anterior, la cual llamamos PROBIODB, y para

cada perfil HMM se asignaron las características funcionales de clasificación de primer, segundo y tercer nivel de acuerdo a las secuencias presentes en cada grupo, además de la bibliografía correspondiente, y los datos de referencias cruzadas de EMBL, NCBI y GO, UNIPROT, ENZIME y PROSITE.

La estructuración de la base de datos fue desarrollada utilizando MONGODB, definiendo una llave primaria (código identificador de 5 letras y 7 dígitos) para cada perfil y secuencia para el acceso a 4 colecciones correspondientes a: 1) Perfiles, que contiene los campos de clasificación funcional de facultades probióticas en tres niveles, nombre de gen, nombre de proteína, procesos biológicos, procesos funcionales, compartimento celular y función molecular; 2) Secuencias, que contiene los campos de clasificación funcional de facultades probióticas en tres niveles, nombre de gen, nombre de proteína, procesos biológicos, procesos funcionales, compartimento celular, función molecular y organismo; 3) publicaciones, que contiene la información bibliográfica asociada a cada secuencia de proteína en el perfil, presentando el título, resumen, autores y revista); 4) Referencias cruzadas, que contiene referencias cruzadas para las bases de datos de EMBL, NCBI, GO, UNIPROT, ENZIME y PROSITE. Además, para el caso de la colección de perfiles se estandarizaron los campos de información con la mayor ocurrencia en cada campo de las secuencias en cada perfil.

4.3. Algoritmo de anotación funcional automatizada de características probióticas

Para la anotación funcional automatizada de características probióticas se construyó un algoritmo en lenguaje de programación Python 3.7 llamado PBDBsearch, que realiza una anotación a nivel de nucleótidos utilizando PRODIGAL para la identificación de regiones codificantes (CDS), y la anotación funcional mediante la búsqueda por homología con HMMER y BLAST utilizando la base de datos construida PROBIODB (de perfiles HMM y secuencias correspondientemente), considerando los marcos de identidad, cobertura y puntaje de alineamiento determinados en esta investigación. Además, en este programa se desarrollaron gráficos para visualizar los resultados de anotación, utilizando las bibliotecas de matplotlib en Python.

4.4. Microorganismos y secuenciación genómica

Se evaluaron 4 cepas bacterianas del género *Lactobacillus* sin caracterizar, proporcionadas por el Laboratorio de Patogenicidad Bacteriana del Departamento de Microbiología, Universidad de Concepción. La extracción de ADN se realizó utilizando el kit UltraClean ® Microbial DNA Isolation (M.OBIO, EEUU), siguiendo el protocolo recomendado por el fabricante, y las muestras fueron secuenciadas utilizando la biblioteca Truseq DNA PCR libre en NovaSeq para lecturas de 150pb con 2Gb por muestra en el laboratorio MACROGEN, Seúl, Corea.

4.5. Ensambles de genoma y análisis bioinformáticos

La calidad de secuenciación se determinó utilizando FASTQC. Las lecturas con calidad inferior a 30 (puntaje de PHred) fueron removidas, se evaluó la presencia de adaptadores, y el análisis de contaminación se realizó con fastq-screen. Las estimaciones de genoma se realizaron utilizando el ensamblador basado en grafos de cadena A5, *scaffolding* con SSPACE y corrección de aperturas con GAPfiller. Finalmente, se eliminaron secuencias contiguas de largo inferior a 300 pb y se estimó un ensamble consenso utilizando CAP3. La calidad y las estadísticas de ensamble fueron determinadas usando códigos en lenguaje Python y el mapeo de las lecturas fue realizado con Bowtie-2.

La identificación de las especies de bacterias se realizó asignando filogenias con RAxML con el modelo GTR con optimización de tasas de sustitución, modelo de tasa de heterogeneidad GAMMA, con 500 réplicas de *bootstrap*, utilizando los alineamientos de secuencias de nucleótidos concatenados de los genes 16S, GyrA, GyrB, RpoB', y RpoB, genes identificados a partir de la anotación de 161 genomas de microorganismos del orden de los Lactobacillales con PROKKA descritos en la tabla TS2 y 8 grupos externos.

Finalmente, la identificación de características probióticas fue realizada utilizando el sistema de caracterización desarrollado en este trabajo de tesis: PBDBsearch utilizando la base de datos PROBIODB.

5 RESULTADOS

5.1. Construcción de base de datos

5.1.1 Proposición de categorías relevantes.

La estructura de clasificación determinada a través del análisis bibliográfico de características probióticas comprende tres niveles de clasificación. El primer nivel de clasificación consiste en una clasificación general de las características y consta de tres estados: 1) Excluyentes, de características funcionales que no poseen los microorganismos probióticos; 2) incluyentes intrínsecas, de características funcionales que poseen los microorganismos probióticos que no dependen de las condiciones; 3) incluyentes extrínsecas, de características funcionales que poseen los microorganismos probióticos que dependen de las condiciones y requieren de un examen más exhaustivo para su clasificación (por ejemplo, los mecanismos funcionales que influyen en el sistema inmune del hospedador).

El segundo nivel de clasificación se estructuró definiendo 10 categorías funcionales mayores para las actividades asociadas a probióticos de forma general, y por último un tercer nivel de 19 subcategorías de características funcionales más específicas para las características generales más diversas. La estructuración de las 10 categorías y las 19 subcategorías se muestran a continuación junto con el número de secuencias y perfiles de HMM resultantes luego de la construcción de la base de datos con 38,684 secuencias y 13,031 perfiles de HMM:

Categoría excluyentes:

E1. Resistencia a antibióticos: Contiene 2,262 secuencias de proteínas y 720 perfiles HMM asociados con resistencia a antibióticos y no posee subcategorías.

E2. Patogenicidad: Contiene 6,596 secuencias de proteínas y 2,272 perfiles HMM asociados con patogénesis y no posee subcategorías.

E3. Producción de compuestos tóxicos: Contiene 3,452 secuencias de proteínas y 1,112 perfiles HMM asociados y posee 1 subcategoría:

E3.1. Producción de aminas biógenas: Contiene 3,452 secuencias de proteínas y 1,112 perfiles HMM.

Categoría incluyentes intrínseca:

I1. Sobrevivencia al tracto intestinal: Contiene 3,452 secuencias de proteínas y 1,112 perfiles HMM asociados y posee 3 subcategorías:

I1.1. Resistencia a pH ácido: Contiene 6,641 secuencias de proteínas y 1,453 perfiles HMM.

I1.2. Resistencia a sales biliares: Contiene 22 secuencias de proteínas y 13 perfiles HMM.

I1.3. Actividad ureasa: Contiene 1,656 secuencias de proteínas y 138 perfiles HMM.

I2. Producción de compuestos bioactivos: Contiene 10,257 secuencias de proteínas y 3,328 perfiles HMM asociados y posee 3 subcategorías:

I2.1. Producción vitaminas del complejo B: Contiene 6,328 secuencias de proteínas y 1,951 perfiles HMM.

I2.2. Actividad lactasa: Contiene 522 secuencias de proteínas y 141 perfiles HMM.

I2.3. Degradación de lípidos y ácidos grasos: Contiene 3,407 secuencias de proteínas y 1,236 perfiles HMM.

I3. Competitividad bacteriana: Contiene 9,382 secuencias de proteínas y 4,255 perfiles HMM asociados, y posee 6 subcategorías:

I3.1. Producción de bacteriocinas: Contiene 24 secuencias de proteínas y 18 perfiles HMM.

I3.2. Sistemas toxina-antitoxina: Contiene 31 secuencias de proteínas y 27 perfiles HMM.

I3.3. Defensa bacteriana: Contiene 5,714 secuencias de proteínas y 2,561 perfiles HMM.

I3.4. Producción de peróxidos: Contiene 190 secuencias de proteínas y 85 perfiles HMM.

I3.5. Adhesión y formación de biopelículas: Contiene 211 secuencias de proteínas y 175 perfiles HMM.

I3.6. Resistencia a estrés oxidativo: Contiene 3,212 secuencias de proteínas y 1,389 perfiles HMM.

Categoría incluyente extrínseca:

En esta clasificación se encuentran principalmente las secuencias de proteínas bacterianas con la capacidad de influir o interactuar con el sistema inmune del huésped, y debido a la complejidad y variabilidad de los efectos y mecanismos involucrados, dependen de un análisis más exhaustivo para su discriminación en términos de beneficios en el huésped, y comprende las siguientes categorías:

IE.1. Regulación positiva del sistema inmune: Contiene 40 secuencias de proteínas y 37 perfiles HMM asociados, y no posee subcategorías.

IE.2. Regulación negativa del sistema inmune: Contiene 1790 secuencias de proteínas y 575 perfiles HMM asociados y no posee subcategorías.

IE.3. Asociados a receptores Toll-like: Contiene 46 secuencias de proteínas y 36 perfiles HMM asociados y no posee subcategorías.

IE.4. No clasificados de modulación del sistema inmune: Contiene 1771 secuencias de proteínas y 570 perfiles HMM asociados y no posee subcategorías.

La descripción de los patrones de función molecular de GO, los campos de texto, nombres de gen y códigos enzimáticos se encuentran en las tablas A-II, A-III y A-IV de la sección de anexos.

5.1.2. Determinación de parámetros para la construcción de la base de datos.

En la primera fase de agrupamiento se establecieron los marcos de referencia para seleccionar aquellas estrategias I-C (identidad-cobertura) cuyos grupos estuvieran compuestos por una clasificación homogénea en los campos de nombre de gen y nombre de proteína. Para esto decidimos evaluar el nivel de redundancia en la categorización de las secuencias en estos campos, agrupando las secuencias en 100/100% I-C, para la determinación de secuencias redundantes con más de una asignación en estos dos campos. Además, debido a la poca flexibilidad en términos de variabilidad de secuencia (requerida para la estimación de parámetros en la fase II) que presenta esta clase de agrupamiento, se evaluó de la misma forma el agrupamiento 95/95% I-C. Contrariamente a lo esperado, donde consideramos que en el agrupamiento de las secuencias 100/100% I-C no debiese existir una sobre clasificación en los campos de nombre de gen y proteína, y este debería ser mínimo en el agrupamiento 95/95% I-C, encontramos altos niveles de redundancia para ambos casos (Figura 3 y 4, Tabla I).

Encontramos que existe un alto nivel de redundancia en la clasificación y asignación de claves para los campos de nombre de gen y nombre de proteína al agrupar las secuencias con un marco de 100/100% I-C con 994 de 38,140 grupos sobre referenciados en el campo nombre de gen, y 5,420 de 38,140 grupos sobre referenciados en el campo nombre de proteína, determinando que para una misma secuencia el valor máximo de entradas para nombre de proteína y nombre de gen fueron de 44 y 55 respectivamente en un mismo grupo. Para el caso de la estrategia de agrupamiento 95/95% se encontraron 2,905 de 51,431 grupos sobre referenciados en el campo nombre de gen y 12,491 de 51,431 grupos sobre referenciados para el campo nombre de proteína, siendo los máximos valores encontrados 28 y 52 en un mismo grupo respectivamente (Tabla I).

Se puede observar que si bien existe una relación proporcional entre el número de entradas y el tamaño de los grupos (Figura 3 y 4, B y D), existe un alto nivel de redundancia en ambos campos de clasificación, donde el campo nombre de proteína se encuentra más sobre referenciado.

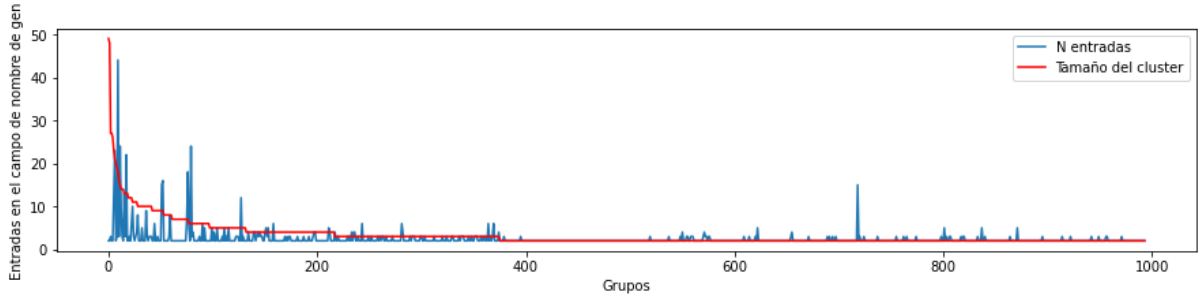
Si bien para la estrategia de agrupamiento 95/95% I-C se observan valores menores de sobre referenciación que en la estrategia 100/100% I-C (Tabla I, Figura 4), se debe tomar en cuenta que en este caso se estandarizaron las entradas válidas para cada secuencia designando como valores de calificación positiva todos todas las entradas al grupo al que pertenece esta secuencia en la estrategia de agrupamiento 100/100% I-C, no obstante el campo nombre de proteína presenta valores similares (Tabla I).

Debido a la heterogeneidad en ambos campos de clasificación es que decidimos estandarizar las entradas válidas en la evaluaciones del resto de las estrategias de agrupamiento, y la necesidad de considerar un rango de variabilidad (necesario para la construcción de perfiles HMM en la fase II) establecimos el agrupamiento 95/95% I-C como marco referencia para evaluación de las demás combinaciones I-C, asumiendo que el inducir este nivel de variabilidad (5%) no afectaría en términos de asignación en los campos nombre de gen o nombre de proteína.

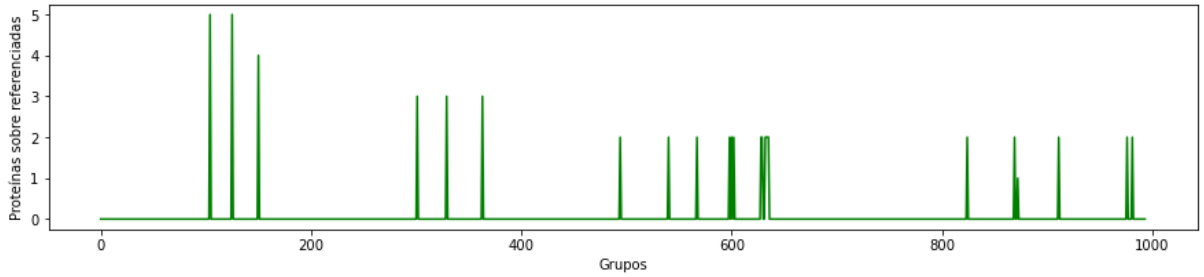
Tabla I: Valores máximos de redundancia para las estrategias de agrupamiento. Muestra los valores máximos de redundancia en las claves de nombre de gen y nombre de proteína en las estrategias de clustering a 100% identidad - 100% de cobertura y 95% identidad - 95% de cobertura.

Estrategia de agrupamiento i-c	Valor máximo de claves de gen en el clúster	Valor máximo proteínas sobre referenciadas para nombre de gen	Mayor tamaño de clúster con más de una entrada para nombre de gen	Valor máximo de claves de proteína en cluster	Valor máximo proteínas sobre referenciadas para nombre de proteína	Mayor tamaño de cluster con más de una entrada para nombre de proteína
100i-100c	44	5	49	55	16	114
95i-95c	28	25	95	52	19	95

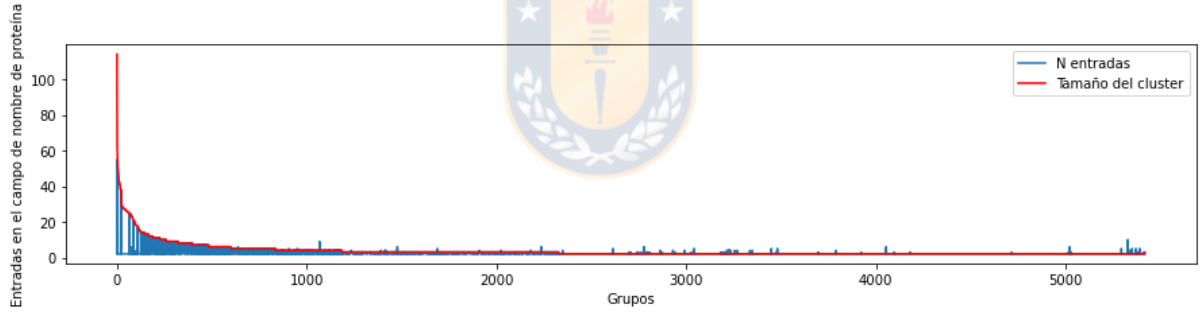
A



B



C



D

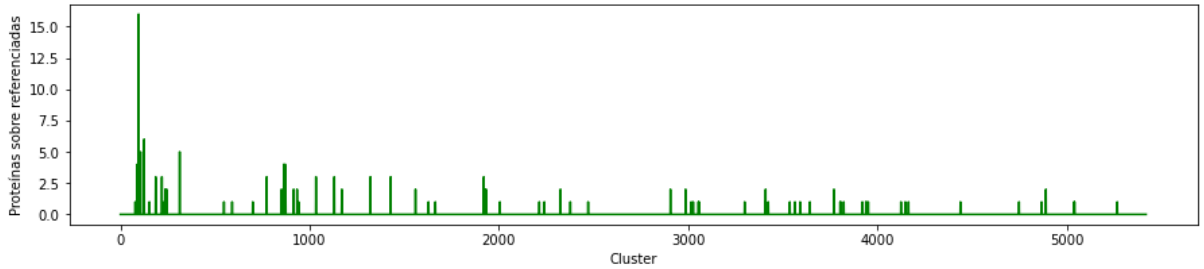
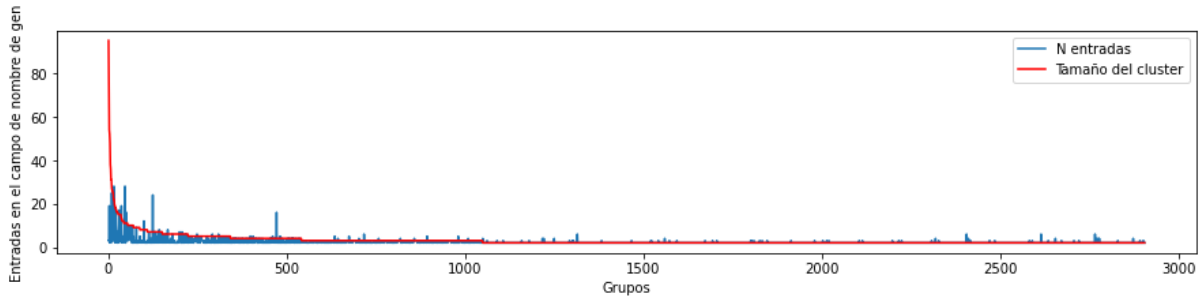
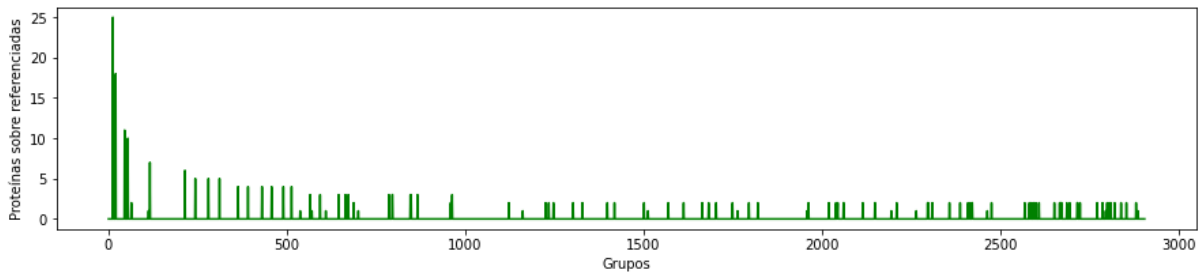


Figura 3: Redundancia para las claves de nombre de gen y nombre de proteína en clustering de secuencias 100% identidad - 100% de cobertura. A) Muestra el número de entradas asignadas para nombre de gen en cada cluster (verde) con más de un valor asignado ordenados por longitud, además del tamaño de cluster (rojo). B) Muestra el número de secuencias que poseen más de una clave para nombre de proteína dentro de cada cluster (verde), ordenados por longitud. C) Muestra el número de entradas asignadas para nombre de proteína en cada cluster (azul) con más de un valor asignado ordenados por longitud, además del tamaño de cluster (rojo). D) Muestra el número de proteínas que poseen más de una clave para nombre de proteína.

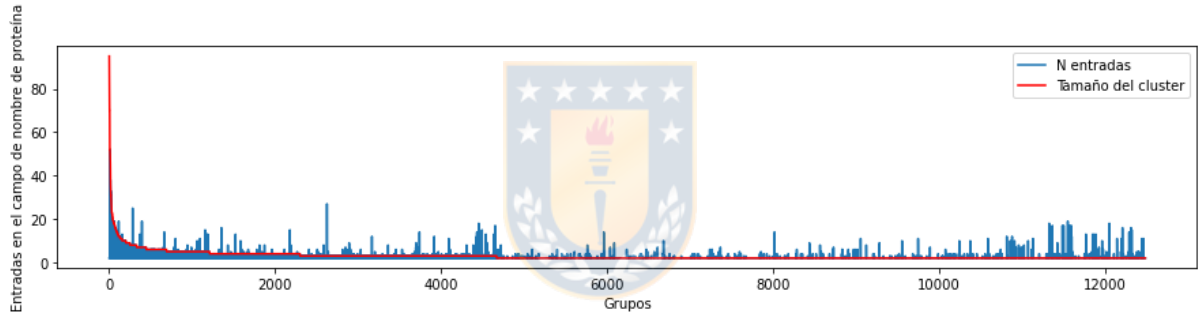
A



B



C



D

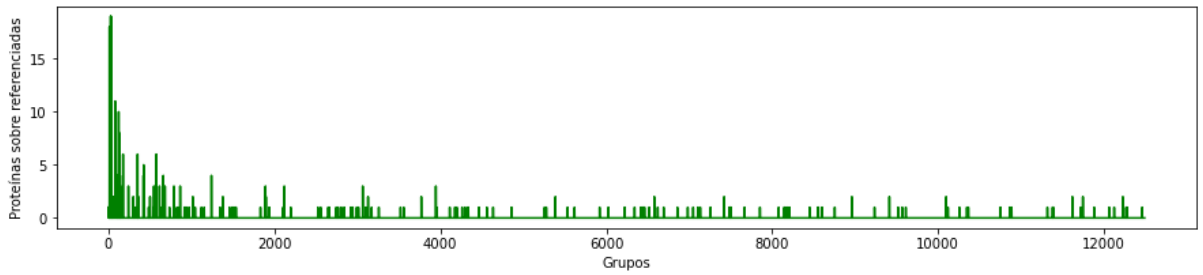


Figura 4: Redundancia para las claves de nombre de gen y nombre de proteína en clustering de secuencias 95% identidad - 95 % de cobertura. A) Muestra el número de entradas asignadas para nombre de gen en cada cluster (verde) con más de un valor asignado ordenados por longitud, además del tamaño de cluster (rojo). B) Muestra el número de secuencias que poseen más de una clave para nombre de proteína dentro de cada cluster (verde), ordenados por longitud. C) Muestra el número de entradas asignadas para nombre de proteína en cada cluster (azul) con más de un valor asignado ordenados por longitud, además del tamaño de cluster (rojo). D) Muestra el número de proteínas que poseen más de una clave para nombre de proteína.

