14-8-2026

Análisis bioinformático de bacteriófagos activos sobre Staphylococcus aureus: su aplicación para el diseño de fagos a medida.

Tesis para optar el título de "Doctora en Ciencias Biomédicas"

LIC. SOLEDAD T. CARRASCO
DIRECTOR: HÉCTOR R. MORBIDONI
LABORATORIO DE MICROBIOLOGÍA MOLECULAR, F.C.M, U.N.R.

AGRADECIMIENTOS

Esta Tesis no tendría comienzo si no fuese por mi querido director, el Dr. Morbidoni, quien, hace más de una década, me abrió las puertas del laboratorio para acompañarme en cada etapa de mi formación y, aún más difícil, no dejó que me rindiera cuando la vida no sonreía. Tampoco podría tener un final si no me hubiese presionado un poco cada día para terminar de escribirla y cerrar esta etapa, recordándome que, para comenzar nuevos proyectos, también es necesario aprender a cerrar los anteriores.

Menos aún lo hubiese logrado sin el apoyo de mis compañeros de laboratorio, quienes durante tantos años me aconsejaron, me acompañaron y me brindaron un mate cuando era necesario. Una mención especial para Geo, con quien compartimos frustraciones, esperas y esa sensación de que los resultados siempre tardaban un poco más de lo que queríamos.

A Luciano D'Attilio, quien, sin saberlo, me impulsó hace más de una década a seguir el camino de la bioinformática y que, aún hoy, continúa estando presente de todas las formas posibles. Una de esas personas de las que siempre escuché palabras de aliento y que, de una manera u otra, estuvieron presentes en cada paso.

A mis amigas, Maru, Ari, Pame, Juli, Aye, Guille y Luli. Ellas me sostuvieron en mis peores momentos, conocen mi peor cara y, aun así, siguen a mi lado, compartiendo cada logro y cada batalla. A mis primeras amigas, mis primas, mi familia, Laura y Regi: no existen palabras suficientes para definir el lazo que nos une.

A mis papás, que siempre me impulsaron a perseguir mis sueños y creyeron en mí incluso cuando yo no lo hacía. Y a mis papás postizos, mis padrinos, quienes me regalaron una segunda familia y estuvieron siempre para acompañarme.

A mi hermano y a mi cuñada, uno de los mejores equipos que conozco. Son todo terreno, de esos que siempre tienen la palabra justa en el momento indicado. Son mucho más importantes en mi vida de lo que probablemente imaginan. Además, son los padres de mis tres amores: Bauti, Gapi y Juani.

A Martín, porque si de equipos hablamos, juntos hemos logrado que este funcione lo más parecido a la perfección que se puede. Gracias por darme la confianza que muchas veces no logro encontrar en mí misma y por regalarme el amor más grande de todos: nuestra hija.

Aini, Pepi, Pepina… Cuando pensé que el desafío más grande de mi vida iba a ser realizar un doctorado, llegaste vos para mostrarme que lo más importante en esta vida no es un título, sino las personas que aparecen en todos los renglones de este agradecimiento. En tu caso, ninguna dedicatoria estaría a la

altura. Todavía no se inventaron palabras capaces de describir el amor que siento por vos. Simplemente, si algún día lees esta Tesis, quiero que sepas que es el reflejo de que, aun con el alma rota, luchamos igual. Como podemos. Sin alcanzar la perfección, pero llegando a la meta.

Y quizás eso sea, después de todo, lo que realmente significa terminar esta Tesis: no haberlo hecho todo perfecto, sino haber llegado hasta acá con todos ustedes.

ÍNDICE

Índice de tablas

Índice de figuras

Abreviaturas y Símbolos

- API (Interfaz de Programación de Aplicaciones): conjunto de reglas y protocolos que permite que diferentes sistemas de software se comuniquen entre sí.
- ADN: Ácido desoxirribonucleico.
- ARN: Ácido ribonucleico.
- ARNt (ARN de transferencia): molécula de ARN que participa en la síntesis de proteínas.
- Query: consulta o solicitud específica para obtener datos de una base de datos.
- CBD (endolisinas): Cell wall-binding domain - dominio de unión a la pared celular. Dominio de una enzima cuya función principal es unirse a componentes de la pared celular.
- CD: Coding DNA, región codificante de un gen que se traduce a proteína.
- Datos crudos: información original tal como se ha capturado, que no han sido sujetos al procesamiento o manipulación posterior.