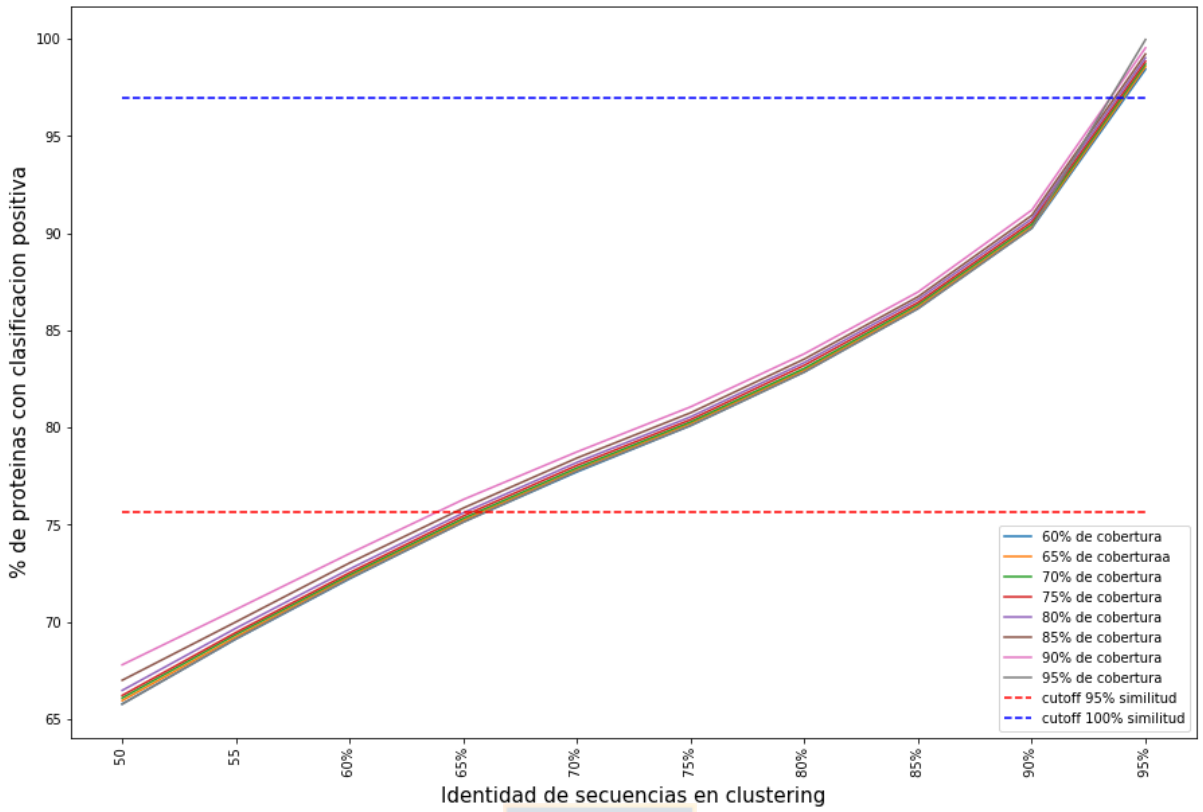


Figura 5: Proporción de grupos con asignación homogénea para el campo de nombre de gen en los distintos agrupamientos I-C. Muestra el porcentaje de proteínas anotadas correctamente dentro de las entradas válidas obtenidas desde la referencia (95/95% I-C) para los distintos valores de identidad y cobertura. Los valores de clasificación obtenidos para el agrupamiento 100%/100% I-C (97.2%, línea azul) y el agrupamiento de referencia 95/95% I-C (75.9%) se muestran marcados con líneas punteadas.

Al realizar la comparación frente a las referencias 100/100 I-C y 95/95 I-C para los campos nombre de gen y nombre de proteína pudimos constatar que sólo utilizando valores de restricción de identidad del 95% en el campo de nombre de gen se logra alcanzar valores iguales o superiores en términos de calificación positiva con respecto al agrupamiento 100/100 I-C (97.2% de grupos bien anotados), y utilizando valores de identidad superiores al 65% los alineamientos presentan valores iguales o superiores con respecto a la referencia 95/95 I-C (75.9% de grupos bien anotados).

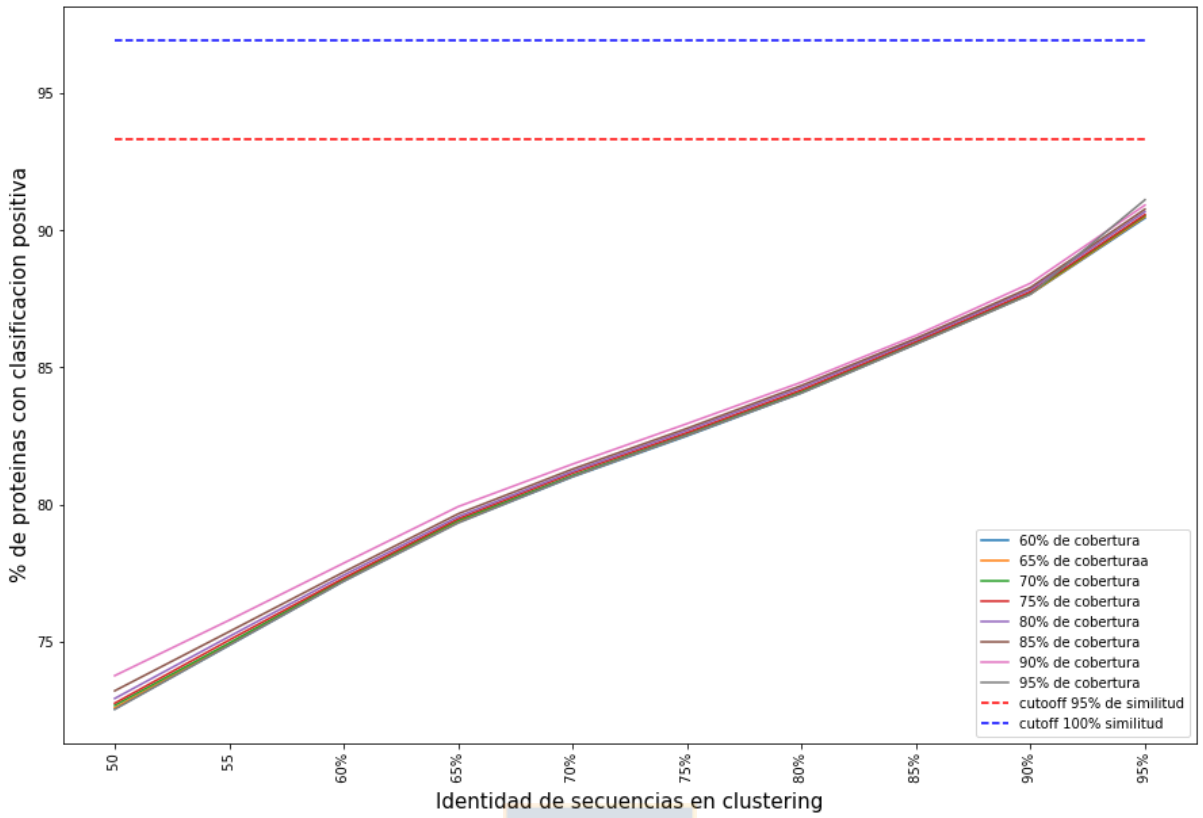


Figura 6: Proporción de grupos con asignación homogénea para el campo de nombre de proteína en los distintos agrupamientos I-C. Muestra el porcentaje de proteínas clasificadas correctamente dentro de las entradas válidas obtenidas desde la referencia (95/95%) para las distintas estrategias de agrupamiento de identidad y cobertura. Los valores de clasificación obtenidos para el agrupamiento 100%/100% I-C (96.3%, línea azul) y el agrupamiento de referencia 95/95% I-C (93.4%) se muestran marcados con líneas punteadas.

En el caso del campo de nombre de proteína, ninguno de los valores de identidad y cobertura evaluados logró superar los valores de las referencias 100/100 I-C y 95/95 I-C (96.3% y 93.4% respectivamente), esto refleja que en este campo de clasificación existe una mayor ambigüedad al momento de asignar estos valores de anotación a las secuencias, presentando múltiples sinónimos para una misma secuencia producto de discrepancias en caracteres específicos, adicción de identificadores (por ejemplo, Proteína A y Proteína A2), e incluso variantes de secuencias idénticas de acuerdo a su línea taxonómica. Debido a que en este caso el campo nombre de proteína no nos permitió realizar una discriminación en primera instancia en la correcta asignación de secuencias dentro de los grupos se seleccionó el campo de nombre de gen para la selección preliminar de las estrategias de agrupamiento, descartando todas las estrategias de agrupamiento en las que se utilizaron valores iguales inferiores a 65% de identidad.

Por otro lado, no se observaron mayores cambios con respecto variaciones de cobertura (Figuras 5 ,6 y 8) debido a la utilización de identidad global (que considera toda la secuencia objetivo) en los agrupamientos con CD-HIT, ya que al filtrar las comparaciones por este parámetro se restringen de forma recíproca los rangos de cobertura que muestran las secuencias, mostrando una relación directa entre identidad (global) y cobertura.

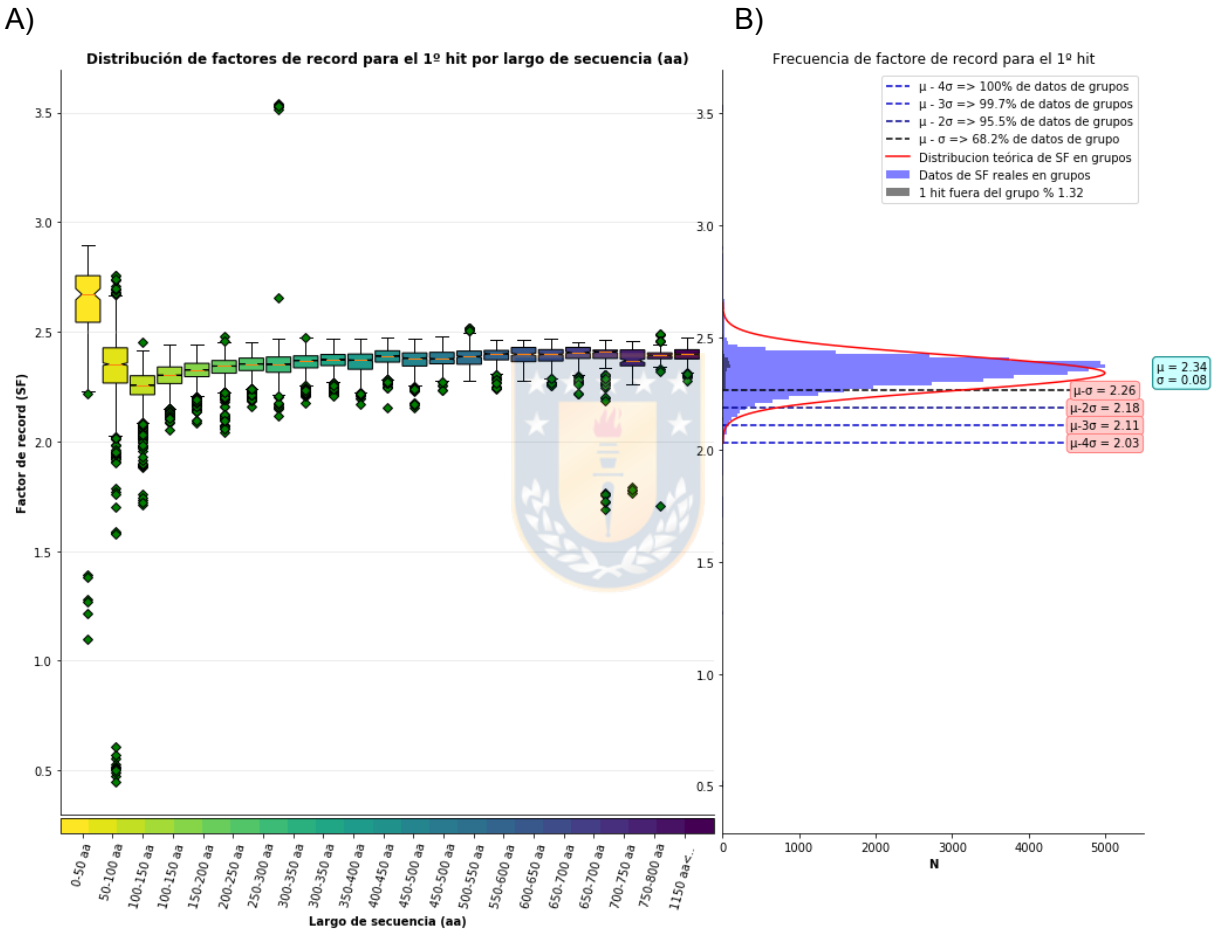


Figura 7: Distribución de puntaje normalizados obtenidos para el primer hit de cada secuencia de los grupos de referencia 95/95 I-C analizada contra la base de datos del agrupamiento de referencia 95/95 I-C. A) Muestra la distribución de los puntajes normalizados del primer hit de secuencias en grupos en rangos de longitud de secuencia. B) histograma del set de datos correspondiente al total de los puntajes normalizados. La distribución real de los datos (azul) se acopla con una curva de distribución normal teórica (línea roja). Además, se muestran los valores de media y desviación estándar correspondientes. La proporción de las secuencias cuyo primer hit no corresponde al grupo raíz se muestra en color amarillo (1.32% de un total 46303 secuencias).

Para la evaluación del comportamiento de los puntajes normalizados del primer *hit* utilizando una base de datos de perfiles HMM creada con las mismas secuencias analizadas, determinamos que los puntajes normalizados obtenidos para el primer *hit* dentro de los rangos de los distintos valores I-C evaluados se comportan siguiendo una distribución normal, donde el promedio y desviación estándar para la referencia 95/95 I-C corresponden a 2.34 y 0.08, respectivamente (Figura 7). Esto nos permitió establecer un valor inferior de corte (4 desviaciones estándar bajo el promedio) de 2.03, el cual utilizamos para discriminar y seleccionar los demás parámetros de agrupamiento en función de la correcta asignación de las secuencias de cada grupo.

Realizando el mismo procedimiento de evaluación de parámetros en los siguientes agrupamientos (evaluación de cada secuencia contra base de datos de perfiles HMM de los grupos), descartamos todas aquellas estrategias de agrupamiento que presentaban al menos un cuartil por debajo del mínimo 2.03, donde el valor mínimo de identidad (70%) no alcanzó a ser excluido (Figura 8). Finalmente, decidimos utilizar un valor de cobertura de 90% para el agrupamiento con CD-HIT, con el objetivo de considerar la mayor parte de las secuencias, de forma que los perfiles HMM resultantes reflejaran la mayor probabilidad de la secuencia de proteínas completas sin caer en un marco redundante de asignación redundante (95% de cobertura). El puntaje normalizado de corte de 1.48≈1.5 (correspondiente el puntaje normalizado mínimo teórico según la distribución de estos datos) para la asignación de las secuencias en un perfil HMM.

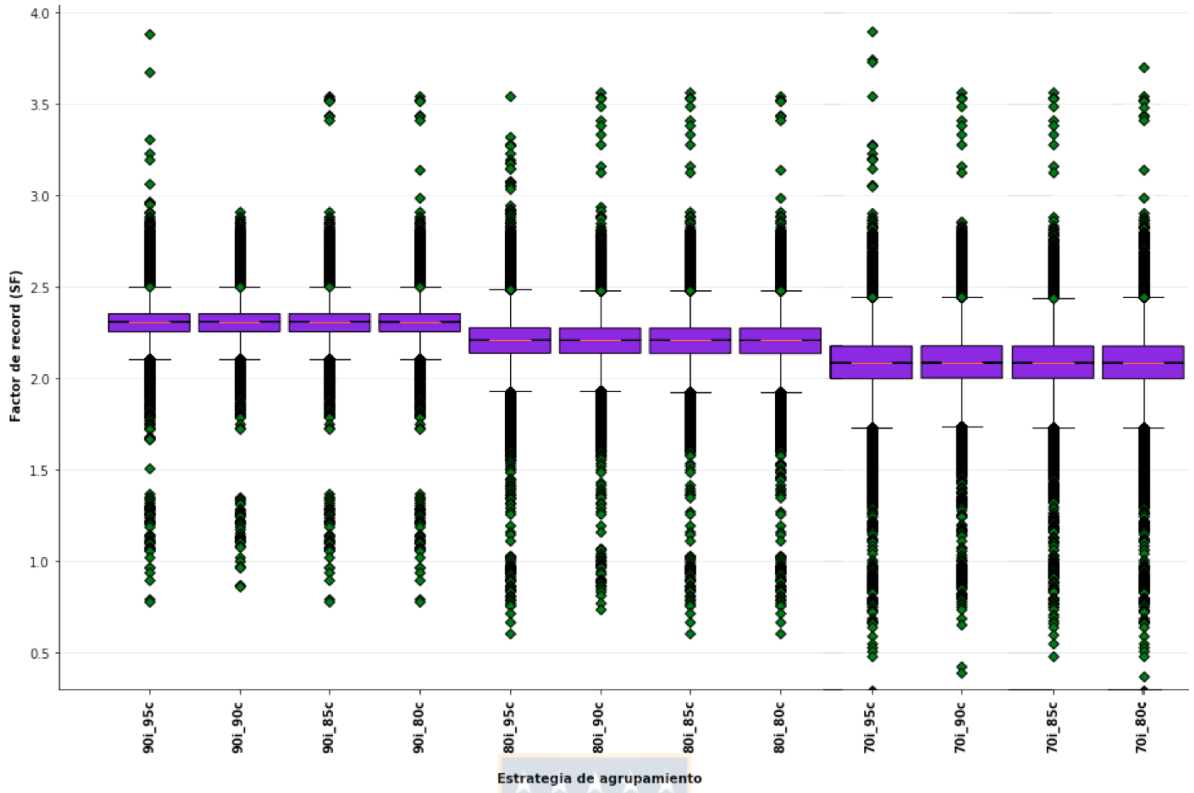


Figura 8: Distribución de factores de puntaje para el primer hit de cada estrategia agrupamiento. Muestra la distribución agrupada de SF en el análisis de perfiles de secuencia de las estrategias de agrupamiento de la fase II.

Tabla II: Estadísticas de estrategias de agrupamiento. Muestra los valores de número de secuencias, desviación estándar, media, secuencias correctamente asignadas, secuencias mal asignadas, número de secuencias por el rango de las estrategias de agrupamiento.

Estrategia de agrupamiento i-c	Secuencias en Grupos	Desviación estándar	Media de SF	correctamente asignadas	SF 2 - 1.5	SN 1.5 - 1	SF < 1	Mal asignados
90i_90c	28544	0,095	2,299	28404	94	40	6	388
90i_85c	28662	0,097	2,299	28480	139	37	6	383
90i_80c	28700	0,097	2,298	28512	145	37	6	393
80i_95c	47507	0,122	2,202	45963	1473	32	39	1007
80i_90c	66007	0,113	2,202	63916	2047	25	19	1381
80i_85c	66544	0,117	2,201	64239	2244	22	39	1314
80i_80c	66814	0,117	2,200	64434	2311	30	39	1373
70i_95c	112642	0,145	2,083	83266	29143	159	73	4818
70i_90c	109575	0,138	2,086	81994	27454	78	49	4673
70i_85c	111227	0,140	2,085	82838	28239	86	64	4840
70i_80c	117812	0,143	2,063	84573	29886	112	85	5012

5.2. Sistema de anotación funcional automatizada de características probióticas

Se desarrolló un sistema de anotación automatizado para la caracterización funcional de genes probióticos construyendo un software en lenguaje Python 3.7 que llamamos PBDBsearch (Figura 9). De forma general el software involucra 2-3 etapas de procesamiento según las opciones que se utilicen, realizando identificación de regiones codificantes (CDS) con PRODIGAL, la anotación funcional mediante la búsqueda por homología basado en puntajes de similaridad con HMMER y BLAST utilizando las bases de datos de perfiles y proteínas de PROBIODB, y finalmente la generación de salidas gráficas desde los resultados. Con el objetivo de ofrecer una mayor versatilidad de uso, decidimos habilitar la opción para escoger la etapa de inicio de los análisis dependiendo de los datos de entrada disponibles (genomas, transcriptomas, secuencias proteínas), pudiendo utilizar los formatos de entrada de archivos estándar FASTA o GBK.

La identificación de CDS con PRODIGAL es realizada con el procedimiento estándar para un solo genoma, considerando genes cerrados y utilizando la tabla de

traducción número 11 utilizada para bacterias, arqueas, virus procarióticos y proteínas de cloroplasto (Nakamura, Gojobori, y Ikemura 2000). Posteriormente la anotación funcional de genes probióticos es realizada utilizando BLAST filtrando los resultados por 70% de identidad, 90% de cobertura, y considerando un valor de error de 0.00001 de forma paralela se utiliza HMMER para los alineamientos con los perfiles HMM y los resultados son filtrados considerando un puntaje normalizado mínimo de 1.5 y un valor de error de 0.00001. Finalmente, para la representación gráfica del genoma se desarrolló un módulo basado en matplotlib que dibuja el genoma de forma lineal (Figuras 15-22), marcando las regiones codificantes según la categoría de actividad probiótica a la que pertenecen, pudiendo ampliar una región específica si se requiere.

Las opciones están disponibles en PBDBsearch y su forma de utilización son las siguientes:

i : Archivo de secuencias de entrada. El formato de entrada predeterminado es un archivo estándar fasta.

if : Formato de archivo de entrada. Utilice fa para fasta y gb para archivos genbank.

sf : Comenzar desde la etapa. El valor predeterminado es 0, use 1 para comenzar desde el archivo de anotación (faa o gbk), 2 solo dibujar genes (se requiere un archivo tabulado de ubicación de genes de la etapa anterior).

db : Directorio de ruta de la base de datos.

t : Número de subprocesos.

o : Nombre del archivo de salida.

ow : Sobrescribir archivos de salida.

bi : Filtro de identidad en BLAST, predeterminado es 70%.

bc : Filtro de cobertura en BLAST, predeterminado es 90.

ns : Filtro de puntuación normalizada en HMMER. Defecto es 1.5.

pt : Código de tabla de traducción que se utilizará. El valor predeterminado es 11, los números del 1 al 33 son válidos.

pm : Modo para establecer en PRODIGAL ("single" o "meta"). El valor predeterminado es "single".

b : Mantener archivo de salida de resultados de Blast. el valor predeterminado es 0 (desactivado). Use 1 para habilitar resultados de BLAST sin procesar en la salida.

hm : Mantener el archivo de salida de resultados HMMER. el valor predeterminado es 0 (desactivado). Use 1 para habilitar resultados HMMER sin procesar en la salida.

cr : Identificadores de referencia cruzada en el archivo de salida. Los números del 0 al 5 son válidos. El valor predeterminado es "0,3" (Uniprot y RefSeq). Utilizar 0 para Uniprot, 1 para KEGG, 2 para EMBL, 3 para RefSeq, 4 para Pfam, 5 para PROSITE.

zf : Archivo de entrada zoom-ranges. Guarda imágenes ampliadas de la gráfica de resultados de salida. Es un archivo separado por tabulaciones

con 3 columnas, 1: posición inicial, 2: posición final, 3: ID de cadena ("3" para cadena de 3 ', "5" para cadena de 5' y "53" para ambas).

h : Muestra el mensaje de opciones.