- Disminución Media de la Precisión (Mean Decrease in Accuracy): medida que informa cuánto disminuye la precisión del modelo si eliminamos una variable o permutamos sus valores.
- Disminución Media de Gini (Mean Decrease in Gini): mide cuánto contribuye cada variable a la homogeneidad de los nodos en el árbol de decisión.
- EAD: Enzymatically active Domain – dominio de acción enzimática cuya función principal es catalizar una reacción.
- N/A: del inglés Not available se refiere a los datos que no están disponibles.
- GBF: GeneBank Flatfile
- ORF: Open Reading Frame, es la secuencia comprendida entre un codón de inicio de secuencia y uno de terminación de la misma.
- PCR: Reacción en cadena de la polimerasa, técnica de laboratorio para amplificar pequeñas cantidades de material genético.
- Pipeline: cadena de procesos que se conectan de tal manera que la salida de cada uno es la entrada del siguiente.
- Singletons: grupos conformados por un solo individuo.

RESUMEN

Staphylococcus aureus es un patobionte de humanos y otros mamíferos, capaz de causar un amplio espectro de enfermedades, desde cuadros leves hasta infecciones severas. Actualmente, se encuentra en la lista prioritaria de patógenos de la Organización Mundial de la Salud (OMS) debido al creciente problema de resistencia a los antimicrobianos. Frente a esta amenaza, ha resurgido el interés por los bacteriófagos, enemigos naturales de las bacterias y, por ende, potenciales agentes terapéuticos.

La aplicación de bacteriófagos o sus proteínas en ámbitos como la medicina, la industria alimentaria o la química requiere una caracterización rigurosa de los mismos mediante herramientas bioinformáticas, que permita detectar posibles riesgos -como genes de virulencia- y facilitar su diseño y mejora funcional. Durante el desarrollo de esta tesis, se anotaron y caracterizaron catorce genomas de fagos activos sobre *S. aureus* aislados en el Laboratorio de Microbiología Molecular de la Facultad de Cs. Médicas de la Universidad de Rosario.

Además, se desarrollaron distintos protocolos de trabajo (pipelines) bio-informáticos que permiten automatizar distintos análisis y se implementaron algoritmos específicos para la caracterización y clasificación de dichos genomas y de genes y proteínas particulares. Cada pipeline fue diseñado de forma modular permitiendo integrar información de distintas fuentes, ya sea interna (como la salida de otros análisis) o externa, y posibilitando la selección de manera flexible de los procedimientos a aplicar.

INTRODUCCIÓN

Resistencia a antimicrobianos

La resistencia antimicrobiana (RAM) es la capacidad de los microorganismos de desplegar mecanismos específicos que le permiten sobrevivir frente a sustancias que anteriormente le resultaban nocivas como antibióticos y desinfectantes; esto conlleva un aumento tanto de la morbilidad como de la mortalidad de los infectados como en la propagación de las enfermedades (1). En la actualidad, la resistencia bacteriana a los antibióticos representa uno de los mayores desafíos para la salud pública mundial. Se predice que para el año 2050 se perderán anualmente diez millones de vidas por esta causa (2).

El uso inadecuado y excesivo de antibióticos favorece la selección de bacterias resistentes a los mismos a través de mutaciones genéticas o por la adquisición de genes de resistencia mediante transferencia horizontal. Entre los desafíos que plantea la RAM se encuentran, por un lado, la necesidad de tratamientos clínicos noveles, y por otro, el diseño de desinfectantes eficaces aplicables en sectores diversos con impacto en el ser humano (sector agrícola-ganadero, industria alimentaria y entornos hospitalarios, entre otros). Además, la RAM representa una carga económica significativa debido a tratamientos prolongados y hospitalizaciones recurrentes, lo cual genera un impacto directo sobre la estabilidad de los sistemas de salud, especialmente en países con sistemas públicos de salud (1,3,4).

La gravedad del tema es de tal magnitud que, desde hace años, la Organización Mundial de la Salud (OMS) publica una lista de patógenos prioritarios resistentes a antibióticos, con el fin de fomentar la investigación y desarrollo (I+D) de nuevos componentes antimicrobianos para combatir dicho problema a nivel mundial. Entre los patógenos catalogados con prioridad elevada, tanto por su alta morbilidad como por su mortalidad asociada a la pérdida de eficacia de los tratamientos convencionales, se encuentra *Staphylococcus aureus (S. aureus)* resistente a Meticilina (MRSA), el cual también ha demostrado capacidad de adquirir resistencia a otros agentes antimicrobianos, generando infecciones difíciles de tratar.

Resaltando su relevancia, *S. aureus* se encuentra, además, entre las bacterias agrupadas bajo el término ESKAPE, acrónimo que incluye *a Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa* y *Enterobacter spp.* -bacterias grampositivas y gramnegativas responsables de la mayoría de las infecciones nosocomiales (5).

Staphylococcus aureus

Aspectos generales

Staphylococcus es un género bacteriano compuesto por cocos Gram positivos, catalasa positivos, responsables de múltiples infecciones en humanos y en animales. Son bacterias no móviles, no esporuladas y generalmente carentes de cápsula. *Staphylococcus aureus* es la especie de nuestro interés que, en humanos, se encuentra habitualmente colonizando la piel y las mucosas nasales sin causar enfermedad. Sin embargo, bajo ciertas condiciones, puede tornarse patogénico (por ejemplo, ante un aumento de carga bacteriana, evasión de barreras defensivas o inoculación en tejidos) causando una amplia gama de infecciones, desde eczema cutáneos (de presentación única, recurrente o crónica) hasta infecciones más graves como neumonía, endocarditis o sepsis.

Las infecciones a causa de *S. aureus* se dan en dos contextos: nosocomial y comunitario. En el primer caso, al tratarse de un patógeno intrahospitalario, las infecciones tienden a ser severas, ya que afectan pacientes con comorbilidades, inmunocomprometidos o con heridas expuestas. En el ámbito comunitario, la prevención de contagio es más difícil debido a su diseminación por aerosoles y por contacto persona-persona de portadores asintomáticos (6,7).

Evolución de la resistencia a antimicrobianos

A nivel mundial, el tratamiento de las infecciones por *S. aureus* se basa en la terapia con antibióticos. El descubrimiento de la penicilina en 1928 por Alexander Fleming dio comienzo a la "era de los antibióticos" iniciando así el tratamiento de las infecciones bacterianas. Sin embargo, menos de dos décadas después comenzaron a registrarse cepas resistentes a este antibiótico, marcando el inicio del proceso de resistencia bacteriana. Este fenómeno, aunque continuo, se ha conceptualizado en términos de "olas de resistencia", que definen períodos entre la introducción de un antibiótico y la aparición de resistencia al mismo (Figura 1) (8).

Como se mencionó anteriormente, la primera ola comenzó con la resistencia a la penicilina, conferida por la modificación de los sitios blancos en las bacterias donde esta droga se une para ejercer su acción antibacteriana (PBP, del inglés penicilin-binding protein). Posteriormente, vemos una segunda ola que inicia con la introducción de la meticilina y la aparición de cepas resistentes (MRSA, Methicillin-Resistant *S. aureus*) en los siguientes meses. Dichas cepas no presentan las variantes típicas de PBP sensibles a beta-lactámicos, sino que expresan la proteína PBP2a -una proteína codificada por el gen *mec*A, localizado en un casete cromosomal estafilocócico (SCC*mec*, por sus siglas en inglés) ubicado cerca del origen de la replicación del cromosoma (7).