Los archivos de salida de PBDBsearch son los siguientes:

Sequences.fnn: Archivo de salida de la anotación de genes en formato fasta estándar de secuencias de nucleótidos.

Sequences.faa: Archivo de salida de la anotación de genes en formato fasta estándar de secuencias de proteína.

blast_filtred.tsv: Archivo de resultados filtrados de BLAST en formato tabulado con 13 columnas (secuencia, mejor acierto, largo de secuencia, inicio, término, inicio del mejor acierto, fin del mejor acierto, valor de error, puntuación, puntuación bits, Identidad, y cobertura)

hmmer_filtred.tsv: Archivo de resultados filtrados de HMMER en formato tabulado con 6 columnas (secuencia, mejor perfil, puntaje normalizado, valor de error, puntaje, bias).

Profiles_results.tsv: Archivo de resultados de anotación funcional de perfiles HM en formato tabulado de 18-23 columnas (secuencia, mejor acierto, categoría, subcategoría, nombre de proteína, nombre de gen, función, función molecular GO, compartimento celular GO, proceso biológico GO, proteínas en grupo del perfil, identificadores PUBMED, mejor hit por BLAST, nombre proteína del mejor hit, nombre del gen del mejor hit, organismo del mejor hit, identificadores PUBMED del mejor hit, y las columnas 18-23 con códigos de referencias cruzadas para el mejor hit).

Proteins_results.tsv: Archivo de resultados de anotación funcional con BLAST en formato tabulado de 13-18 columnas (secuencia, mejor acierto, categoría, subcategoría, nombre de proteína, nombre de gen, organismo, función, función molecular GO, compartimento celular GO, proceso biológico GO, identificadores PUBMED, y las columnas 13-28 con códigos de referencias cruzadas).

proteins.png y proteins.svg: Archivos de imagen en formato PNG y SVG de la representación gráfica de los genes en la búsqueda por homología con BLAST.

profiles.png y profiles.svg: Archivos de imagen en formato PNG y SVG de la representación gráfica de los genes en la búsqueda por homología con HMMER.

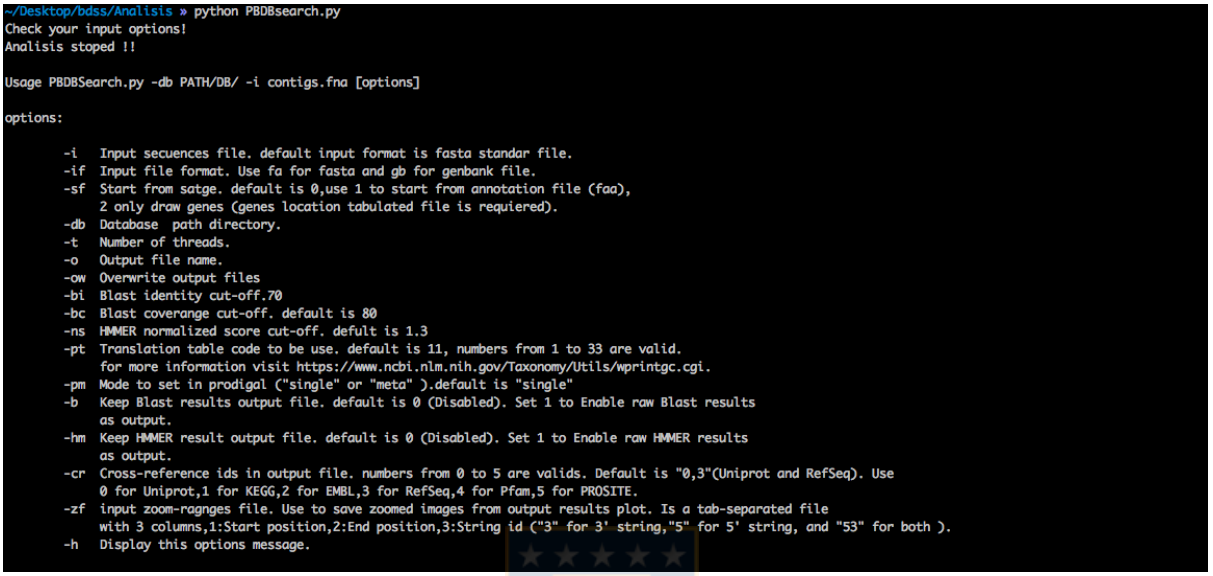


Figura 9: Visualización en terminal del panel de ayuda de PBDBsearch. Muestra la forma de uso de PBDBsearch además de las distintas opciones disponibles.

5.3. Ensambles de genoma y análisis bioinformáticos

5.3.1. Ensambles de genoma

Considerando el tipo de lecturas secuenciadas para resolución de los genomas bacterianos (lecturas pareadas de 150 pb) se logró obtener genomas integrales con poca segmentación (Tabla III). No obstante, para los cuatro casos no se logró determinar el genoma completo, obteniendo el mejor de los casos un ensamble de 34 *contigs* (*Lactobacillus* L90). Si bien para todos los ensambles se utilizó número de lecturas para obtener una profundidad sobre 1200x por base, lo cual supera el muestreo estándar sobre 100x, y valores de N50 cercanos o sobre 50kb, considerados como parámetros de un genoma de buena calidad (Chaisson, Wilson, y Eichler 2015; Miller, Koren, y Sutton 2010), en este caso el tamaño de las lecturas utilizadas influyó considerablemente en la segmentación.En un análisis más exhaustivo de las regiones terminales de los *contigs* obtenidos, encontramos que la mayor parte de estas regiones constituían elementos transponiles o regiones en tándem mayores a 200 pb, produciendo que al momento de intentar ensamblar este tipo de lecturas se localicen en dos o más regiones del genoma, generando quiebres en la contigüidad, a causa de la imposibilidad de determinar cuál de las secciones terminales continúa la secuencia, lo cual podría mejorarse utilizando lecturas de secuenciación larga (Chaisson, Wilson, y Eichler 2015).

Tabla III: Estadísticas de ensamblaje de 4 cepas de *Lactobacillus*. Muestra las estadísticas de los genomas ensamblados de las cuatro especies de *Lactobacillus* (*Lactobacillus* L26, *Lactobacillus* L33, *Lactobacillus* L90, *Lactobacillus* L134).

Ensamble	L26	L33	L90	L134
Contigs	93	70	34	46
Tamaño de genoma	1972381	1916400	32934444	2943062
Contig mayor	100918	168630	429317	291377
N50	44446	58118	186205	131538
Lecturas	28320442	32341348	29570688	33262346
Lecturas corregidas	27978449	31941513	29233376	32941105
% de lecturas usadas	98.79	98.76	98.86	99.03
Media de cobertura	1933	2324	1200	1628
%GC	51.99	52.42	44.37	46.61

Para *Lactobacillus spp*. L26 se obtuvo un ensamble de genoma de 1,972,381 pb con un N50 de 44,446 y un %GC de 51.99 con 1917 secuencias codificantes (Figura 10). Para *Lactobacillus spp.* L33 se obtuvo un ensamble de genoma de 1,916,400 pb con un N50 de 58,118 y un %GC de 52.42 con 1,867 secuencias codificantes (Figura 11). Para *Lactobacillus spp.* L90 se obtuvo un ensamble de genoma de 3,293,444 pb con un N50 de 186,205 y un %GC de 44.37 con 3,092 secuencias codificantes (Figura 12). Para *Lactobacillus spp.* L134 se obtuvo un ensamble de genoma de 2,943,062 pb con un N50 de 131538 y un %GC de 46.61 con 2,768 secuencias codificantes (Figura 12).

5.3.2. Identificación de especies

El análisis de filogenia utilizando las secuencias de nucleótidos concatenados de los genes 16S, GyrA, GyrB, RpoB'y RpoB ubicó a las cuatro especies analizadas en ramas profundas dentro del género *Lactobacillus* (Figura 9), determinando que las cepas L26 y L33 pertenecen a la especie *Lactobacillus fermentum*, L90 a la especie *Lactobacillus plantarum*, y L134 a la especie *Lactobacillus rhamnosus.* Para todos los casos el %GC al igual que el tamaño del genoma es concordante con los encontrados en la literatura, siendo para *L. fermentum* con un tamaño de genoma de 2.01 Mb y 51.8%GC (Falasconi et al. 2020), *L. plantarum* con un tamaño de genoma de 3.2 Mb y 44.5%GC (Kleerebezem et al. 2003), y L134 a la especie *L. rhamnosus* con un tamaño de genoma de 2.8 Mb y 46.8%GC.

Por otro lado, encontramos que la mayor parte de los grupos externos utilizados se ubicaron correctamente fuera del clado perteneciente a *Lactobacillus* a excepción del género *Pediococcus*, donde las 5 especies quedaron agrupadas en una rama dentro del género *Lactobacillus* cercana a las especies *L. pentosus*, *L. plantarum*, y *L. paraplantarum*. Si bien podemos observar que la rama de *Pediococcus spp.* presenta un valor de confianza relativamente bajo (de valor 38) , se ha descrito que este género en particular presenta una alta similitud con microorganismos *Lactobacillales*, donde análisis filogenéticos de genes de copia única y 16S ubican de la misma forma como un grupo integral no periférico dentro género *Lactobacillus*, explicada por sus similitudes con respecto a su ecología y aspectos metabólicos (Zheng et al. 2015, Franz et al. 2006, Bulgasem et al. 2016, F. Chen et al. 2017).

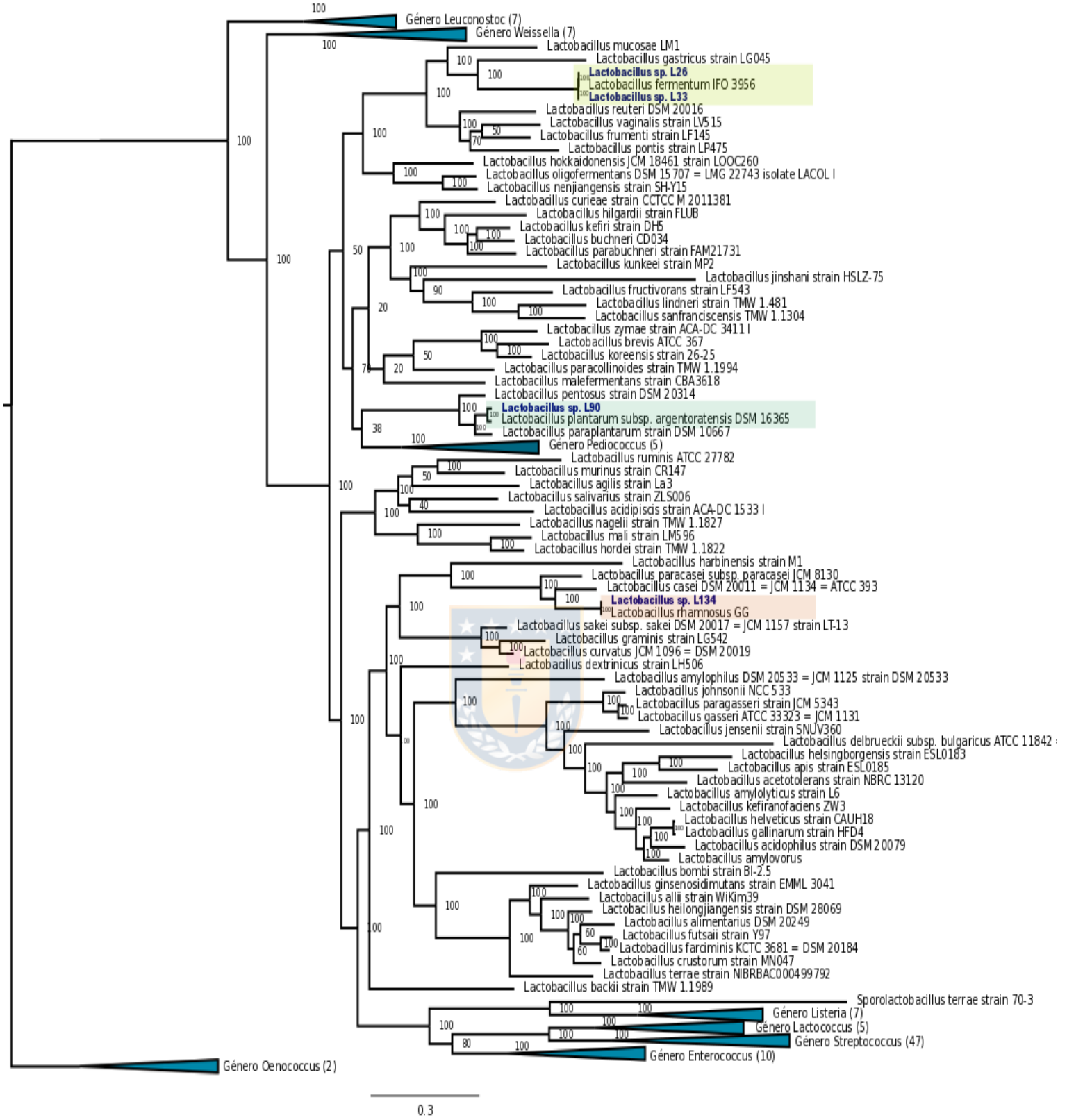


Figura 10: Árbol filogenético de las especies de *Lactobacillus*. Muestra la clasificación filogenética de las 4 cepas de *Lactobacillus* analizadas (*Lactobacillus* sp. L26, *Lactobacillus* sp. L33, *Lactobacillus* sp. L90, *Lactobacillus* sp. L134). Los grupos externos se muestran colapsados (Triángulos celestes), y el valor de confianza de los nodos se encuentra marcada en cada subdivisión.

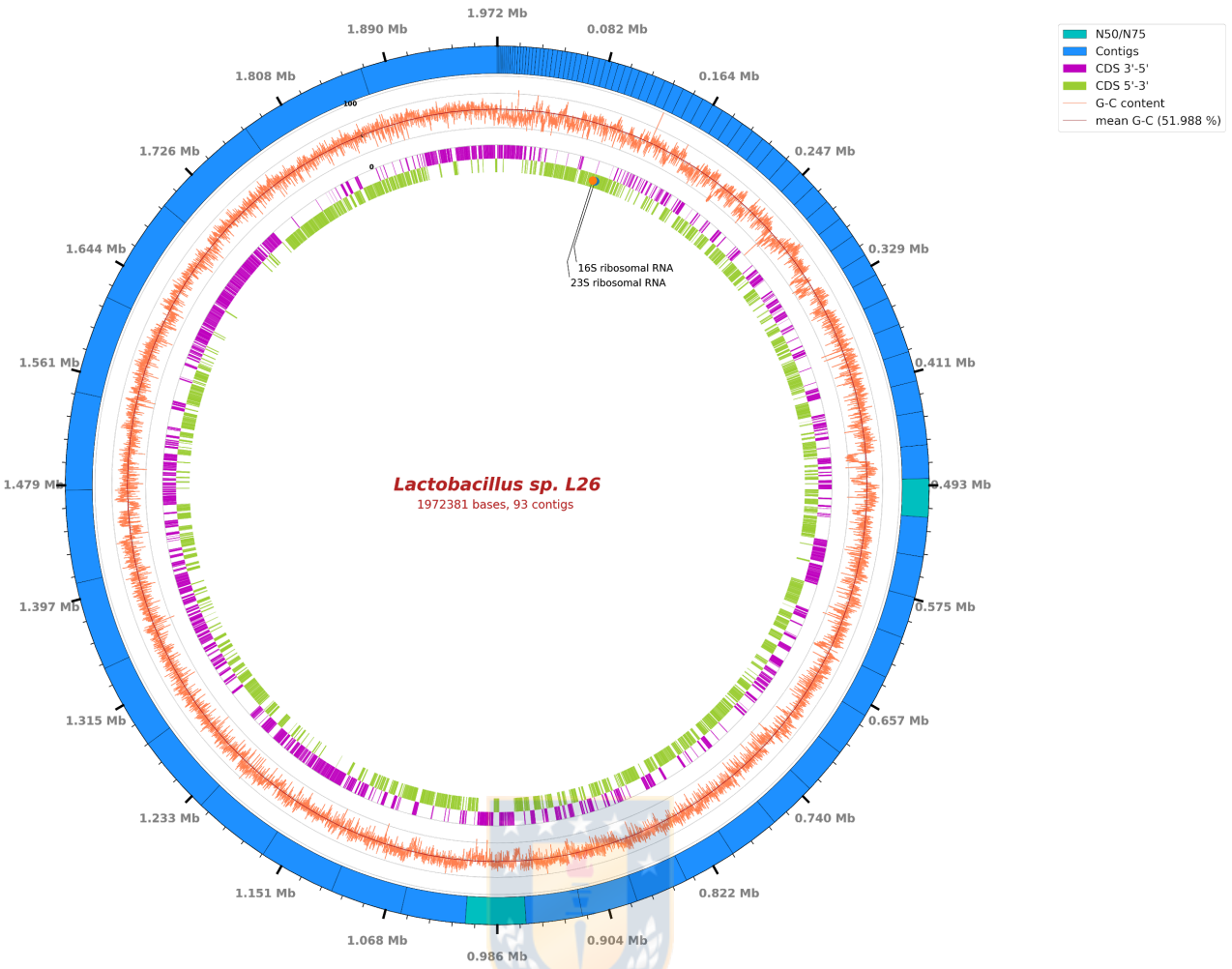


Figura 11: Genoma draft de *Lactobacillus* sp L26. Muestra la proporción de tamaño de los 93 *contigs* (celeste) ordenados de mayor a menor (celeste) en el genoma de 1,972,381pb y un N50 de 44,446 bases (color cian). El anillo concéntrico interior muestra la gráfica escalada (0%-100%) de la proporción de Guanina-Citosina calculada cada 100 pares de bases (naranja) a lo largo de todo el genoma, además de la media de 51.99% G-C (línea roja). En el interior se demarcan las 1,917 secuencias codificantes para las direcciones 5' (magenta) y 3' (verde).

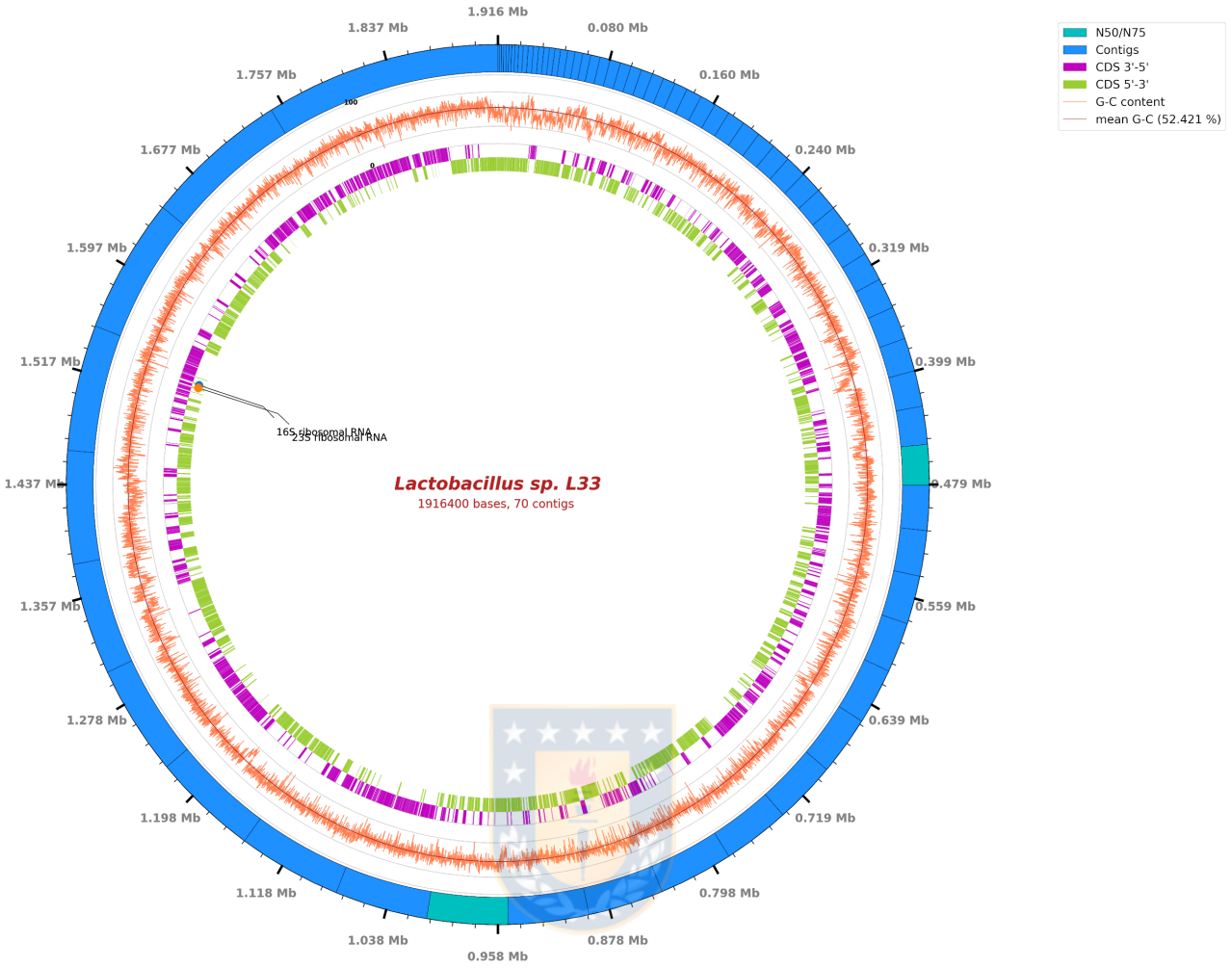


Figura 12: Genoma draft de *Lactobacillus* sp L33. Muestra la proporción de tamaño de los 70 *contigs* (celeste) ordenados de mayor a menor (celeste) en el genoma de 1,916,400 pb y un N50 de 58,118 bases (color cian). El anillo concéntrico interior muestra la gráfica escalada (0%-100%) de la proporción de Guanina-Citosina calculada cada 100 pares de bases (naranja) a lo largo de todo el genoma, además de la media de 52.42%% G-C (línea roja). En el interior se demarcan las 1,867 secuencias codificantes para las direcciones 5' (magenta) y 3' (verde).

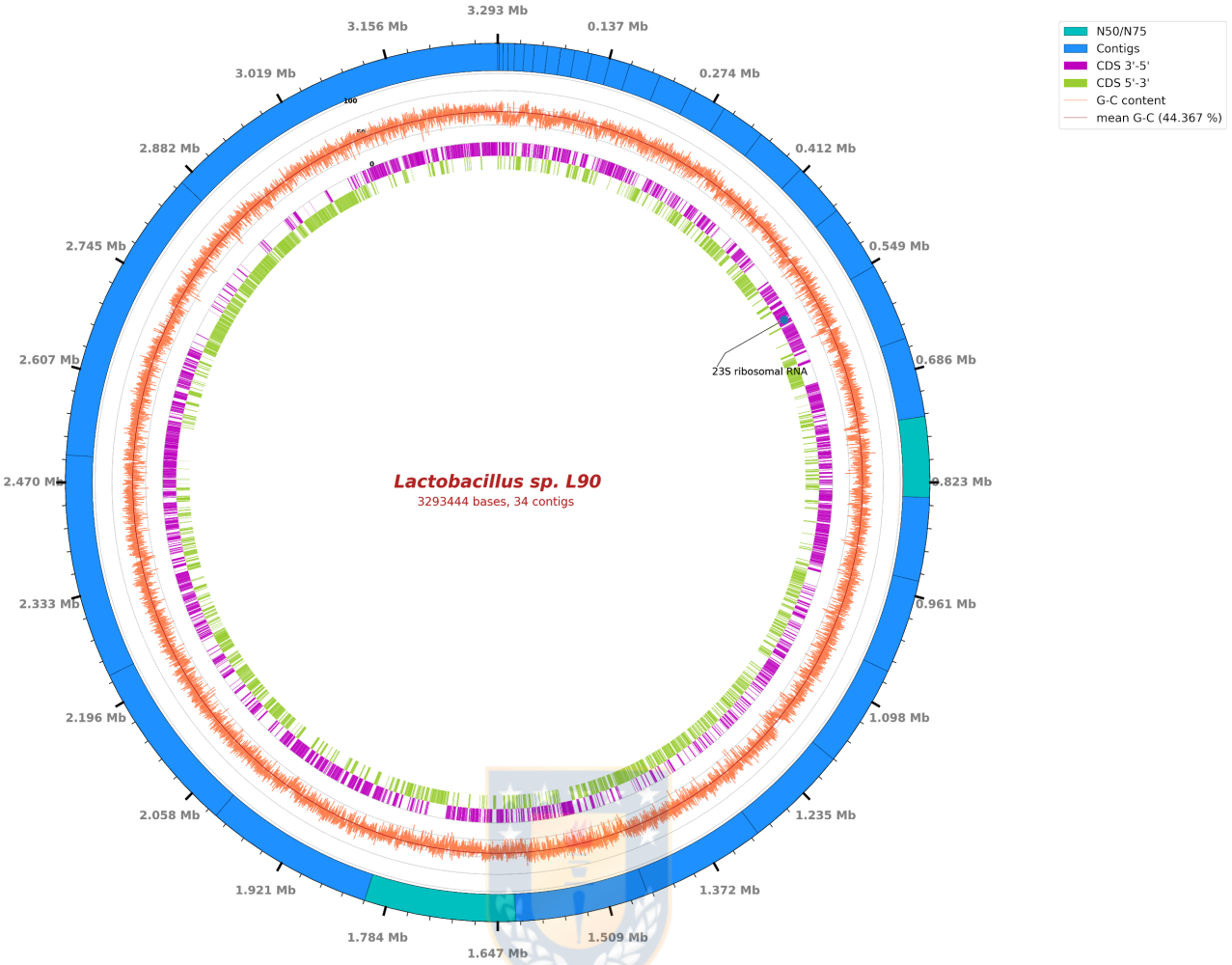


Figura 13: Genoma draft de *Lactobacillus* sp L90. Muestra la proporción de tamaño de los 34 *contigs* (celeste) ordenados de mayor a menor (celeste) en el genoma de 3,293,4444 pb y un N50 de 186,205 bases (color cian). El anillo concéntrico interior muestra la gráfica escalada (0%-100%) de la proporción de Guanina-Citosina calculada cada 100 pares de bases (naranja) a lo largo de todo el genoma, además de la media de 44.37% G-C (línea roja). En el interior se demarcan las 3,092 secuencias codificantes para las direcciones 5' (magenta) y 3' (verde).

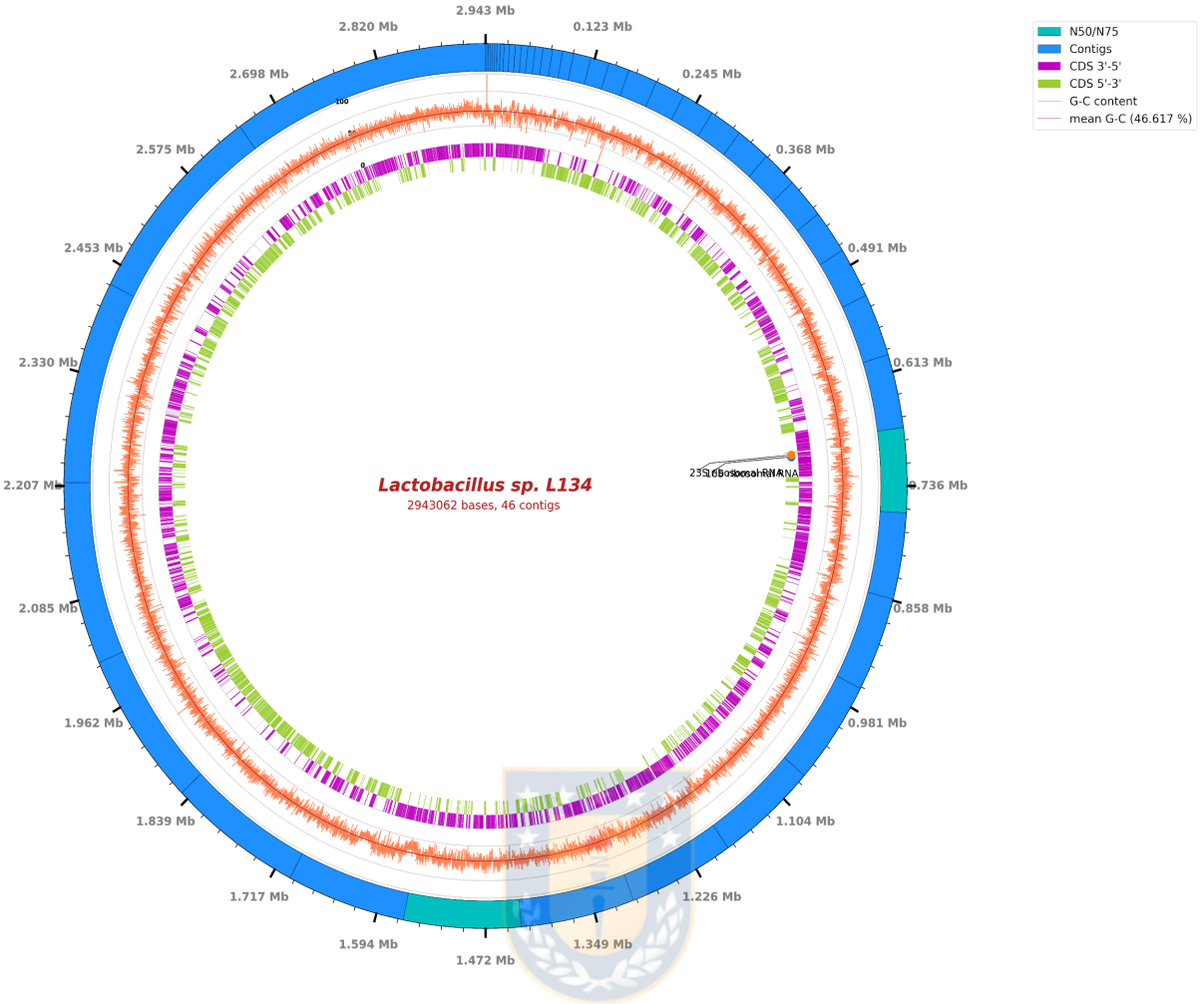


Figura 14: Genoma draft de *Lactobacillus* sp L134. Muestra la proporción de tamaño de los 46 contigs (celeste) ordenados de mayor a menor (celeste) en el genoma de 2,943,062 pb y un N50 de 131,538 bases (color cian). El anillo concéntrico interior muestra la gráfica escalada (0%-100%) de la proporción de Guanina-Citosina calculada cada 100 pares de bases (naranja) a lo largo de todo el genoma, además de la media de 51.99% G-C (línea roja). En el interior se demarcan las 2,768 secuencias codificantes para las direcciones 5' (magenta) y 3' (verde).

5.4. Anotación funcional de características probióticas

Para todas las especies de *Lactobacillus* analizadas se encontraron genes asociados a distintas facultades de organismos probióticos (tablas IV - XI, figuras 15-22) en ambas estrategias de identificación usadas con PBDBsearch (comparaciones de homología por alineamiento con perfiles HMM y alineamientos de secuencia). No obstante, determinamos que la búsqueda por alineamientos con perfiles HMM entregó mejores resultados en cuanto a la cantidad de genes

identificados. Para el caso de *L. fermentum* L26 se encontraron 12 y 26 genes, en *L. fermentum* L33 se encontraron 14 y 28 genes, *L. plantarum* L90 se encontraron 25 y 40, y en *L. rhamnosus* se encontraron 25 y 35 genes, a través de la búsqueda por homología de los alineamientos de secuencia y perfiles HMM, respectivamente.

Cabe destacar que para todos los organismos analizados, gran parte de los genes identificados corresponden a la categoría de resistencia a pH ácido en ambos métodos de identificación. Además, todos los organismos analizados presentaron genes asociados a la degradación de lactosa, producción de vitaminas del complejo B, adhesión y formación de biopelículas y resistencia a estrés oxidativo. Por otro lado, encontramos que sólo dos de las cuatro especies analizadas no poseen genes de resistencias a antibióticos, ambas de la especie *L. fermentum*, y de éstas sólo una no posee genes de producción de aminas biógenas (*L. fermentum* L26). Sin embargo, todas las especies mostraron la presencia de genes asociados a patogenicidad.

Tabla IV: Genes asociados a actividad probiótica de *Lactobacillus fermentum* L26 identificados por alineamiento de secuencias. Se muestran 12 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre del gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
Contig27_22	PBDBCPR004572	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
Contig27_23	PBDBCPR033630	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
Contig27_48	PBDBCPR014302	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Arginine deiminase	arcA
Contig12_30	PBDBCPR003693	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig12_31	PBDBCPR009208	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig12_32	PBDBCPR013857	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig12_33	PBDBCPR038002	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig12_35	PBDBCPR022475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
Contig11_5	PBDBCPR029907	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
scaffold_31_3	PBDBCPR038464	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
scaffold_13_14	PBDBCPR025260	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
scaffold_15_7	PBDBCPR027524	Patogenicidad	Patogenicidad	Enolase	eno

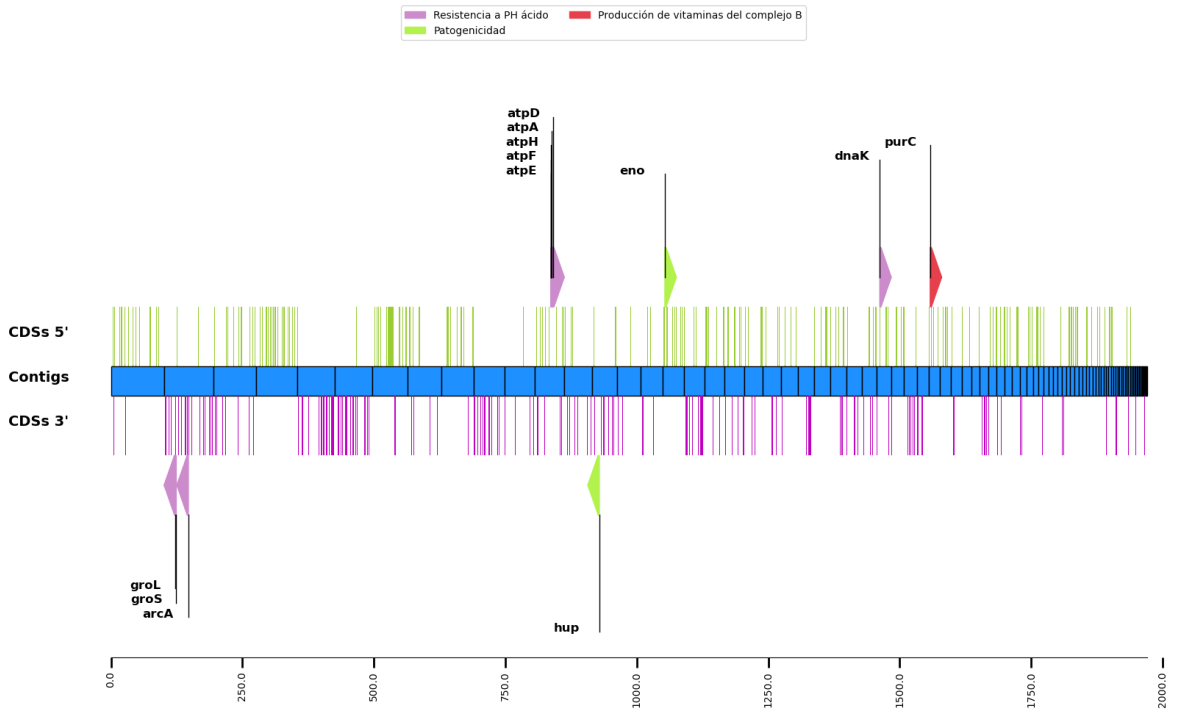


Figura 15: Distribución de genes identificados por alineamiento de secuencias asociados a actividad probiótica en el genoma de *Lactobacillus fermentum* L26. El genoma lineal de 1.9 MB se muestra representado en el centro con 91 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 12 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla V: Genes asociados a actividad probiótica de *Lactobacillus fermentum* L26 identificados por alineamiento de perfiles HMM. Se muestran 26 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
Contig27_22	PBDBCFS0001354	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
Contig27_23	PBDBCFS0002475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
Contig27_48	PBDBCFS0001974	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Arginine deiminase	arcA
Contig12_25	PBDBCFS0005845	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Uracil phosphoribosyltransferase	upp
Contig12_30	PBDBCFS0005192	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig12_31	PBDBCFS0003095	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig12_32	PBDBCFS0006658	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig12_33	PBDBCFS0002153	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig12_35	PBDBCFS0002322	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
scaffold_15_12	PBDBCFS0009088	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	SsrA-binding protein	smpB

Contig11_5	PBDBCFS0000004	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
Contig27_3	PBDBCFS0003000	Competitividad bacteriana	Resistencia a estrés oxidativo	Thioredoxin reductase	trxB
Contig2_3	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
Contig2_5	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
Contig10_7	PBDBCFS0013193	;Interacción con sistema inmune	No clasificados de modulación del sistema inmune	UvrABC system protein B	uvrB
Contig11_22	PBDBCFS0003535	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	6,7-dimethyl-8-ribityllumazine synthase	ribH
scaffold_31_3	PBDBCFS0014281	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
Contig1_48	PBDBCFS0000274	Patogenicidad	Patogenicidad	Ribonuclease Y	rny
scaffold_13_14	PBDBCFS0000636	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
scaffold_15_7	PBDBCFS0000239	Patogenicidad	Patogenicidad	Enolase	eno
scaffold_17_10	PBDBCFS0013114	Patogenicidad	Patogenicidad	ATP-dependent Clp protease	clpL
Contig2_15	PBDBCFS0009849	Competitividad bacteriana	Defensa bacteriana	CRISPR-associated endoribonuclease Cas2	cas2
scaffold_40_3	PBDBCFS0006128	Competitividad bacteriana	Defensa bacteriana	CRISPR system Cms endoribonuclease Csm3	csm3
Contig9_2	PBDBCFS0005866	Competitividad bacteriana	Adhesión y formación de biopelículas	ATP-dependent Clp protease proteolytic subunit 1	clpP1
Contig5_7	PBDBCFS0003860	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM
Contig5_8	PBDBCFS0004730	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase large subunit	lacL

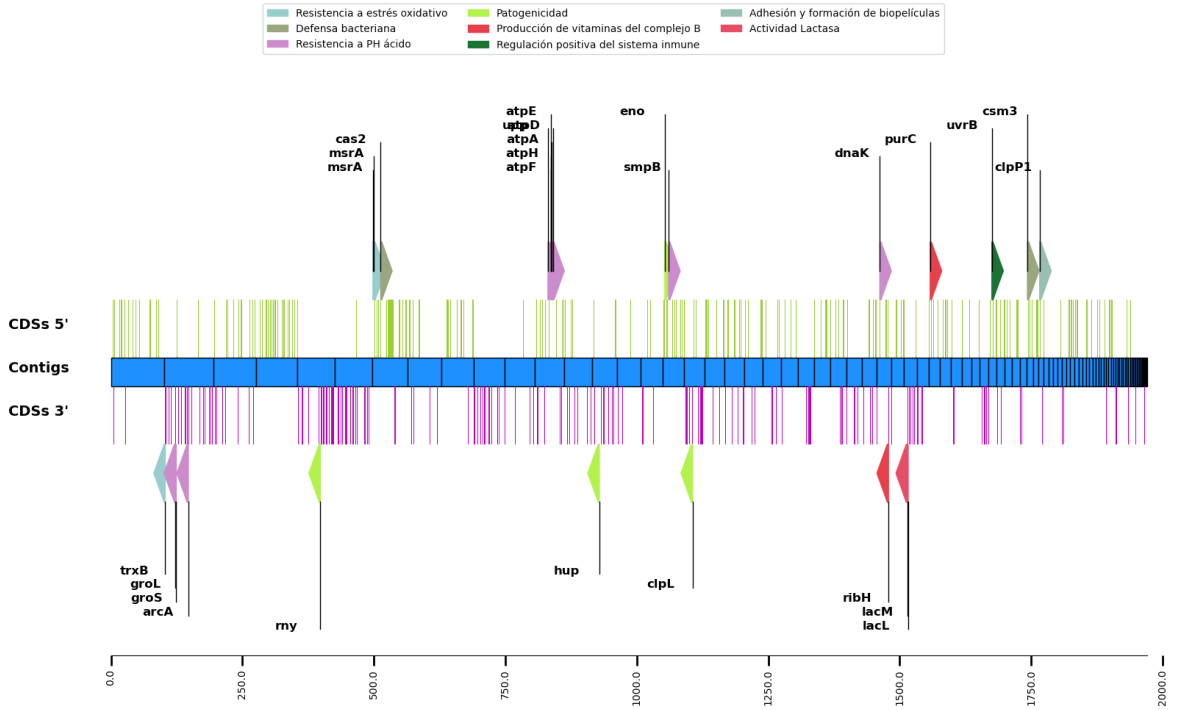


Figura 16: Distribución de genes identificados por alineamiento de perfiles HM asociados a actividad probiótica en el genoma de *Lactobacillus fermentum* L26. El genoma lineal de 1.9 MB se muestra representado en el centro con 91 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 26 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla VI: Genes asociados a actividad probiótica de *Lactobacillus fermentum* L33 identificados por alineamiento de secuencias. Se muestran 14 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
Contig3_69	PBDBCPR003693	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig3_70	PBDBCPR009208	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig3_71	PBDBCPR013857	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig3_72	PBDBCPR038002	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig3_74	PBDBCPR022475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
scaffold_6_25	PBDBCPR014302	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Arginine deiminase	arcA
scaffold_6_50	PBDBCPR033630	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
scaffold_6_51	PBDBCPR004572	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
scaffold_33_4	PBDBCPR029907	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
scaffold_18_17	PBDBCPR038464	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	SAICAR synthetase	purC
scaffold_38_10	PBDBCPR030408	Producción de compuestos tóxicos	Producción de aminas biógenas	Putative agmatine deiminase	aguA
scaffold_38_12	PBDBCPR023258	Producción de compuestos tóxicos	Producción de aminas biógenas	Putrescine carbamoyltransferase	ptcA
scaffold_4_41	PBDBCPR025260	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
scaffold_10_5	PBDBCPR027524	Patogenicidad	Patogenicidad	Enolase	eno

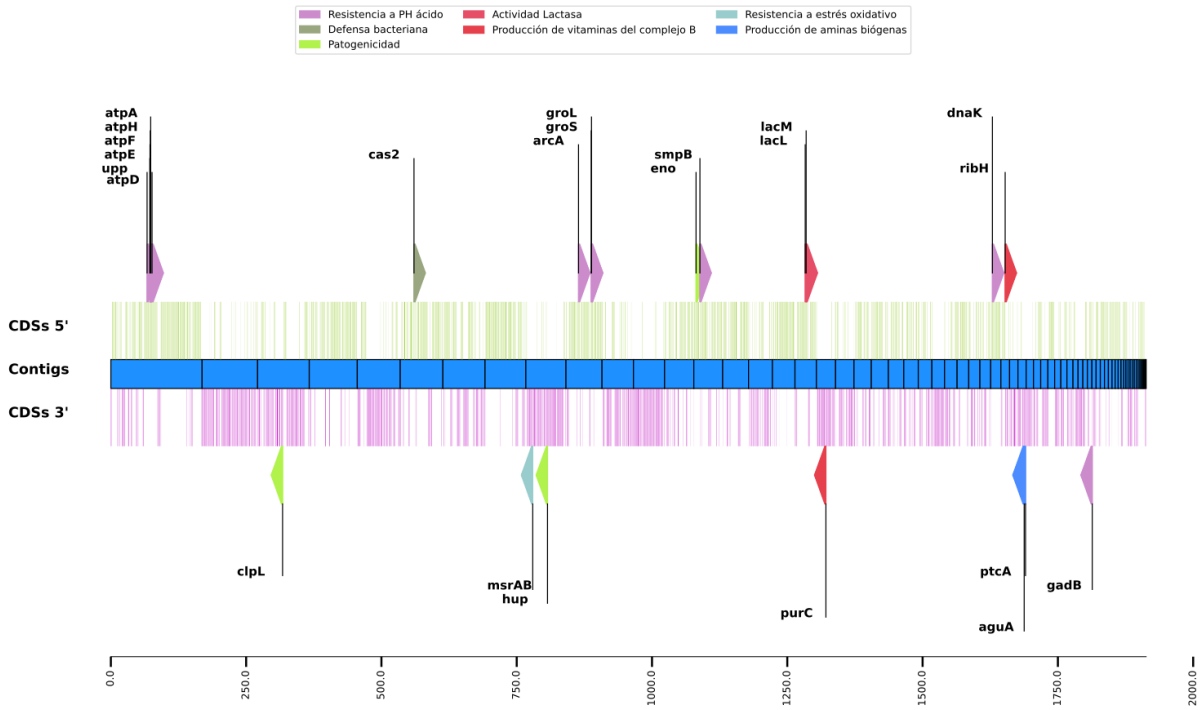


Figura 17: Distribución de genes identificados por alineamiento de secuencias asociados a actividad probiótica en el genoma de *Lactobacillus fermentum* L33. El genoma lineal de 1.91 MB se muestra representado en el centro con 70 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 14 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla VII: Genes asociados a actividad probiótica de *Lactobacillus fermentum* L33 identificados por alineamiento de perfiles HMM. Se muestran 28 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
Contig3_64	PBDBCFS0005845	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Uracil phosphoribosyltransferase	upp
Contig3_69	PBDBCFS0005192	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig3_70	PBDBCFS0003095	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig3_71	PBDBCFS0006658	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig3_72	PBDBCFS0002153	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig3_74	PBDBCFS0002322	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
scaffold_6_25	PBDBCFS0001974	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Arginine deiminase	arcA
scaffold_6_50	PBDBCFS0002475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
scaffold_6_51	PBDBCFS0001354	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
scaffold_10_10	PBDBCFS0009088	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	SsrA-binding protein	smpB
scaffold_33_4	PBDBCFS0000004	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
scaffold_50_2	PBDBCFS0001873	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Glutamate decarboxylase	gadB

scaffold_2_3	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
scaffold_2_5	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
scaffold_30_2	PBDBCFS0003000	Competitividad bacteriana	Resistencia a estrés oxidativo	Thioredoxin reductase	trxB
scaffold_30_11	PBDBCFS0013193	Interacción con sistema inmune	Regulación positiva del sistema inmune; No clasificados de modulación del sistema inmune	UvrABC system protein B	uvrB
scaffold_18_17	PBDBCFS0014281	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
scaffold_34_11	PBDBCFS0003535	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	6,7-dimethyl-8-ribityllumazine synthase	ribH
scaffold_38_10	PBDBCFS0002203	Producción de compuestos tóxicos	Producción de aminas biógenas	Putative agmatine deiminase	aguA1
scaffold_38_12	PBDBCFS0002640	Producción de compuestos tóxicos	Producción de aminas biógenas	Putrescine carbamoyltransferase	ptcA
Contig3_122	PBDBCFS0000274	Patogenicidad	Patogenicidad	Ribonuclease Y	rny
Contig30_36	PBDBCFS0013114	Patogenicidad	Patogenicidad	ATP-dependent Clp protease ATP-binding subunit ClpL	clpL
scaffold_4_41	PBDBCFS0000636	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
scaffold_10_5	PBDBCFS0000239	Patogenicidad	Patogenicidad	Enolase	eno
scaffold_2_20	PBDBCFS0009849	Competitividad bacteriana	Defensa bacteriana	CRISPR-associated endoribonuclease	cas2
scaffold_30_19	PBDBCFS0005866	Competitividad bacteriana	Adhesión y formación de biopelículas	ATP-dependent Clp protease proteolytic subunit 1	clpP1
Contig10_16	PBDBCFS0004730	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase large subunit	lacL
Contig10_17	PBDBCFS0003860	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM