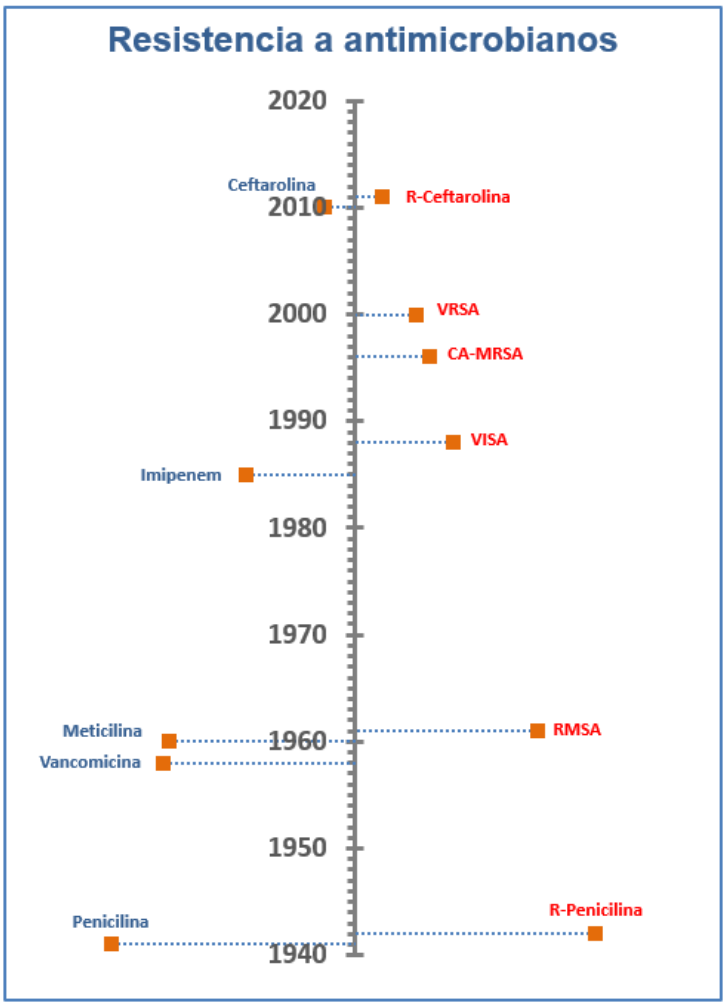

Figura 1: Línea histórica del aumento de resistencia a antimicrobianos en Staphylococcus aureus.

Posteriormente, nuevos clones de MRSA fueron apareciendo seguidos por el uso de vancomicina y su subsiguiente resistencia (VRSA). Este último antibiótico generó una presión selectiva tan intensa que provocó el resurgimiento de cepas MRSA en contextos comunitarios (CA-MRSA), cuando previamente eran exclusivos del entorno hospitalario. Por último, podemos observar como el uso de nuevos antibióticos genera nuevas resistencias.

En bacterias, la resistencia puede surgir por mutaciones genéticas espontáneas, seleccionándose aquellas que otorgan ventajas frente al agente antimicrobiano. No obstante, las formas más estudiadas y de mayor trascendencia evolutiva son las adquiridas a través de la transferencia horizontal de genes de una bacteria resistente a una que no lo es. Entre los elementos genéticos móviles que median esa transferencia se encuentran los profagos, plásmidos, islas de patogenicidad y transposones, los cuales pueden modificar tanto la resistencia a los antibióticos como la virulencia bacteriana (8).

Finalmente, el uso de concentraciones sub-óptimas de antibióticos elimina las bacterias sensibles a la droga administrada permitiendo la supervivencia y multiplicación de las resistentes, lo que acelera la pérdida de la eficacia de estos tratamientos (9).

Factores de virulencia

Los factores de virulencia son aquellos que permiten la invasión y la resistencia a las defensas del huésped. El éxito de *S. aureus* durante la infección depende en gran medida de la regulación coordinada de dichos factores que actúan de manera conjunta a través de la comunicación célula a célula; esta capacidad bacteriana se conoce como percepción de quórum (QS, *quorum-sensing*). Específicamente para esta especie, el sistema se denomina *agr* (por *accessory gene regulator*) y regula la expresión de adhesinas de exotoxinas. Este sistema también es responsable de la producción de biopelículas (*biofilms*): comunidades bacterianas asociadas en una matriz compuesta por polisacáridos, proteínas, ácidos nucleicos y lípidos. Esta matriz, permite la adhesión a distintas superficies bióticas y abióticas, y protege a la comunidad bacteriana frente a los mecanismos inmunes de defensa del hospedador (por ejemplo, la fagocitosis). Las biopelículas también juegan un rol crucial al disminuir el ingreso de los antibióticos al citoplasma bacteriano, favoreciendo así la persistencia bacteriana y la acumulación de mutaciones útiles para el objetivo final de la resistencia (7,10).

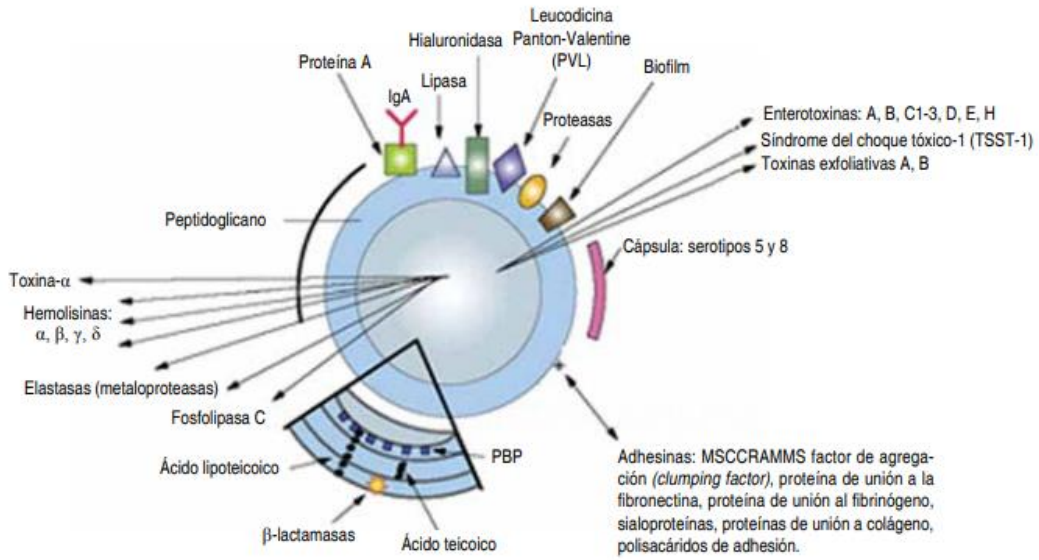

Figura 2: Factores de virulencia en Staphylococcus aureus. Extraído de Cervantes-García, et al.

Como se puede observar en la figura 2, existen múltiples factores de virulencia, algunos de ellos, como las adhesinas y las lipasas, se encuentran como proteínas de superficie con funciones del metabolismo de dicha pared celular o participando en la adhesión al tejido del hospedador facilitando la evasión del sistema inmune. Otros factores de virulencia, como las toxinas, están

codificados en elementos genéticos móviles, como bacteriófagos o islas genómicas, lo cual es clave, ya que la transferencia horizontal de estos elementos es el pilar de la evolución y adaptabilidad bacteriana, especialmente en términos de patogenicidad y competencia (11,12).

En este contexto, los bacteriófagos cumplen un rol central no solo como agentes de transferencia horizontal de genes de virulencia, sino también como moduladores de la evolución bacteriana, influyendo en la adaptación, patogenicidad y diversidad genómica de *S. aureus*.

Bacteriófagos

Necesidad de nuevas estrategias frente a la resistencia antimicrobiana

La necesidad urgente de desarrollar nuevas estrategias terapéuticas ha impulsado la investigación en el uso de bacteriófagos como una alternativa viable y prometedora (13).

Los bacteriófagos, o fagos, poseen la capacidad intrínseca de infectar específicamente a sus huéspedes bacterianos y la ventaja de co-evolucionar con ellos, lo que los convierte en herramientas potencialmente poderosas en la lucha contra infecciones bacterianas resistentes a antibióticos. Sin embargo, su uso terapéutico, o de algunas de sus enzimas (por ejemplo, endolisinas), exige un entendimiento profundo de sus interacciones con las bacterias hospedadoras, así como de su biología y genética. En este contexto, la bioinformática emerge como una disciplina clave, proporcionando las herramientas y métodos para analizar grandes volúmenes de datos genómicos y funcionales de fagos, facilitando su diseño y optimización para aplicaciones terapéuticas (14).

La terapia fágica tiene una historia centenaria, iniciada tras el descubrimiento de los fagos a principios del siglo XX, con los primeros resultados publicados por el bacteriólogo francés Félix d'Hérelle. Sin embargo, su desarrollo se vio eclipsado por el auge de los antibióticos en 1940 luego del descubrimiento de la penicilina. No obstante, en la ex Unión Soviética y en Europa del Este, las terapias fágicas continuaron desarrollándose activamente. Con el aumento de la resistencia antibiótica, estas terapias han resurgido como una opción terapéutica viable. En los últimos años, se ha registrado un renovado interés y un crecimiento significativo en la investigación y aplicación de fagos en la medicina humana, la industria alimentaria y la agricultura (15–17).

En la medicina humana, pese a las largas fases de investigación requeridas para la aprobación de nuevos medicamentos, los fagos ya se han utilizado exitosamente en casos de infecciones bacterianas resistentes a los antibióticos. Por ejemplo, en Georgia, el *Eliava Institute of Bacteriophages, Microbiology and Virology* (IBMV) ha sido pionero en el desarrollo de productos terapéuticos, profilaxis y diagnóstico durante décadas, con numerosos casos de éxito

documentados. Además, en Europa y Estados Unidos, se están llevando a cabo ensayos clínicos para evaluar su eficacia y seguridad frente a distintas infecciones.