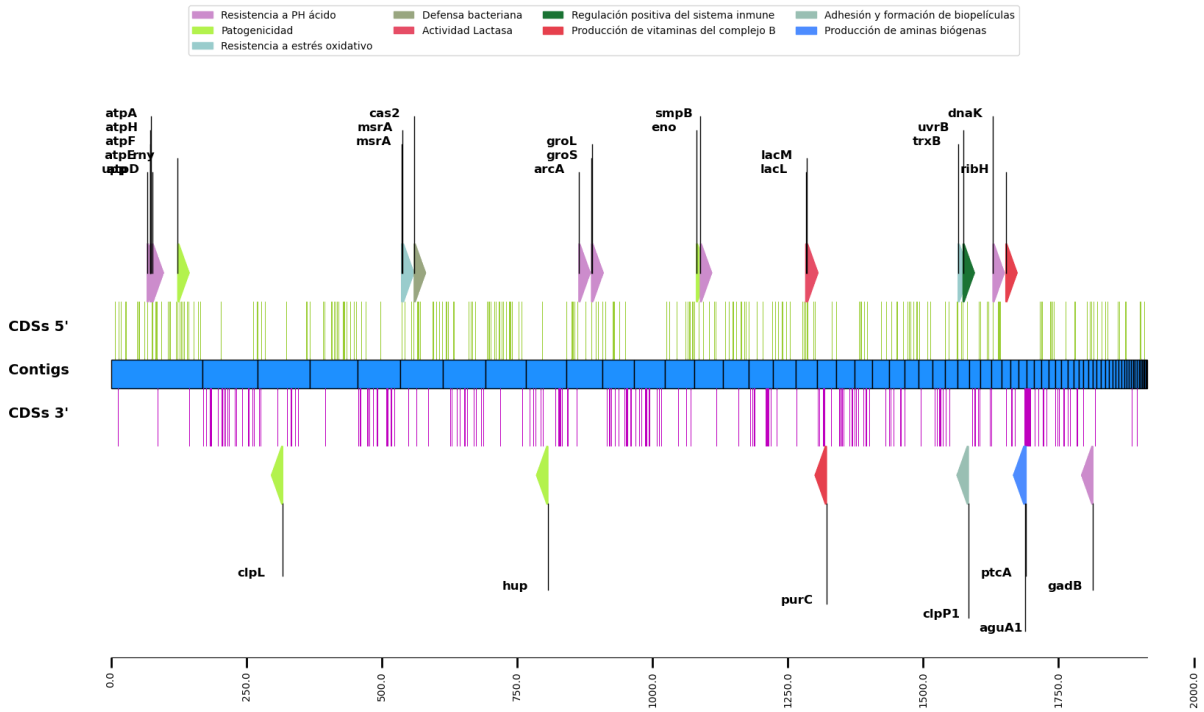


Figura 18: Distribución de genes identificados por alineamiento de perfiles HM asociados a actividad probiótica en el genoma de *Lactobacillus fermentum* L33. El genoma lineal de 1.91 MB se muestra representado en el centro con 70 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 28 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla VIII: Genes asociados a actividad probiótica de *Lactobacillus plantarum* L90 identificados por alineamiento de secuencias. Se muestran 25 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
scaffold_0_266	PBDBCPR029962	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
scaffold_0_267	PBDBCPR029326	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
scaffold_5_234	PBDBCPR002079	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
scaffold_20_16	PBDBCPR018584	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
scaffold_20_17	PBDBCPR010106	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
scaffold_20_18	PBDBCPR018324	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
scaffold_20_19	PBDBCPR038002	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
scaffold_20_21	PBDBCPR022475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
scaffold_2_61	PBDBCPR019303	Resistencia a antibióticos	Resistencia a antibióticos	Glucosamine-6-phosphate deaminase	nagB
Contig10_72	PBDBCPR013305	Resistencia a antibióticos	Resistencia a antibióticos	N-acetylneuraminate lyase	nanA
Contig10_73	PBDBCPR037702	Resistencia a antibióticos	Resistencia a antibióticos	Putative N-acetylmannosamine-6-phosphate	nanE

				2-epimerase	
scaffold_2_80	PBDBCPR010620	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoserine aminotransferase	serC
scaffold_3_112	PBDBCPR034413	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	6,7-dimethyl-8-ribityllumazine synthase	ribH
Contig13_15	PBDBCPR000752	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
Contig10_90	PBDBCPR011408	Competitividad bacteriana	Producción de peróxidos	Pyruvate oxidase	pox5
scaffold_1_176	PBDBCPR013835	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase alpha chain	trpA
scaffold_1_177	PBDBCPR008951	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase beta chain	trpB
scaffold_1_178	PBDBCPR010161	Producción de compuestos tóxicos	Producción de aminas biógenas	N-(5'-phosphoribosyl)anthranilate isomerase	trpF
scaffold_1_179	PBDBCPR038579	Producción de compuestos tóxicos	Producción de aminas biógenas	Indole-3-glycerol phosphate synthase	trpC
scaffold_1_180	PBDBCPR009079	Producción de compuestos tóxicos	Producción de aminas biógenas	Anthranilate phosphoribosyltransferase	trpD
scaffold_0_213	PBDBCPR027524	Patogenicidad	Patogenicidad	Enolase	eno
scaffold_5_343	PBDBCPR025260	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
Contig6_1	PBDBCPR036343	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM
Contig6_2	PBDBCPR015721	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase large subunit	lacL
Contig7_1	PBDBCPR036343	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM

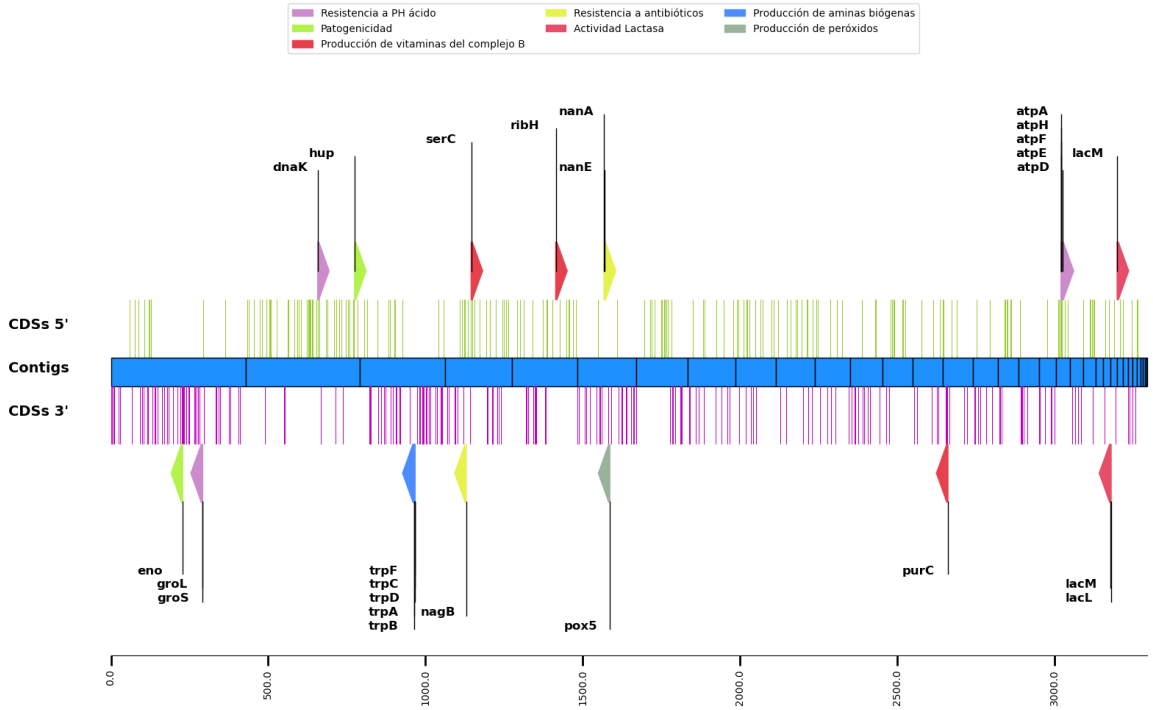


Figura 19: Distribución de genes identificados por alineamiento secuencias asociados a actividad probiótica en el genoma de *Lactobacillus plantarum* L90. El genoma lineal de 3.29 MB se muestra representado en el centro con 34 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 25 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla IX: Genes asociados a actividad probiótica de *Lactobacillus plantarum* L90 identificados por alineamiento de perfiles HM. Se muestran 40 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
scaffold_0_207	PBDBCFS0009088	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	SsrA-binding protein	smpB
scaffold_0_266	PBDBCFS0001354	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
scaffold_0_267	PBDBCFS0005382	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
scaffold_5_234	PBDBCFS0000004	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
scaffold_9_37	PBDBCFS0001873	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Glutamate decarboxylase	gadB
scaffold_20_13	PBDBCFS0005845	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Uracil phosphoribosyltransferase	upp
scaffold_20_16	PBDBCFS0005218	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
scaffold_20_17	PBDBCFS0005746	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
scaffold_20_18	PBDBCFS0006654	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
scaffold_20_19	PBDBCFS0002153	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA

scaffold_20_21	PBDBCFS0002322	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
Contig10_79	PBDBCFS0010183	Competitividad bacteriana; Resistencia a antibióticos	Resistencia a estrés oxidativo; Resistencia a antibióticos	Catalase	katA
scaffold_0_239	PBDBCFS0003000	Competitividad bacteriana	Resistencia a estrés oxidativo	Thioredoxin reductase	trxB
scaffold_5_256	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
scaffold_1_23	PBDBCFS0010587	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrAB
scaffold_1_24	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
scaffold_2_61	PBDBCFS0003105	Resistencia a antibióticos	Resistencia a antibióticos	Glucosamine-6-phosphate deaminase	nagB
Contig10_72	PBDBCFS0013832	Resistencia a antibióticos	Resistencia a antibióticos	N-acetylneuraminate lyase	nanA
Contig10_73	PBDBCFS0004319	Resistencia a antibióticos	Resistencia a antibióticos	Putative N-acetylmannosamine-6-phosphate 2-epimerase	nanE
scaffold_0_230	PBDBCFS0013193	Interacción con sistema inmune; Interacción con sistema inmune	Regulación positiva del sistema inmune; No clasificados de modulación del sistema inmune	UvrABC system protein B	uvrB
scaffold_2_80	PBDBCFS0015141	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoserine aminotransferase	serC
scaffold_3_112	PBDBCFS0009143	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	6,7-dimethyl-8-ribityllumazine synthase	ribH
Contig13_15	PBDBCFS0014804	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
Contig10_90	PBDBCFS0012827	Competitividad bacteriana	Producción de peróxidos	Pyruvate oxidase	pox5
scaffold_1_176	PBDBCFS0013634	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase alpha chain	trpA
scaffold_1_177	PBDBCFS0001872	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase beta chain	trpB
scaffold_1_178	PBDBCFS0006035	Producción de compuestos tóxicos	Producción de aminas biógenas	N-(5'-phosphoribosyl)anthranilate isomerase	trpF
scaffold_1_179	PBDBCFS0013535	Producción de compuestos tóxicos	Producción de aminas biógenas	Indole-3-glycerol phosphate synthase	trpC
scaffold_1_180	PBDBCFS0016360	Producción de compuestos tóxicos	Producción de aminas biógenas	Anthranilate phosphoribosyltransferase	trpD
scaffold_0_37	PBDBCFS0013011	Patogenicidad	Patogenicidad	ATP-dependent Clp protease	clpC
scaffold_0_213	PBDBCFS0000239	Patogenicidad	Patogenicidad	Enolase	eno
scaffold_0_214	PBDBCFS0014563	Patogenicidad	Patogenicidad	Triosephosphate isomerase	tpiA
scaffold_5_343	PBDBCFS0000636	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
Contig10_84	PBDBCFS0006741	Patogenicidad	Patogenicidad	Putative AgrB-like protein	
Contig10_85	PBDBCFS0013114	Patogenicidad	Patogenicidad	ATP-dependent Clp protease	clpL
scaffold_12_58	PBDBCFS0000274	Patogenicidad	Patogenicidad	Ribonuclease Y	rny
scaffold_0_219	PBDBCFS0005866	Competitividad bacteriana	Adhesión y formación de biopelículas	ATP-dependent Clp protease proteolytic subunit 1	clpP1
Contig6_1	PBDBCFS0015903	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM
Contig6_2	PBDBCFS0012508	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase large subunit	lacL
Contig7_1	PBDBCFS0015903	Producción de compuestos bioactivos	Actividad Lactasa	Beta-galactosidase small subunit	lacM

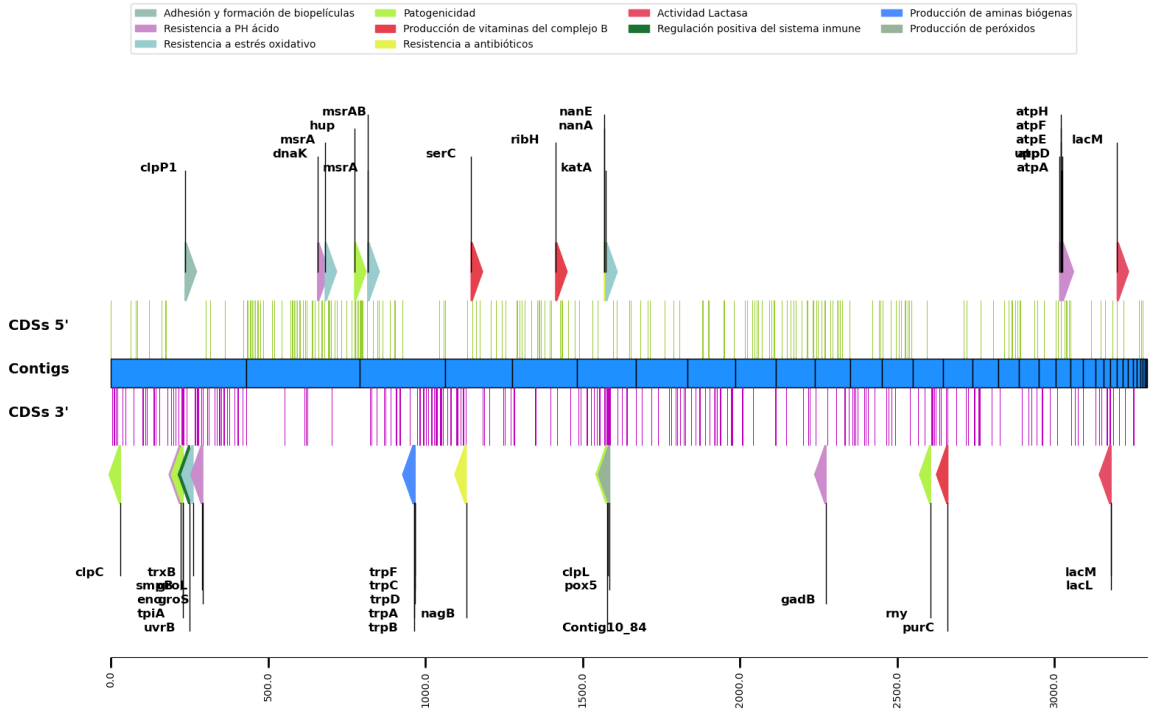


Figura 20: Distribución de genes identificados por alineamiento de perfiles HMM asociados a actividad probiótica en el genoma de *Lactobacillus plantarum* L90. El genoma lineal de 3.29 MB se muestra representado en el centro con 34 *contigs* (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 40 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla X: Genes asociados a actividad probiótica de *Lactobacillus Rhamnosus* L134 identificados por alineamiento de secuencias. Se muestran 20 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
umpscaff_1_126	PBDBCPR014314	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
Contig3_153	PBDBCPR035918	Producción de compuestos bioactivos	Actividad Lactasa	6-phospho-beta-galactosidase	lacG
Contig3_154	PBDBCPR035918	Producción de compuestos bioactivos	Actividad Lactasa	6-phospho-beta-galactosidase	lacG
Contig10_61	PBDBCPR022475	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
Contig10_64	PBDBCPR038002	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig10_65	PBDBCPR001256	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig10_66	PBDBCPR031855	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig10_67	PBDBCPR024337	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig10_69	PBDBCPR005668	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Uracil phosphoribosyltransferase	upp

Contig25_67	PBDBCPR036185	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
Contig5_74	PBDBCPR018119	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase alpha chain	trpA
Contig5_75	PBDBCPR005853	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase beta chain	trpB
Contig5_76	PBDBCPR020146	Producción de compuestos tóxicos	Producción de aminas biógenas	N-(5'-phosphoribosyl)anthranilate isomerase	trpF
Contig5_77	PBDBCPR027851	Producción de compuestos tóxicos	Producción de aminas biógenas	Indole-3-glycerol phosphate synthase	trpC
Contig5_78	PBDBCPR018311	Producción de compuestos tóxicos	Producción de aminas biógenas	Anthranilate phosphoribosyltransferase	trpD
Contig8_26	PBDBCPR021240	Competitividad bacteriana	Defensa bacteriana	CRISPR-associated endoribonuclease Cas2	cas2
Contig8_60	PBDBCPR017214	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
Contig8_61	PBDBCPR013346	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS
umpscaff_18_53	PBDBCPR021551	Resistencia a antibióticos	Resistencia a antibióticos	Glucosamine-6-phosphate deaminase	nagB
Contig17_19	PBDBCPR027708	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK

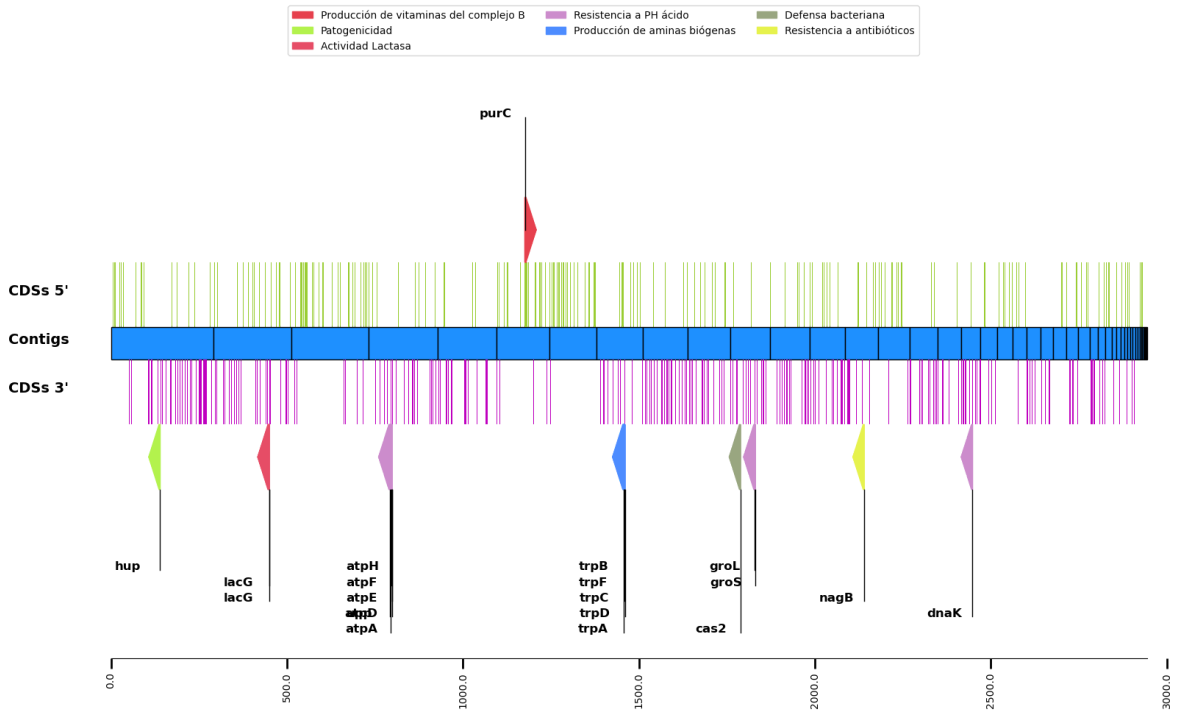


Figura 21: Distribución de genes identificados por alineamiento secuencias asociados a actividad probiótica en el genoma de *Lactobacillus rhamnosus* L134. El genoma lineal de 2.94 MB se muestra representado en el centro con 46 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 25 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla XI: Genes asociados a actividad probiótica de *Lactobacillus rhamnosus* L26 identificados por alineamiento de perfiles HMM. Se muestran 21 genes identificados por alineamiento de secuencias junto con la categoría a la que pertenecen, subcategoría, proteína y nombre de gen.

Secuencia	Mejor hit	Categoría	Subcategoría	Proteína	Gene
Contig10_60	PBDBCFS0008776	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase epsilon chain	atpC
Contig10_61	PBDBCFS0002322	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit beta	atpD
Contig10_63	PBDBCFS0016027	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase gamma chain	atpG
Contig10_64	PBDBCFS0002153	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit alpha	atpA
Contig10_65	PBDBCFS0004147	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit delta	atpH
Contig10_66	PBDBCFS0005020	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit b	atpF
Contig10_67	PBDBCFS0005194	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	ATP synthase subunit c	atpE
Contig10_69	PBDBCFS0005845	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Uracil phosphoribosyltransferase	upp
Contig8_60	PBDBCFS0001354	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	60 kDa chaperonin	groL
Contig8_61	PBDBCFS0001887	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	10 kDa chaperonin	groS

Contig9_6	PBDBCFS0009088	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	SsrA-binding protein	smpB
Contig17_19	PBDBCFS0000004	Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Chaperone protein DnaK	dnaK
umpscaff_1_114	PBDBCFS0005706	Competitividad bacteriana	Resistencia a estrés oxidativo	Peptide methionine sulfoxide reductase	msrA
Contig9_28	PBDBCFS0003000	Competitividad bacteriana	Resistencia a estrés oxidativo	Thioredoxin reductase	trxB
umpscaff_18_53	PBDBCFS0014178	Producción de compuestos bioactivos	Degradación de lípidos y ácidos grasos	Glucosamine-6-phosphate deaminase	nagB
Contig9_25	PBDBCFS0013193	Interacción con sistema inmune;Interacción con sistema inmune	No clasificados de modulación del sistema inmune	UvrABC system protein B	uvrB
Contig25_67	PBDBCFS0003472	Producción de compuestos bioactivos	Producción de vitaminas del complejo B	Phosphoribosylaminoimidazole-succinocarboxamide synthase	purC
Contig5_74	PBDBCFS0013287	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase alpha chain	trpA
Contig5_75	PBDBCFS0001872	Producción de compuestos tóxicos	Producción de aminas biógenas	Tryptophan synthase beta chain	trpB
Contig5_76	PBDBCFS0002227	Producción de compuestos tóxicos	Producción de aminas biógenas	N-(5'-phosphoribosyl)anthranilate isomerase	trpF
Contig5_77	PBDBCFS0002224	Producción de compuestos tóxicos	Producción de aminas biógenas	Indole-3-glycerol phosphate synthase	trpC
Contig5_78	PBDBCFS0002343	Producción de compuestos tóxicos	Producción de aminas biógenas	Anthranilate phosphoribosyltransferase	trpD
umpscaff_1_126	PBDBCFS0000636	Patogenicidad	Patogenicidad	DNA-binding protein HU	hup
Contig9_13	PBDBCFS0000239	Patogenicidad	Patogenicidad	Enolase	eno
Contig9_14	PBDBCFS0014563	Patogenicidad	Patogenicidad	Triosephosphate isomerase	tpiA
Contig9_55	PBDBCFS0000274	Patogenicidad	Patogenicidad	Ribonuclease Y	rny
Contig19_2	PBDBCFS0013011	Patogenicidad	Patogenicidad	ATP-dependent Clp protease ATP-binding subunit ClpC	clpC
umpscaff_40_6	PBDBCFS0013114	Patogenicidad	Patogenicidad	ATP-dependent Clp protease ATP-binding subunit ClpL	clpL
Contig8_26	PBDBCFS0009828	Competitividad bacteriana	Defensa bacteriana	CRISPR-associated endoribonuclease Cas2	cas2
Contig9_18	PBDBCFS0005866	Competitividad bacteriana	Adhesión y formación de biopelículas	ATP-dependent Clp protease proteolytic subunit 1	clpP1
Contig3_139	PBDBCFS0000027	Producción de compuestos bioactivos	Actividad Lactasa	Galactose-6-phosphate isomerase subunit LacB	lacB1
Contig3_140	PBDBCFS0016552	Producción de compuestos bioactivos	Actividad Lactasa	Tagatose 1,6-diphosphate aldolase 2	lacD2
Contig3_153	PBDBCFS0002317	Producción de compuestos bioactivos	Actividad Lactasa	6-phospho-beta-galactosidase	lacG
Contig3_154	PBDBCFS0002317	Producción de compuestos bioactivos	Actividad Lactasa	6-phospho-beta-galactosidase	lacG
Contig4_64	PBDBCFS0016552	Producción de compuestos bioactivos	Actividad Lactasa	Tagatose 1,6-diphosphate aldolase 2	lacD2

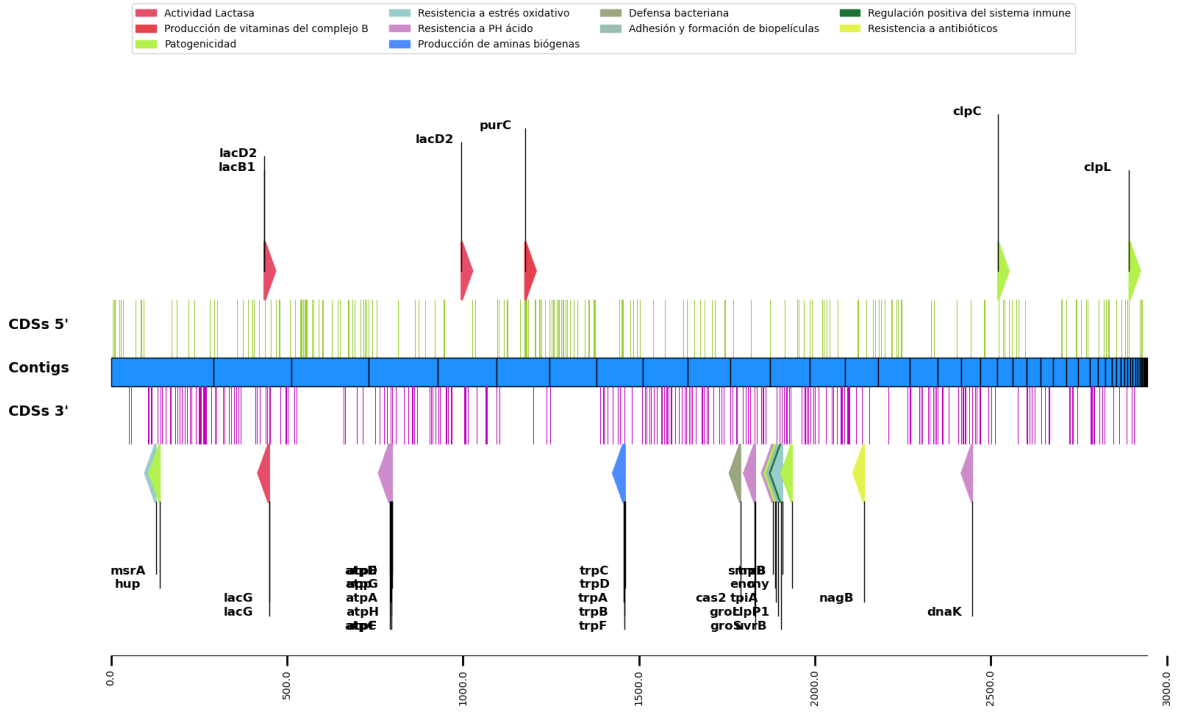


Figura 22: Distribución de genes identificados por alineamiento de perfiles HMM asociados a actividad probiótica en el genoma de *Lactobacillus plantarum* L90. El genoma lineal de 2.94 MB se muestra representado en el centro con 46 contigs (barras central celestes). Las regiones codificantes 5' (verde) y 3' (magenta) están demarcadas sobre y bajo los segmentos del genoma respectivamente. Los 35 genes se ubican marcados con triángulos (colores en la legenda) posicionados y orientados según la dirección en que se encuentran en el genoma.

Tabla XII: Número de genes identificados por categoría en las 4 cepas de *Lactobacillus* en la identificación por alineamiento de secuencias.

Categoría	Subcategoría	*Lactobacillus fermentum L26*	*Lactobacillus fermentum L33*	*Lactobacillus plantarum L90*	*Lactobacillus rhamnosus L134*
Asociados a receptores Toll-like	-	0	0	0	0
Competitividad bacteriana	Producción de bacteriocinas	0	0	0	0
	Sistemas toxina-antitoxina	0	0	0	0
	Defensa bacteriana	0	0	0	1
	Producción de peróxidos	0	0	1	0
	Adhesión y formación de biopelículas	0	0	0	0
	Resistencia a estrés oxidativo	0	0	0	0
No clasificados de modulación del sistema inmune	-	0	0	0	0
Patogenicidad	-	2	2	2	1
Producción de compuestos bioactivos	Producción vitaminas del complejo B	1	1	3	1
	Actividad lactasa	0	0	3	2
	Degradación de lípidos y ácidos grasos	0	0	0	1
Producción de compuestos tóxicos	Producción de aminas biógenas	0	2	5	5
Regulación negativa del sistema inmune	-	0	0	0	0
Regulación positiva del sistema inmune	-	0	0	0	0
Resistencia a antibióticos	-	0	0	3	0
Sobrevivencia al tracto intestinal	Resistencia a pH ácido	9	9	8	9
	Resistencia a sales biliares	0	0	0	0
	Actividad ureasa	0	0	0	0

Tabla XIII: Número de genes identificados por categoría en las 4 cepas de *Lactobacillus* en la identificación por alineamiento de perfiles HMM.

Categoría	Subcategoría	*Lactobacillus fermentum L26*	*Lactobacillus fermentum L33*	*Lactobacillus plantarum L90*	*Lactobacillus rhamnosus L134*
Asociados a receptores Toll-like	-	0	0	0	0
Competitividad bacteriana	Producción de bacteriocinas	0	0	0	0
	Sistemas toxina-antitoxina	0	0	0	0
	Defensa bacteriana	2	1	0	1
	Producción de peróxidos	0	0	1	0
	Adhesión y formación de biopelículas	1	1	1	1
	Resistencia a estrés oxidativo	3	3	5	2
No clasificados de modulación del sistema inmune	-	1	1	1	1
Patogenicidad	-	4	4	7	6
Producción de compuestos bioactivos	Producción vitaminas del complejo B	2	2	3	1
	Actividad Lactasa	2	2	3	5
	Degradación de lípidos y ácidos grasos	0	0	0	0
Producción de compuestos tóxicos	Producción de aminas biógenas	0	2	5	5
Regulación negativa del sistema inmune	-	0	0	0	0
Regulación positiva del sistema inmune	-	0	0	0	0
Resistencia a antibióticos	-	0	0	4	1
Sobrevivencia al tracto intestinal	Resistencia a pH ácido	11	12	11	12
	Resistencia a sales biliares	0	0	0	0
	Actividad ureasa	0	0	0	0

6 DISCUSIÓN

Actualmente, muchas de las metodologías de anotación estándar se basan en la utilización de comparaciones de similitud a través de alineamientos de secuencia (directamente con secuencias o perfiles HMM) utilizando en su subrutina programas como BLAST o HMMER, donde se aplican filtros en parámetros de error, identidad y cobertura para la discriminación de la correcta anotación. Dentro del gran número de las aplicaciones y servidores construidos para realizar la anotación funcional del material genético destacan RAST, PFAM, BLAST2GO, PROKKA, que ofrecen una herramienta esencial para llevar las secuencias a un contexto biológico y obtener una visión general de los sistemas y características funcionales de un organismo. No obstante, sus objetivos son de anotación global o de metabolitos secundarios, lo que no les permite profundizar en la discriminación de características más específicas y la selección y búsqueda de estas se deja a criterio del investigador teniendo que recurrir a otros métodos más específicos para la clasificación de las secuencias. Si bien, se han desarrollado sistemas de anotación automatizados más específicos donde se pueden discriminar algunas características asociadas a facultades probióticas como BATIBASE (para la anotación de bacteriocinas utilizando alineamientos de secuencias) (Hammami, R. et al. 2010), GenDB especializada en patógenos (Meyer, F. et al. 2003), o ARMfinderPlus (para la búsqueda de genes de resistencia a antibióticos) (Feldgarden et al. 2019), ninguno de estos servidores ofrece un sistema unificado de anotación basado en características funcionales probióticas. El sistema desarrollado en esta investigación, donde se han propuesto y ordenado categorías y subcategorías para clasificación de secuencias biológicas en términos de facultades probióticas, ofrece una metodología de anotación similar a ARMfinderPlus, donde las secuencias son anotadas utilizando alineamientos de secuencias contra perfiles HMM. Sin embargo, aparte de la especificidad orientada a microorganismos probióticos de PBDBsearch, otra diferencia entre esta metodología y otros sistemas como ARMfinder o PFAM, es que en este caso construimos una base de datos de perfiles HMM para proteínas completas, y no solo los dominios conservados para la clasificación de las secuencias. Esto nos permite asociar de forma más específica las características funcionales de una secuencia completa y no solo una parte de ella, entregando de

manera integrada la información biológica, funcional y bibliográfica con una mayor precisión que los sistemas de anotación que utilizan alineamientos de secuencias con programas como BLAST.

En cuanto al análisis de microorganismos del género *Lactobacillus*, la búsqueda de genes asociados a las distintas facultades probióticas fue exitosa, encontrando genes asociados a las distintas categorías propuestas en esta investigación. En todos los organismos analizados encontramos genes de resistencia a pH ácido documentados previamente, como los del sistema de bomba de protones F0F1 ATPasa (atpA, atpD, atpE, atpF, atpH) y genes de reparación y protección del ADN (grol, gros, DnaK), además de genes de resistencia a estrés oxidativo como msrA y trxB, que permiten a estos microorganismos sobrevivir dentro del adverso ambiente del tracto intestinal (Cotter y Hill 2003). De la misma forma encontramos genes asociados a beneficios a la salud del huésped en los 4 microorganismos analizados, pertenecientes a las categoría de producción de compuestos bioactivos, específicamente en las subcategoría de producción de vitaminas del complejo B como purC involucrado en la biosíntesis de cobalamina, y ribH involucrada en la biosíntesis de riboflavina (LeBlanc et al. 2011). Además se detectaron genes asociados a la degradación de lactosa como lacM y lacL codificantes para la enzima beta-galactosidasa (Nakayama y Amachi 2002) encontrados en las especies *L. fermentum* y *L. plantarum*, y lacB1, lacD2, lacG encontrados en *L. rhamnosus*.
Por otro lado, se encontró el gen uvrB asociado con la respuesta del sistema inmune del huésped, presente en todas las especies. Este gen codifica para una subunidad del complejo UvrABC que juega un papel fundamental en la reparación de daño del ADN, y estudios en *Mycobacterium tuberculosis* sugieren que mutantes uvrB son susceptibles a óxido nítrico sintasa inducible y otras defensas inmunitarias (Darwin y Nathan 2005).
Se ha descrito que microorganismos probióticos del género *Lactobacillus* son capaces de modular las respuestas inmunes en el huésped, otorgando un estado de alerta frente al ataque de otros microorganismos patógenos (Yan y Polk 2011). No obstante el rol que desempeñan microorganismos probióticos, así como los mecanismos epigenéticos que regulan su interacción con el sistema inmune no han sido descifrados completamente (Llewellyn y Foey 2017).

Curiosamente, encontramos genes asociados a patogenicidad relacionados a estrategias de evasión y virulencia, como *hup* (gen que codifica para la proteína de unión a DNA asociada con la coordinación de patogenicidad (Phan et al. 2015),escrito en la inmunoestimulación del huésped (Vastano et al. 2016), *rny* que afecta la expresión de factores de virulencia a través de *agr* o de forma independiente (Jester, Romby, y Lioliou 2012), clpC y clpL, chaperonas reguladoras de la proteólisis involucradas en la invasión y adhesión celular (Nair, Milohanic, y Berche 2000). La presencia de estos genes en todas las especies analizadas, podría explicar en parte como estos microorganismos son capaces de evadir las respuestas defensivas del huésped, activando el sistema inmune de manera amigable y ser capaces de sobrevivir estableciendo una relación simbionte con su hospedador. No obstante, a pesar de que la presencia de estos genes ha sido reportada en algunas especies probióticas(Zuo, Chen, y Marcotte 2020; Lebeer, Vanderleyden, y De Keersmaecker 2008), no existe información sobre la influencia de los productos genéticos y su relación con los mecanismos de defensa en el huésped.

De acuerdo con nuestros resultados estimamos que el candidato más inocuo de las 4 especies analizadas para ser considerado probiótico es *Lactobacillus fermentum* L26, ya que fue el único donde no se encontró la presencia de genes asociados con la producción de aminas biógenas, no obstante la producción de aminas biógenas por parte de microorganismo probióticos ha sido documentada y se considera que dependiendo de cantidades en que estas sean producidas puede llegar a afectar de forma negativa la salud del huésped (Roselino et al. 2020). Por otro lado, consideramos que los candidato con mayores facultades probióticas fueron *L. plantarum* L90 y *L. rhamnosus* L134, ya que poseen una batería de genes más variada que los del género *L. fermentum*, no obstante *L. plantarum* L90 posee 5 genes de resistencia a antibióticos por lo que clasificamos a *L. rhamnosus* L134 como el candidato con mayores facultades probióticas. Es importante destacar que este tipo de sistemas de anotación basados en métodos *in silico* requieren de validación experimental que sustente los resultados teóricos encontrados, considerando los mecanismos reguladores que actúan para un determinado fenotipo. No obstante, ofrece una visión preliminar para la discriminación de las

facultades probióticas que se buscan en un microorganismo, permitiendo la selección rápida y eficaz previo a estudios basados en métodos experimentales empíricos. Trabajo en progreso está siendo desarrollado para validar, mejorar y ajustar este sistema de anotación mediante: 1) La comparación de las anotaciones con un grupo de bacterias no probióticas y probióticas de los géneros *Lactobacillus sp.* y *Bifidobacterium sp*., 2) El análisis de las variaciones en las secuencias de los genes descritos, 3) La evaluación de la formación de clúster de genes con anotaciones probióticas en los genomas de los microorganismos, 4) El estudio de las interacciones sistémicas de estos genes, y 5) La implementación de los servicios en una aplicación web.

7 CONCLUSIÓN

En esta investigación se ha logrado desarrollar un sistema de anotación funcional de genes específico para la clasificación y selección de microorganismos probióticos basado en categorías de clasificación incluyente y excluyentes para microorganismos probióticos llamado PBDBsearch. Este sistema de clasificación está sustentado en dos elementos esenciales, correspondientes al sistema de anotación de comparaciones por homología basadas en alineamiento (de secuencia y de perfiles HMM), y una base de datos específica construida a partir de genes asociados características funcionales de microorganismos probióticos, seleccionados por el análisis de términos de ontología génica. Además, determinamos a través del análisis del agrupamiento (variando parámetros de identidad y cobertura) de 561,911 secuencias revisadas manualmente (Uniprot-SwissProt) que los ´valores de corte de los parámetros de identidad cobertura y puntaje normalizado de alineamiento para la selección de la anotación funcional de genes y la construcción de la base de datos son de 70%, 90%, y 1.5, respectivamente, que fueron aplicados en nuestra metodología. Esto nos permitió construir una base de datos de perfiles HMM para el análisis de secuencias de proteínas completas a partir de las secuencias de proteínas seleccionadas por ontología de genes, asociadas a características funcionales probióticas, obteniendo una base de datos robusta y precisa para el análisis y anotación de secuencias codificantes, la cual llamamos base de datos PROBIODB, que se compone de 38,684 secuencias y 13,031 perfiles HMM, clasificados en 10 categorías (resistencia a antibióticos, patogenicidad, producción de compuestos tóxicos, sobrevivencia al tracto intestinal, producción de compuestos bioactivos, competitividad bacteriana, regulación positiva del sistema inmune, regulación negativa del sistema inmune, asociados a receptores *toll-like*, y no clasificados de modulación del sistema inmune) y 19 subcategorías funcionales (resistencia a antibióticos, patogénesis y enfermedad, producción de aminas biógenas, resistencia a pH ácido, resistencia a sales biliares, actividad ureasa, producción vitaminas del complejo B, actividad lactasa, degradación de lípidos y ácidos grasos, producción de bacteriocinas, sistemas toxina-antitoxina, defensa bacteriana, producción de peróxidos, adhesión y formación de biopelículas, resistencia a estrés oxidativo, regulación positiva del sistema inmune, regulación negativa del sistema inmune, asociados a receptores

toll-like, no clasificados de modulación del sistema inmune) propuestas desde el análisis bibliográfico de las características de microorganismos probióticos, además de integrar 23 campos de información biológica-funcional obtenida del cruce de información con las bases de datos de UNIPROT, NCBI, EMBL, GO, PUBMED, ENZIME, PFAM, y PROSITE.

Finalmente, logramos establecer un sistema unificado capaz de integrar información biológica, funcional y bibliográfica para anotación funcional de secuencias genómicas de microorganismos en términos de sus facultades probióticas, con el cual se analizaron 4 especies de *Lactobacillus* (*L. fermentum* L26, *L. fermentum* L33, *L. plantarum* L90, y *L. rhamnosus* L134), el cual permitió discriminar *Lactobacillus fermentum* L26 como el potencial probiótico más inocuo y a *Lactobacillus rhamnosus* L134 como el potencial probiótico con mayores facultades.

Este sistema de anotación nos permite priorizar y definir un *ranking* de microorganismo candidatos para acelerar el descubrimiento de nuevos probióticos, a la vez que caracterizamos los componentes moleculares responsables de estas propiedades.

8 BIBLIOGRAFÍA

Aureli, Paolo, Lucio Capurso, Anna Maria Castellazzi, Mario Clerici, Marcello Giovannini, Lorenzo Morelli, Andrea Poli, Fabrizio Pregliasco, Filippo Salvini, y Gian Vincenzo Zuccotti. 2011. “Probiotics and Health: An Evidence-Based Review”. Pharmacological Research 63 (5): 366–76. https://doi.org/10.1016/j.phrs.2011.02.006.

ASALE, RAE-, y RAE. s. f. “homología | Diccionario de la lengua española”. «Diccionario de la lengua española» - Edición del Tricentenario. Accedido 12 de mayo de 2020. https://dle.rae.es/homología.

Azam, Rosa, Soudeh Ghafouri-Fard, Mina Tabrizi, Mohammad-Hossein Modarressi, Reza Ebrahimzadeh-Vesal, Maryam Daneshvar, Maryam Beigom Mobasheri, y Elahe Motevaseli. 2014. “Lactobacillus Acidophilus and Lactobacillus Crispatus Culture Supernatants Downregulate Expression of Cancer-Testis Genes in the MDA-MB-231 Cell Line”. *Asian Pacific Journal of Cancer Prevention* 15 (10): 4255–59. https://doi.org/10.7314/APJCP.2014.15.10.4255.

Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, et al. 2008. “The RAST Server: Rapid Annotations using Subsystems Technology”. *BMC Genomics* 9 (1): 75. https://doi.org/10.1186/1471-2164-9-75.

Bairoch, Amos. 2000. “The ENZYME database in 2000”. *Nucleic Acids Research* 28 (1): 304–5. https://doi.org/10.1093/nar/28.1.304.

Barbieri, Federica, Chiara Montanari, Fausto Gardini, y Giulia Tabanelli. 2019. “Biogenic Amine Production by Lactic Acid Bacteria: A Review”. *Foods* 8 (1). https://doi.org/10.3390/foods8010017.

Bulgasem, Bulgasem, Mohd Lani, Zaiton Hassan, Wan Mohtar, Wan Yusoff, y Sumaya Fnaish. 2016. “Mycobiology Antifungal Activity of Lactic Acid Bacteria Strains Isolated from Natural Honey against Pathogenic Candida Species”. *Mycobiology* 44 (noviembre). https://doi.org/10.5941/MYCO.2016.44.4.302.

Capozzi, Vittorio, Pasquale Russo, María Teresa Dueñas, Paloma López, y Giuseppe Spano. 2012. “Lactic Acid Bacteria Producing B-Group Vitamins: A Great Potential for Functional Cereals Products”. *Applied Microbiology and Biotechnology* 96 (6): 1383–94. https://doi.org/10.1007/s00253-012-4440-2.

Cha, Min-Kyeong, Do-Kyung Lee, Hyang-Mi An, Si-Won Lee, Seon-Hee Shin, Jeong-Hyun Kwon, Kyung-Jae Kim, y Nam-Joo Ha. 2012. “Antiviral Activity of Bifidobacterium AdolescentisSPM1005-A on Human Papillomavirus Type 16”. *BMC Medicine* 10 (1): 72. https://doi.org/10.1186/1741-7015-10-72.

Chaisson, Mark J. P., Richard K. Wilson, y Evan E. Eichler. 2015. “Genetic Variation and the de Novo Assembly of Human Genomes”. *Nature Reviews Genetics* 16 (11): 627–40. https://doi.org/10.1038/nrg3933.

Chen, Chuming, Hongzhan Huang, y Cathy H. Wu. 2017. “Protein Bioinformatics

Databases and Resources". En *Protein Bioinformatics*, editado por Cathy H. Wu, Cecilia N. Arighi, y Karen E. Ross, 1558:3–39. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6783-4_1.

Chen, F., L. Zhu, H. Qiu, F. Chen, L. Zhu, y H. Qiu. 2017. "Isolation and Probiotic Potential of Lactobacillus Salivarius and Pediococcus Pentosaceus in Specific Pathogen Free Chickens". *Brazilian Journal of Poultry Science* 19 (2): 325–32. https://doi.org/10.1590/1806-9061-2016-0413.

Cotter, Paul D., y Colin Hill. 2003. "Surviving the Acid Test: Responses of Gram-Positive Bacteria to Low PH". *Microbiology and Molecular Biology Reviews: MMBR* 67 (3): 429–53, table of contents. https://doi.org/10.1128/mmbr.67.3.429-453.2003.

Croft, David, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, et al. 2011. "Reactome: a database of reactions, pathways and biological processes". *Nucleic Acids Research* 39 (suppl_1): D691–97. https://doi.org/10.1093/nar/gkq1018.

Darwin, K. Heran, y Carl F. Nathan. 2005. "Role for Nucleotide Excision Repair in Virulence of Mycobacterium Tuberculosis". *Infection and Immunity* 73 (8): 4581–87. https://doi.org/10.1128/IAI.73.8.4581-4587.2005.

Do, Chuong B., y Kazutaka Katoh. 2008a. "Protein Multiple Sequence Alignment". En *Functional Proteomics*, editado por Julie D. Thompson, Marius Ueffing, y Christine Schaeffer-Reiss, 484:379–413. Methods in Molecular Biology. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-59745-398-1_25.

———. 2008b. "Protein Multiple Sequence Alignment". En *Functional Proteomics*, editado por Julie D. Thompson, Marius Ueffing, y Christine Schaeffer-Reiss, 484:379–413. Methods in Molecular Biology. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-59745-398-1_25.

Dobson, Alleson, Paul D. Cotter, R. Paul Ross, y Colin Hill. 2012. "Bacteriocin Production: A Probiotic Trait?" *Applied and Environmental Microbiology* 78 (1): 1–6. https://doi.org/10.1128/AEM.05576-11.

Edgar, Robert C. 2004. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Ellegren, Hans. 2008. "Comparative Genomics and the Study of Evolution by Natural Selection". *Molecular Ecology* 17 (21): 4586–96. https://doi.org/10.1111/j.1365-294X.2008.03954.x.

Ernst, Carl, y Cynthia C. Morton. 2013. "Identification and Function of Long Non-Coding RNA". *Frontiers in Cellular Neuroscience* 7. https://doi.org/10.3389/fncel.2013.00168.

Falasconi, Irene, Alessandra Fontana, Vania Patrone, Annalisa Rebecchi, Guillermo Duserm Garrido, Laura Principato, Maria Luisa Callegari, Giorgia Spigno, y Lorenzo Morelli. 2020. "Genome-Assisted Characterization of Lactobacillus Fermentum, Weissella Cibaria, and Weissella Confusa Strains Isolated from Sorghum as Starters for Sourdough Fermentation". *Microorganisms* 8 (9). https://doi.org/10.3390/microorganisms8091388.

Feldgarden, Michael, Vyacheslav Brover, Daniel H. Haft, Arjun B. Prasad, Douglas J. Slotta, Igor Tolstoy, Gregory H. Tyson, et al. 2019. "Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates". *Antimicrobial Agents and Chemotherapy* 63 (11). https://doi.org/10.1128/AAC.00483-19.

Fiddes, Ian T., Joel Armstrong, Mark Diekhans, Stefanie Nachtweide, Zev N. Kronenberg, Jason G. Underwood, David Gordon, et al. 2018. “Comparative Annotation Toolkit (CAT)—Simultaneous Clade and Personal Genome Annotation”. *Genome Research* 28 (7): 1029–38. https://doi.org/10.1101/gr.233460.117.

Flint, Harry J., Karen P. Scott, Sylvia H. Duncan, Petra Louis, y Evelyne Forano. 2012. “Microbial degradation of complex carbohydrates in the gut”. *Gut Microbes* 3 (4): 289–306. https://doi.org/10.4161/gmic.19897.

Fitch, Walter M. 1970. “Distinguishing Homologous from Analogous Proteins”. *Systematic Biology* 19 (2): 99–113. https://doi.org/10.2307/2412448.

Franz, Charles, Marc Vancanneyt, Katrien Vandemeulebroecke, Marjan Wachter, Ilse Cleenwerck, Bart Hoste, Ulrich Schillinger, y Jean Swings. 2006. “Pediococcus stilesii sp. no., isolated from maize grains”. *International journal of systematic and evolutionary microbiology* 56 (marzo): 329–33. https://doi.org/10.1099/ijs.0.63944-0.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, y Weizhong Li. 2012. “CD-HIT: accelerated for clustering the next-generation sequencing data”. *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

“Gene Ontology Consortium: Going Forward”. 2015. *Nucleic Acids Research* 43 (D1): D1049–56. https://doi.org/10.1093/nar/gku1179.

Henikoff, S, y J G Henikoff. 1992. “Amino acid substitution matrices from protein blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915–19.

Hill, Colin, Francisco Guarner, Gregor Reid, Glenn R. Gibson, Daniel J. Merenstein, Bruno Pot, Lorenzo Morelli, et al. 2014a. “Expert Consensus Document. The International Scientific Association for Probiotics and Prebiotics Consensus Statement on the Scope and Appropriate Use of the Term Probiotic”. *Nature Reviews. Gastroenterology & Hepatology* 11 (8): 506–14. https://doi.org/10.1038/nrgastro.2014.66.

———. 2014b. “The International Scientific Association for Probiotics and Prebiotics Consensus Statement on the Scope and Appropriate Use of the Term Probiotic”. *Nature Reviews Gastroenterology & Hepatology* 11 (8): 506–14. https://doi.org/10.1038/nrgastro.2014.66.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, y Mario Stanke. 2016. “BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.” *Bioinformatics* 32 (5): 767–69. https://doi.org/10.1093/bioinformatics/btv661.

Holland, Heinrich D. 2006. “The Oxygenation of the Atmosphere and Oceans”. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1470): 903–15. https://doi.org/10.1098/rstb.2006.1838.

Holt, Carson, y Mark Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects”. *BMC Bioinformatics* 12 (1): 491. https://doi.org/10.1186/1471-2105-12-491.

Hulo, Nicolas, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S. Langendijk-Genevaux, Marco Pagni, y Christian J. A. Sigrist. 2006. “The PROSITE database”. *Nucleic Acids Research* 34 (suppl_1): D227–30. https://doi.org/10.1093/nar/gkj063.

Ishibashi, N., y S. Yamazaki. 2001. “Probiotics and Safety”. *The American Journal of Clinical Nutrition* 73 (2 Suppl): 465S-470S. https://doi.org/10.1093/ajcn/73.2.465s.

Iyer, Chandra, Astrid Kosters, Gautam Sethi, Ajaikumar B. Kunnumakkara, Bharat B. Aggarwal, y James Versalovic. 2008. “Probiotic Lactobacillus Reuteri Promotes TNF-Induced Apoptosis in Human Myeloid Leukemia-Derived Cells by Modulation of NF-KB and MAPK Signalling”. *Cellular Microbiology* 10 (7): 1442–52. https://doi.org/10.1111/j.1462-5822.2008.01137.x.

Janik, Rafal, Lynsie A. M. Thomason, Andrew M. Stanisz, Paul Forsythe, John Bienenstock, y Greg J. Stanisz. 2016. “Magnetic Resonance Spectroscopy Reveals Oral Lactobacillus Promotion of Increases in Brain GABA, N-Acetyl Aspartate and Glutamate”. *NeuroImage* 125 (enero): 988–95. https://doi.org/10.1016/j.neuroimage.2015.11.018.

Jester, Brian C., Pascale Romby, y Efthimia Lioliou. 2012. “When Ribonucleases Come into Play in Pathogens: A Survey of Gram-Positive Bacteria”. *International Journal of Microbiology* 2012. https://doi.org/10.1155/2012/592196.

Kleerebezem, Michiel, Jos Boekhorst, Richard van Kranenburg, Douwe Molenaar, Oscar P. Kuipers, Rob Leer, Renato Tarchini, et al. 2003. “Complete Genome Sequence of Lactobacillus Plantarum WCFS1”. *Proceedings of the National Academy of Sciences of the United States of America* 100 (4): 1990–95. https://doi.org/10.1073/pnas.0337704100.

Krogh, A., M. Brown, I. S. Mian, K. Sjölander, y D. Haussler. 1994. “Hidden Markov Models in Computational Biology. Applications to Protein Modeling”. *Journal of Molecular Biology* 235 (5): 1501–31. https://doi.org/10.1006/jmbi.1994.1104.

Kumar, Sudhir, y Alan Filipski. 2007. “Multiple Sequence Alignment: In Pursuit of Homologous DNA Positions”. *Genome Research* 17 (2): 127–35. https://doi.org/10.1101/gr.5232407.

Lebeer, Sarah, Peter A Bron, Maria L Marco, Jan-Peter Van Pijkeren, Mary O’Connell Motherway, Colin Hill, Bruno Pot, Stefan Roos, y Todd Klaenhammer. 2018. “Identification of Probiotic Effector Molecules: Present State and Future Perspectives”. *Current Opinion in Biotechnology*, Food biotechnology • Plant biotechnology, 49 (febrero): 217–23. https://doi.org/10.1016/j.copbio.2017.10.007.

Lebeer, Sarah, Jos Vanderleyden, y Sigrid C. J. De Keersmaecker. 2008. “Genes and Molecules of Lactobacilli Supporting Probiotic Action”. *Microbiology and Molecular Biology Reviews: MMBR* 72 (4): 728–64, Table of Contents. https://doi.org/10.1128/MMBR.00017-08.

LeBlanc, J. G., J. E. Laiño, M. Juarez del Valle, V. Vannini, D. van Sinderen, M. P. Taranto, G. Font de Valdez, G. Savoy de Giori, y F. Sesma. 2011. “B-Group Vitamin Production by Lactic Acid Bacteria – Current Knowledge and Potential Applications”. *Journal of Applied Microbiology* 111 (6): 1297–1309. https://doi.org/10.1111/j.1365-2672.2011.05157.x.

Lima, Tania, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Karine Michoud, Catherine Rivoire, Virginie Bulliard, et al. 2009. “HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot”. *Nucleic Acids Research* 37 (suppl_1): D471–78. https://doi.org/10.1093/nar/gkn661.

Llewellyn, Amy, y Andrew Foey. 2017. “Probiotic Modulation of Innate Cell Pathogen

Sensing and Signaling Events". *Nutrients* 9 (10). https://doi.org/10.3390/nu9101156.
Levit, R., G. Savoy de Giori, A. de Moreno de LeBlanc, y J. G. LeBlanc. s. f. "Recent Update on Lactic Acid Bacteria Producing Riboflavin and Folates: Application for Food Fortification and Treatment of Intestinal Inflammation". *Journal of Applied Microbiology* n/a (n/a). Accedido 21 de enero de 2021. https://doi.org/10.1111/jam.14854.
Matera, A. Gregory, Rebecca M. Terns, y Michael P. Terns. 2007. "Non-Coding RNAs: Lessons from the Small Nuclear and Small Nucleolar RNAs". *Nature Reviews Molecular Cell Biology* 8 (3): 209–20. https://doi.org/10.1038/nrm2124.
Ménard, Odile, Valérie Gafa, Nathalie Kapel, Bertrand Rodriguez, Marie-José Butel, y Anne-Judith Waligora-Dupriet. 2010. "Characterization of Immunostimulatory CpG-Rich Sequences from Different Bifidobacterium Species". *Applied and Environmental Microbiology* 76 (9): 2846–55. https://doi.org/10.1128/AEM.01714-09.
Metzler, Dirk. 2003. "Statistical alignment based on fragment insertion and deletion models". *Bioinformatics* 19 (4): 490–99. https://doi.org/10.1093/bioinformatics/btg026.
Miklós, I., G. A. Lunter, y I. Holmes. 2004. "A 'Long Indel' Model For Evolutionary Sequence Alignment". *Molecular Biology and Evolution* 21 (3): 529–40. https://doi.org/10.1093/molbev/msh043.
Miller, Jason R., Sergey Koren, y Granger Sutton. 2010. "Assembly Algorithms for Next-Generation Sequencing Data". *Genomics* 95 (6): 315–27. https://doi.org/10.1016/j.ygeno.2010.03.001.
Morelli, Lorenzo, y Lucio Capurso. 2012. "FAO/WHO Guidelines on Probiotics: 10 Years Later". *Journal of Clinical Gastroenterology* 46 (octubre): S1. https://doi.org/10.1097/MCG.0b013e318269fdd5.
Nair, Shamila, Eliane Milohanic, y Patrick Berche. 2000. "ClpC ATPase Is Required for Cell Adhesion and Invasion of Listeria Monocytogenes". *Infection and Immunity* 68 (12): 7061–68. https://doi.org/10.1128/IAI.68.12.7061-7068.2000.
Nakamura, Y., T. Gojobori, y T. Ikemura. 2000. "Codon Usage Tabulated from International DNA Sequence Databases: Status for the Year 2000". *Nucleic Acids Research* 28 (1): 292. https://doi.org/10.1093/nar/28.1.292.
Nakayama, Toru, y Teruo Amachi. 2002. "β-Galactosidase, Enzymology". En *Encyclopedia of Bioprocess Technology*. American Cancer Society. https://doi.org/10.1002/0471250589.ebt102.
Natale, Darren A., Cecilia N. Arighi, Judith A. Blake, Carol J. Bult, Karen R. Christie, Julie Cowart, Peter D'Eustachio, et al. 2014. "Protein Ontology: a controlled structured network of protein entities". *Nucleic Acids Research* 42 (D1): D415–21. https://doi.org/10.1093/nar/gkt1173.
Notredame, Cédric. 2007. "Recent Evolutions of Multiple Sequence Alignment Algorithms". *PLOS Computational Biology* 3 (8): e123. https://doi.org/10.1371/journal.pcbi.0030123.
Peek, Richard M., y Martin J. Blaser. 2002. "Helicobacter Pylori and Gastrointestinal Tract Adenocarcinomas". *Nature Reviews Cancer* 2 (1): 28–37. https://doi.org/10.1038/nrc703.
Phan, Ngoc Quang, Takashi Uebanso, Takaaki Shimohata, Mutsumi Nakahashi, Kazuaki Mawatari, y Akira Takahashi. 2015. "DNA-Binding Protein HU Coordinates Pathogenicity in Vibrio parahaemolyticus". *Journal of*

Bacteriology 197 (18): 2958–64. https://doi.org/10.1128/JB.00306-15.
Rabe, L. K., y S. L. Hillier. 2003. “Optimization of Media for Detection of Hydrogen Peroxide Production by Lactobacillus Species”. *Journal of Clinical Microbiology* 41 (7): 3260–64. https://doi.org/10.1128/JCM.41.7.3260-3264.2003.
Racedo, Silvia, Julio Villena, Marcela Medina, Graciela Agüero, Virginia Rodríguez, y Susana Alvarez. 2006. “Lactobacillus Casei Administration Reduces Lung Injuries in a Streptococcus Pneumoniae Infection in Mice”. *Microbes and Infection* 8 (9): 2359–66. https://doi.org/10.1016/j.micinf.2006.04.022.
Ramón y Cajal, Santiago, Miguel F. Segura, y Stefan Hümmer. 2019. “Interplay Between ncRNAs and Cellular Communication: A Proposal for Understanding Cell-Specific Signaling Pathways”. *Frontiers in Genetics* 10 (abril). https://doi.org/10.3389/fgene.2019.00281.
Roselino, Mariana Nougalli, Leonardo Fonseca Maciel, Veronica Sirocchi, Matteo Caviglia, Gianni Sagratini, Sauro Vittori, María Pía Taranto, y Daniela Cardoso Umbelino Cavallini. 2020. “Analysis of Biogenic Amines in Probiotic and Commercial Salamis”. *Journal of Food Composition and Analysis* 94 (diciembre): 103649. https://doi.org/10.1016/j.jfca.2020.103649.
Ruas-Madiedo, Patricia, Miguel Gueimonde, Abelardo Margolles, Clara G. de los REYES-GAVILÁN, y Seppo Salminen. 2006. “Exopolysaccharides Produced by Probiotic Strains Modify the Adhesion of Probiotics and Enteropathogens to Human Intestinal Mucus”. *Journal of Food Protection* 69 (8): 2011–15. https://doi.org/10.4315/0362-028X-69.8.2011.
Sandes, Sávio, Luige Alvim, Bruno Silva, Leonardo Acurcio, Cinara Santos, Márcia Campos, Camila Santos, Jacques Nicoli, Elisabeth Neumann, y Álvaro Nunes. 2017. “Selection of New Lactic Acid Bacteria Strains Bearing Probiotic Features from Mucosal Microbiota of Healthy Calves: Looking for Immunobiotics through in Vitro and in Vivo Approaches for Immunoprophylaxis Applications”. *Microbiological Research* 200 (julio): 1–13. https://doi.org/10.1016/j.micres.2017.03.008.
Schickel, R., B. Boyerinas, S.-M. Park, y M. E. Peter. 2008. “MicroRNAs: Key Players in the Immune System, Differentiation, Tumorigenesis and Cell Death”. *Oncogene* 27 (45): 5959–74. https://doi.org/10.1038/onc.2008.274.
Schomburg, Ida, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, y Dietmar Schomburg. 2004. “BRENDA, the enzyme database: updates and major new developments”. *Nucleic Acids Research* 32 (suppl_1): D431–33. https://doi.org/10.1093/nar/gkh081.
“Scientific Opinion on Risk Based Control of Biogenic Amine Formation in Fermented Foods”. 2011. *EFSA Journal* 9 (10): 2393. https://doi.org/10.2903/j.efsa.2011.2393.
Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation”. *Bioinformatics (Oxford, England)* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.
Silva, B. C., L. R. C. Jung, S. H. C. Sandes, L. B. Alvim, M. R. Q. Bomfim, J. R. Nicoli, E. Neumann, y A. C. Nunes. 2013. “In Vitro Assessment of Functional Properties of Lactic Acid Bacteria Isolated from Faecal Microbiota of Healthy Dogs for Potential Use as Probiotics”. Text. Wageningen Academic Publishers. 1 de septiembre de 2013. https://doi.org/info:doi/10.3920/BM2012.0048.

Sippl, Manfred J. 1999. “Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids, Edited by R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson. 1998. Cambridge: Cambridge University Press. 356 Pp. £55.00 ($80.00) (Hardcover); £19.95 ($34.95) (Paper).” *Protein Science* 8 (3): 695–695. https://doi.org/10.1110/ps.8.3.695.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, y Mike Tyers. 2006. “BioGRID: a general repository for interaction datasets”. *Nucleic Acids Research* 34 (suppl_1): D535–39. https://doi.org/10.1093/nar/gkj109.

Stein, Lincoln. 2001. “Genome Annotation: From Sequence to Biology”. *Nature Reviews Genetics* 2 (7): 493–503. https://doi.org/10.1038/35080529.

Takagi, Akimitsu, Mitsuyoshi Kano, y Chiaki Kaga. 2015. “Possibility of Breast Cancer Prevention: Use of Soy Isoflavones and Fermented Soy Beverage Produced Using Probiotics”. *International Journal of Molecular Sciences* 16 (5): 10907–20. https://doi.org/10.3390/ijms160510907.

The UniProt Consortium. 2015. “UniProt: a hub for protein information”. *Nucleic Acids Research* 43 (D1): D204–12. https://doi.org/10.1093/nar/gku989.

Thibaud-Nissen, Françoise, Alexander Souvorov, Terence Murphy, Michael DiCuccio, y Paul Kitts. 2013. *Eukaryotic Genome Annotation Pipeline*. *The NCBI Handbook [Internet]. 2nd Edition*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/sites/books/NBK169439/.

Van Tassell, Maxwell L., y Michael J. Miller. 2011. “Lactobacillus Adhesion to Mucus”. *Nutrients* 3 (5): 613–36. https://doi.org/10.3390/nu3050613.

Vastano, Valeria, Annunziata Pagano, Alessandra Fusco, Gianluca Merola, Margherita Sacco, y Giovanna Donnarumma. 2016. “The Lactobacillus Plantarum Eno A1 Enolase Is Involved in Immunostimulation of Caco-2 Cells and in Biofilm Development”. *Advances in Experimental Medicine and Biology* 897: 33–44. https://doi.org/10.1007/5584_2015_5009.

Verhoeven, Veronique, Nathalie Renard, Amin Makar, Paul Van Royen, John-Paul Bogers, Filip Lardon, Marc Peeters, y Marc Baay. 2013. “Probiotics Enhance the Clearance of Human Papillomavirus-Related Cervical Lesions: A Prospective Controlled Pilot Study”. *European Journal of Cancer Prevention* 22 (1): 46–51. https://doi.org/10.1097/CEJ.0b013e328355ed23.

Villena, Julio, Eriko Chiba, Yohsuke Tomosada, Susana Salva, Gabriela Marranzino, Haruki Kitazawa, y Susana Alvarez. 2012. “Orally Administered Lactobacillus Rhamnosus Modulates the Respiratory Immune Response Triggered by the Viral Pathogen-Associated Molecular Pattern Poly(I:C)”. *BMC Immunology* 13 (1): 1–15. https://doi.org/10.1186/1471-2172-13-53.

Villena, Julio, y Haruki Kitazawa. 2014. “Modulation of Intestinal TLR4-Inflammatory Signaling Pathways by Probiotic Microorganisms: Lessons Learned from Lactobacillus Jensenii TL2937”. *Frontiers in Immunology* 4. https://doi.org/10.3389/fimmu.2013.00512.

Wilkins, Marc R., Keith L. Williams, Ron D. Appel, y Denis F. Hochstrasser. 2013. *Proteome Research: New Frontiers in Functional Genomics*. Springer Science & Business Media.

Yamaguchi, Yoshihiro, Jung-Ho Park, y Masayori Inouye. 2011. “Toxin-Antitoxin Systems in Bacteria and Archaea”. *Annual Review of Genetics* 45 (1): 61–79. https://doi.org/10.1146/annurev-genet-110410-132412.

Yan, Fang, y D.B. Polk. 2011. “Probiotics and immune health”. *Current opinion in gastroenterology* 27 (6): 496–501.

https://doi.org/10.1097/MOG.0b013e32834baa4d.
Yao, Kecheng, Linghai Zeng, Qian He, Wei Wang, Jiao Lei, y Xiulan Zou. 2017. “Effect of Probiotics on Glucose and Lipid Metabolism in Type 2 Diabetes Mellitus: A Meta-Analysis of 12 Randomized Controlled Trials”. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 23 (junio): 3044–53. https://doi.org/10.12659/msm.902600.
Yu, Ai-Qun, y Lianqin Li. 2016. “The Potential Role of Probiotics in Cancer Prevention and Treatment”. *Nutrition and Cancer* 68 (4): 535–44. https://doi.org/10.1080/01635581.2016.1158300.
Zhang, Zemin, y William I. Wood. 2003. “A profile hidden Markov model for signal peptides generated by HMMER”. *Bioinformatics* 19 (2): 307–8. https://doi.org/10.1093/bioinformatics/19.2.307.
Zheng, Jinshui, Lifang Ruan, Ming Sun, y Michael Gänzle. 2015. “A Genomic View of Lactobacilli and Pediococci Demonstrates That Phylogeny Matches Ecology and Physiology”. *Applied and Environmental Microbiology* 81 (20): 7233–43. https://doi.org/10.1128/AEM.02116-15.
Zuo, Fanglei, Shangwu Chen, y Harold Marcotte. 2020. “Engineer Probiotic Bifidobacteria for Food and Biomedical Applications - Current Status and Future Prospective”. *Biotechnology Advances* 45 (diciembre): 107654. https://doi.org/10.1016/j.biotechadv.2020.107654.

8 ANEXOS

Tabla A-I: Características de beneficios contra enfermedades a la salud de algunas especies probióticas. Extraído de Aureli et al. 2011.

Desorden	Cepa	Dosis
Tratamiento de la diarrea infecciosa en niños	*L. rhamnosus* GG	10^{10}-10^{11} **ufc**
	L. reuteri ATCC 55730	10^{10}-10^{11} **ufc x 2/d**
	S. cerevisiae (*boulardii*)	10^{9} **ufc x 3/d**
Tratamiento de la diarrea infecciosa en adultos	*Enterococcus faecium* LAB SF68	10^{8} **ufc x 3/d**
Prevención de diarrea asociada a antibióticos	*S. cerevisiae (boulardii)*	10^{9} **ufc x 2/d**
	L. rhamnosus GG	10^{10} **ufc x 1-2/d**
	B. lactis Bb12 + *S. thermophilus*	10^{7} + 10^{6} g**/fórmula**
	Enterococcus faecium LAB SF68	10^{8} ufc x 2/d
	S. cerevisiae (*boulardii*)	1g o 3 x 10^{10} ufc x 1/d
	L. rhamnosus GG	10^{10}-10^{11} **ufc x 2/d**
	L. casei DN-114 001 en leche fermentada con *L. bulgaricus* +*S. thermophilus*	10^{10} **ufc x 2/d**
	B. clausii	2 x 10^{9} esporas x 3/d
	L. acidophilus CL1285 + *L. casei*	5 x 10^{10} ufc x 1/d
Prevención de la infección nosocomial por rotavirus en niños	*L. rhamnosus* GG	10^{10}-10^{11} ufc x 2/d
	B. lactis Bb12 + *S. thermophilus*	10^{8} + 10^{7} ufc/g fórmula
	B. lactis Bb12	10^{9} ufc x 2/d
	L. reuteri ATCC 55730	10^{9} ufc x 2/d
Prevención de la infección por *C. difficile* en adultos	*L. casei* DN-114 001 en leche fermentada con *L. bulgaricus* +*S. thermophilus*	10^{10} ufc x 2/d
	S. cerevisiae (boulardii)	2 x 10^{10} ufc x 1/d
Adyuvante en terapias para la erradicación de *Helicobacter Pylori*	*L. rhamnosus* GG	6 x 10^{9} ufc x 2/d
	B. clausii	2 x 10^{9} esporas x 3/d
	S. cerevisiae (boulardii)	1g o 5 x 10^{9} ufc x d
	L. casei DN-114 001 en leche fermentada con *L. bulgaricus* +*S. thermophilus*	10^{10} ufc x 2/d
Reducción de los síntomas del síndrome del intestino irritable	*B. infantis* 35624	10^{8} ufc x 1/d
	L. rhamnosus GG	6 x 10^{9} ufc x 2/d
	B. longum, *B. infantis, B. breve, L.acidophilus*, *L. casei*, *L.delbrueckii* subsp. *bulgaricus*, *L. plantarum*, *S. salivarius* subsp. *thermophilus*	4.5 x 10^{11} ufc x 2/d
	L. rhamnosus GG, *L. rhamnosus* LC705, *B.breve Bb99* y *Pfreudenreichii* subsp. *shermanii* JS	10^{10} ufc x 1/d

	B. animalis DN-173 0101 en leche fermentada con *L. bulgaricus* + *S. thermophilus*	10^{10} ufc x 2/d
Remisión de la colitis ulcerosa	*E. coli Nissle* 1917	5 x 10^{10} ufc x 2/d
Remisión de la pouchitis	*B. longum*, *B. infantis, B. breve, L.acidophilus*, *L. casei*, *L.delbrueckii* subsp. *bulgaricus*, *L. plantarum*, *S. salivarius* subsp. *thermophilus*	4.5 x 10^{11} ufc x
	B. infantis, S. salivarius subsp. *thermophilus*, *B. bifidum*	3.5 X 10^{8} ufc por cepa x 1/d
Prevención de la enterocolitis necrotizante	*L. acidophilus* + *B. infantis*	10^{9} ufc por cepa x 2/d

Tabla A-II: Criterios de búsqueda por códigos de GO utilizados para la búsqueda de secuencias asociadas a las características de GOPRODB.

Categoría	Subcategoría	Criterio de búsqueda por función molecular GO	Código GO	Secuencias encontradas
Resistencia a antibióticos	-	proceso catabólico de antibióticos amiloglucosidasa	30649	4
		péptidos de proceso catabólico de antibióticos	30652	0
		proceso catabólico de antibióticos betalactamasa	30655	81
		proceso catabólico de antibióticos exógenos	42740	0
		proceso catabólico de antibióticos endógenos	42741	0
		proceso catabólico de antibióticos	17001	2209
		complejo tBC	1990203	0
		proceso catabólico de penicilina	42317	0
		regulación positiva de la catálisis de penicilina	33248	53
		regulación negativa de la catálisis de penicilina	33249	0
		regulación de la catálisis de penicilina	33247	0
		inmunidad a bacteriocinas	30153	0
		catálisis de bacteriocinas	46225	0
		proceso catabólico de isoflavonoides fitoalexina	46290	0
		proceso catabólico de flavonoides fitoalexina	46286	0
Patogenicidad	-	patogénesis	9405	2770
Sobrevivencia al tracto intestinal	Resistencia a PH ácido	respuesta celular a PH ácido	71468	107
	Resistencia a sales biliares	proceso catabólico de sales biliares	30573	22
	Actividad ureasa	proceso catabólico de urea	43419	828
Producción de compuestos bioactivos	Producción vitaminas del complejo B (cobalamina)	biosíntesis de cobalamina	9236	1876
		biosíntesis aeróbica de cobalamina	19250	1
		biosintesis anaeróbica de cobalamina	19251	22
		actividad cobalamina-5 sintasa	8818	247
		actividad adenosilcobalamina-GDP ribazoltransferasa	51073	247
		biosíntesis de alfa-ribazol	97290	0
	Producción vitaminas del complejo B (riboflavina)	biosintesis de riboflavina	9231	1191
		complejo riboflavina sintasa	9349	621
		actividad proteasa FMN-fosfato	103027	0
		actividad FMN-hidrolasa	90711	1
		actividad FAD-AMP-liasa cíclica	34012	5

	Producción vitaminas del complejo B (biotina)	biosíntesis de biotina	9102	1418
		actividad biotina sintasa	4076	594
	Producción vitaminas del complejo B (piridoxina)	biosíntesis de vitamina B6	42819	1838
		biosíntesis de piridoxina	8615	1293
		actividad piridoxina 5-fosfato-sintasa	33856	259
	Actividad Lactasa	actividad beta-galactosidasa (isomerización de lactosa)	103033	0
		actividad beta-galactosidasa	4565	218
		actividad lactasa	16	1269
		proceso catabólico de lactosa usando glucósido-3-deshidrogenasa	19513	0
		proceso catabólico de lactosa via tagatosa-6-fosfato	19512	253
		proceso catabólico de lactosa via UDP-galactosa	19515	0
		proceso catabólico de lactosa	5990	261
	Degradación de lípidos y ácidos grasos	proceso catabólico de ácidos grasos de cadena media	51793	14
		proceso de degradación de ácidos grasos	9062	922
		proceso catabólico de ácidos grasos monoinsaturados	1903965	1
		proceso catabólico de derivados de ácidos grasos	1901569	43
		proceso catabólico de ácidos grasos de cadena corta	19626	86
		proceso catabólico de ácidos grasos de cadena muy larga	42760	19
		proceso catabólico de ácidos grasos de cadena larga	42758	29
		proceso anaeróbico de catálisis de ácidos grasos	1990486	0
		metabolismo del simbionte de carbohidratos del huésped	52175	0
		catabolismo del simbionte de carbohidratos del huésped	52015	0
		metabolismo del huésped de carbohidratos del simbionte	52406	2
		catabolismo del huésped de carbohidratos del simbionte	52353	2
		catabolismo de lípidos intestinales	44258	5
		catabolismo de lípidos intestinales	16042	3387
Competitividad bacteriana	Producción de bacteriocinas	proceso biosintético de bacteriocinas	30152	24
	Sistemas toxina-antitoxina	complejo toxina-antitoxina	110001	3
		unión tipo II de par toxina-antitoxina	97351	28
	Defensa bacteriana	respuesta de defensa a bacterias gram negativa	50829	793
		respuesta de defensa a bacterias gram positivas	50830	808
		respuesta de defensa a hongos	50832	1434
		respuesta de defensa a virus	51607	1166
		respuesta de defensa bacterias	42742	3898
		respuesta de defensa a oocytes	2229	0
	Producción de peróxidos	biosíntesis de especies reactivas de oxígeno	1903409	168
		búsqueda por nombre de código de actividad enzimática	-	-
	Adhesión y formación de biopelículas	formación de biopelícula sumergida de una sola especie	90609	145
		formación de biopelículas en superficie	90607	0

		formación de biopelícula sumergida de múltiples especies	90608	10
		formación de biopelícula sumergida	90605	163
		formacion de biopeliculas en superficie de una sola especie	90606	5
		formación de biopelículas de múltiples especies en o sobre el huésped	44401	0
		formacion de biopeliculas en sustrato inanimado	44400	0
		formacion de biopeliculas de una sola especie en o sobre el huésped	44407	8
		formación de biopelículas de múltiples especies	44399	10
		ensamble de la matriz de biopelícula	98785	0
		organización de la matriz de biopelículas	98784	0
		adhesión celular involucrada la formación de biopelículas de múltiples especies	43710	10
		adhesión celular involucrada la formación de biopelículas de una sola especie en el huésped	47307	1
		adhesión celular involucrada la formación de biopelículas de una sola especie	43709	75
		adhesión celular involucrada en la formación de biopelículas	43708	93
		formación de biopelículas	42710	210
		matriz bacterial de biopelículas	97311	1
		matriz bacterial de biopelículas en superficie	97313	0
		componente de matriz bacterial de biopelículas	97312	0
		formacion de biopeliculas de una sola especie	44010	192
		formacion de biopeliculas de una sola especie en sustrato inanimado	44011	71
Regulación positiva del sistema inmune	-	regulación positiva del sistema de secreción de citoquinas por el simbionte	52035	1
		regulación positiva de la respuesta de defensa dependiente de genes de resistencia	52527	0
		regulación positiva de la respuesta de defensa del huésped	52509	0
		regulación positiva de la respuesta inmune del huésped	52556	0
		inducción de la respuesta inmune del huésped	52559	14
		regulación positiva de la producción de ROS relacionada con defensa	52369	0
		inducción de la respuesta inmune innata del huésped	52390	1
		regulación positiva de la cascada I-Kappa kinasa NF-kappa B del huésped por el simbionte	85033	1
		inducción de la respuesta de defensa del huésped por el simbionte	44416	24
		inducción de la respuesta inmune innata del huésped por virus	46738	0
		inducción de la respuesta humoral del huésped por virus	46736	0
		inducción de la respuesta inmune mediada por células del huésped por virus	46737	0
		inducción de la respuesta inmune del huésped por virus	46730	0
		regulación positiva de la producción de citoquinas por virus	44832	1
		inducción de efectores dependientes de la respuesta inmune del huésped por el simbionte	80815	0

		regulación positiva del comportamiento de búsqueda del huésped	32540	0
		regulación positiva de la endocitosis mediada por receptores	44078	2
		regulación positiva de la respuesta de defensa mediada por ácido salicílico	52074	0
		regulación positiva de las vías de transducción de señales MAP kinasa por el simbionte	52079	0
		regulación positiva del nivel de proteínas relacionada con defensa del huésped por simbionte	33664	1
		regulación positiva por simbionte de la transducción de señal mediada por proteínas quinasa del huésped	75131	1
		regulación positiva de la transducción de señal mediada por MAP quinasa en respuesta al huésped	75172	0
		regulación positiva por simbionte de transducción de señales mediada por proteínas quinasa en la respuesta del huésped	75169	0
		regulación positiva de niveles de citoquinas en el huésped	1990223	0
Regulación negativa del sistema inmune	-	regulación negativa por simbionte de la muerte celular programada relacionada con la defensa del huésped	34054	0
		actividad de señuelo del receptor PAMP	140320	0
		supresión por virus de la actividad del factor de transcripción NF-kappaB del huésped	39644	112
		supresión por virus de la activación de la célula asesina natural del huésped	39672	1
		evasión por virus de la actividad de las células NK del huésped	39671	8
		evasión por virus de la actividad de las células dendríticas del huésped	39673	4
		supresión por virus de la actividad de citoquinas del huésped	39518	14
		supresión por virus de la actividad de la proteína tirosina quinasa del huésped	39512	63
		supresión por virus de la cascada JAK-STAT del huésped	39514	208
		supresión por virus de la actividad del receptor de interferón del huésped	39511	3
		supresión por virus de la actividad del receptor de reconocimiento de patrones del huésped	39509	220
		supresión por virus del procesamiento del antígeno del huésped y presentación del antígeno peptídico a través de MHC clase II	39505	50
		supresión por virus de la vía de señalización mediada por interferón tipo I del huésped	39502	725
		supresión por virus de la producción de interferón tipo I del huésped	39501	12
		supresión por virus de la respuesta inmune adaptativa del huésped	39504	90
		supresión por virus de la respuesta inmune innata del huésped	39503	1216
		supresión por virus de la vía de señalización MDA-5 del huésped	39539	42
		supresión por virus de la vía de señalización RIG-I del huésped	39538	192

		supresión por virus de la vía de señalización del receptor del reconocimiento del patrón citoplasmático inducido por el virus del huésped	39537	422
		supresión por virus de la transducción de señales mediada por TRAF del huésped	39527	54
		supresión por virus de la actividad IRF7 del huésped	39557	59
		supresión por virus de la actividad quimiocina del huésped	39553	14
		supresión por virus de la actividad MDA-5 del huésped	39554	42
		supresión por virus de la actividad RIG-I del huésped	39540	192
		supresión por virus de la actividad MAVS del huésped	39545	213
		supresión por virus de la actividad IRF3 del huésped	39548	124
		supresión por virus de la actividad TRAF del huésped	39547	52
		supresión por virus de la activación del complemento del huésped	39573	19
		supresión por virus de la actividad ISG15 del huésped	39579	68
		supresión por virus de la actividad del host TYK2	39574	38
		supresión por virus de la actividad del host JAK1	39576	28
		supresión por virus de la actividad IRF9 del huésped	39560	5
		supresión por virus de la actividad STAT del huésped	39562	183
		supresión por virus de la actividad STAT2 del huésped	39564	116
		supresión por virus de la actividad STAT1 del huésped	39563	144
		supresión por el virus de la actividad de la proteína quinasa del huésped	39584	245
		supresión por virus de la actividad PKR del huésped	39580	174
		supresión por virus del procesamiento y presentación del antígeno del huésped	39588	89
		regulación negativa por simbionte de la respuesta inmune innata del huésped	52170	1226
		modulación por simbionte de los niveles de ácido salicílico en el huésped	52023	0
		regulación negativa por simbionte de la vía de transducción de señal mediada por ácido salicílico relacionada con la defensa del huésped	52003	0
		regulación negativa por simbionte de la respuesta de defensa mediada por ácido salicílico del huésped	52004	0
		regulación negativa por simbionte de la respuesta inflamatoria del huésped	52036	1
		regulación negativa por simbionte de la respuesta de defensa del huésped	52037	1227
		regulación negativa por simbionte de la respuesta inmune innata del huésped inducida por el patrón molecular asociada microorganismos	52034	0
		regulación negativa por simbionte de la respuesta inmune del huésped	52562	1324

		mantenimiento de la tolerancia simbionte a las moléculas de defensa del huésped	75145	0
		modificación por simbionte de la ubiquitinación de proteínas del huésped	75346	0
		supresión por virus de la vía de señalización de receptores toll-like del huésped	39722	9
		supresión por virus de la actividad IKBKE del host	39724	32
		supresión por virus de la actividad TBK1 del huésped	39723	9
		regulación negativa por simbionte de la secreción de citoquinas del huésped	140133	0
		regulación negativa por simbionte de la cascada de señales I-kappa NF-kappa B del huésped	85034	0
		supresión de las defensas del huésped	44414	1230
		evasión de las defensas del huésped	44413	0
		evasión o tolerancia de las defensas del huésped	44415	0
		supresión por virus del procesamiento y presentación péptidos de antígenos través de MHC clase I en el huésped	46776	68
		supresión por virus de la actividad de interferón intracelular del huésped	46774	4
		supresión por virus de la producción de citoquinas del huésped	46775	13
		evasión o tolerancia por virus de la respuesta inmune del huésped	30683	1600
		evasión o tolerancia de la respuesta de defensa del huésped	30682	1731
		evasión activa de la respuesta inmune del huésped	42783	21
		evasión activa de la respuesta inmune del huésped a través de la regulación del sistema del complemento del huésped	42784	5
		evasión activa de la respuesta inmune del huésped a través de la regulación de la red de citoquinas del huésped	42785	1
		evasión activa de la respuesta inmune del huésped a través de la regulación del procesamiento y presentación del antígeno del huésped	42786	1
		evasión pasiva de la respuesta inmune del huésped	42782	0
		evasión o tolerancia de la respuesta inmune del huésped	20012	1667
		regulación negativa del comportamiento de búsqueda del huésped	32539	1
		evasión o tolerancia de las defensas del huésped por virus	19049	1632
		regulación positiva por simbionte de la vía de transducción de señales mediada por ácido salicílico relacionada con la defensa del huésped	52072	0
		regulación negativa por simbionte de la vía de transducción de señal mediada por MAP quinasa relacionada con la defensa del huésped	52078	0
		regulación negativa por simbionte de la respuesta inmune mediada por células T del huésped	52085	6
		regulación negativa por simbionte de la respuesta inmune mediada por células B del huésped	52086	0
		regulación negativa por simbionte de la respuesta inmune mediada por células huésped	52083	8

		regulación negativa por simbionte del nivel de proteínas relacionadas con la defensa del huésped	33663	1
		regulación negativa por simbionte de la producción ROS del huésped relacionadas con la defensa	33661	0
		regulación negativa por simbionte de la respuesta de defensa dependiente del genes de resistencia del huésped	33660	0
		regulación negativa por simbionte de la transducción de señal mediada por proteínas quinasa del huésped	75132	0
		regulación negativa de la transducción de señal mediada por proteína quinasa en la respuesta del huésped	75170	0
		regulación negativa de la transducción de señal mediada por MAP quinasa en la respuesta del huésped	75173	0
No clasificados de modulación del sistema inmune	-	actividad ligando del receptor derivado de patógenos	140295	0
		modulación por virus de la señalización NIK / NF-kappaB del huésped	61765	1
		modulación por simbionte de la respuesta inmune mediada por células B del huésped	52154	0
		modulación por simbionte de la respuesta inmune innata del huésped inducida por patrones moleculares asociados a microorganismos	52157	3
		modulación por simbionte de la respuesta de defensa dependiente de genes de resistencia del huésped	52158	0
		modulación por simbionte de la respuesta inmune mediada por células del huésped	52155	9
		modulación por simbionte de la respuesta inmune mediada por células T del huésped	52156	6
		modulación por simbionte de la producción de especies ROS relacionadas con la defensa del huésped	52164	3
		modulación por simbionte de la producción de fitoalexina del huésped	52165	0
		modulación por simbionte de la respuesta inmune innata del huésped	52167	1235
		modulación por simbionte de la respuesta inflamatoria del huésped	52032	3
		modulación por simbionte de la respuesta de defensa del huésped	52031	1324
		modulación por simbionte de la respuesta inmune del huésped	52553	1371
		modulación por virus de la respuesta inmune del huésped	75528	1325
		regulación de la transducción de señal de respuesta mediada por MAP quinasa del huésped'	75171	0
		regulación de la transducción de señal mediada por proteína quinasa en la respuesta del huésped	75168	0
		respuesta a las moléculas de defensa del huésped	75140	0
		modulación por simbionte de la muerte celular programada relacionada con la defensa del huésped	34053	72

		modulación por simbionte de la cascada I-kappaB quinasa / NF-kappaB del huesped	85032	4
		modulación por virus de la producción de citoquinas del huésped	44831	14
		modulación por simbionte de endocitosis mediada por receptores del huésped	44077	2
		modulación por simbionte de la vía de transducción de señal mediada por MAP quinasa relacionadas con la defensa del huésped	52080	0
		modulación por simbionte del nivel de proteínas relacionadas con la defensa del huésped	33662	2
		regulación de la respuesta de defensa al virus por virus	50690	1627
Asociados a receptores toll-like	-	búsqueda por texto asociado a función molecular	-	-

Tabla A-III: Criterios de búsqueda por códigos de ENZIME utilizados para la búsqueda de secuencias asociadas a las características de GOPRODB. Muestra los criterios de búsqueda de la base de datos ENZIME utilizados para la recopilación de genes, y el número de secuencias encontradas.

Categoría	Subcategoría	Criterio de búsqueda por enzima	Código ENZIME	Secuencias encontradas
Sobrevivencia al tracto intestinal	Resistencia a pH ácido	Glutamato descarboxilasa	EC 4.1.1.15	43
		Arginina deaminasa	EC 3.5.3.6	148
	Actividad ureasa	Ureasa	EC 3.5.1.5	828
Competitividad bacteriana	Producción de peróxidos	Piruvato oxidasa	EC 1.2.3.3	93
		NADH oxidasa	EC 1.6.3.3	1
		NADH oxidasa	EC 1.6.3.4	3
		L - lactato oxidasa	EC 1.1.3.2	3
		Flavina reductasa	EC 1.5.1.36	16

Tabla A-IV: Criterios búsqueda de campo de texto utilizados para la búsqueda de secuencias asociadas a las características de GOPRODB. Muestra los criterios de búsqueda de texto utilizados para la recopilación de genes, y el número de secuencias encontradas. Aquellas búsquedas por nombre de gen fueron filtradas por dicha categoría y las demás filtradas manualmente.

Categoría	Subcategoría	Criterio de búsqueda por campo de texto	Secuencias encontradas
Sobrevivencia al tracto intestinal	Resistencia a PH ácido	nombre de gen: dnak	800
		nombre de gen: groel	1051
		nombre de gen: groes	703
		nombre de gen: hdea	10
		nombre de gen: hdeb	3
		f1f0 atpase	2948
Asociados a receptores toll-like	-	toll-like	46