Existen diversas patentes registradas para el uso de fagos en tratamientos humanos. Empresas como *Phage Therapeutics* y *Adaptive Phage Therapeutics* han desarrollado fagos específicos para tratar infecciones causadas por *Staphylococcus aureus* y otras bacterias resistentes. Más allá de la medicina humana, los fagos también aplican en la industria alimentaria para prevenir y controlar infecciones bacterianas en productos perecederos, extendiendo así su vida útil y mejorando la seguridad alimentaria. La empresa *Intralytix,* por ejemplo, ha desarrollado fagos empleados en la industria cárnica y aviar para reducir la contaminación por *Listeria*, *Salmonella y Shigella* (18–20).

En agricultura y ganadería, el uso de fagos gana terreno debido a su capacidad para controlar infecciones bacterianas sin dejar residuos químicos ni afectar la calidad del producto. Se han utilizado con éxito para tratar infecciones en cultivos como tomate y papa, y para prevenir infecciones en ganado y aves de corral, reduciendo así el uso de antibióticos y otros químicos (21,22).

Se espera que las terapias fágicas complementen –e incluso reemplacen en algunos casos- a los antibióticos tradicionales, especialmente en el tratamiento de infecciones resistentes. Su combinación con otros antimicrobianos podría mejorar la eficacia terapéutica y reducir la aparición de nuevas resistencias. Un estudio reciente proyecta que, para el año 2030, la terapia fágica podría consolidarse como una opción común para tratar diversas infecciones bacterianas, respaldada por herramientas bioinformáticas avanzadas.

La personalización de las terapias fágicas mediante el uso de fagos específicos para cada paciente es una de las áreas más prometedoras de la medicina de precisión. El análisis genómico del patógeno específico que infecta a un paciente permitirá seleccionar fagos más efectivos, aumentando las tasas de éxito terapéutico. En este escenario, la integración de la bioinformática en la práctica clínica será fundamental, exigiendo el desarrollo de plataformas capaces de analizar de forma rápida y precisa genomas bacterianos y fágicos, facilitando la implementación de estas terapias en hospitales y clínicas (23).

Características generales

Se estima que existen 10^{31} fagos en la Tierra, presentes en todos los nichos ecológicos conocidos. Su abundancia y diversidad hacen que su clasificación dependa de múltiples criterios. Así, cada vez que un fago es aislado, se lo compara con los ya descriptos utilizando un esquema de clasificación múltiple, para luego poder describir las características únicas del individuo en cuestión (12,24).

Los genomas fágicos presentan un alto grado de organización: los genes suelen disponerse en clústeres con función semejante interrumpidos por genes con funciones conservadas entre especies. Además, su información genética está densamente empaquetada, superponiendo genes y evitando largos espacios inter-génicos vacíos (gaps). Si bien fagos de un mismo género comparten clústeres con genes conservados, las variaciones únicas les otorgan características distintivas (25).

Según el tipo de ácido nucleico sea ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN) y el número de hebras (simple [ss] o doble [ds]) los fagos agrupan en cuatro tipos: dsADN, ssADN, dsARN o ssARN. Estas categorías tienen valor descriptivo, pero no taxonómico. Actualmente, existen cuatro criterios principales de clasificación: por ciclo de vida, morfología, taxonomía y supergrupo (26). Sin embargo, debido al avance de las tecnologías de secuenciación y la reducción de sus costos, la clasificación basada en morfología está en declive.

Clasificación según su ciclo de vida

El ciclo de infección de los fagos comienza con la adsorción, una unión específica y reversible, en primera instancia, de los mismos a la célula hospedadora gracias a la interacción de la cola del primero con los receptores de superficie de la última. Este proceso culmina con la absorción irreversible y la inyección del genoma a la bacteria blanco. A partir de allí, se pueden seguir cuatro vías distintas: lítica, lisogénica, pseudolisogénica o crónica (Figura 3).

Los fagos virulentos, líticos estrictos, luego de la infección utilizan la maquinaria genética del hospedador para replicar su material genético y ensamblar nuevas partículas fágicas que son liberadas gracias a la acción de enzimas del fago. Los fagos temperados, en cambio, pueden ingresar al ciclo lisogénico, el cual consiste en convertirse en un profago que puede mantenerse en reposo como un episoma circular o integrarse al cromosoma bacteriano replicándose con el ADN huésped para ser heredado por toda la progenie. Distintos cambios moleculares, desencadenados especialmente por condiciones ambientales como la densidad bacteriana o el estrés nutricional, activan o desactivan la transcripción de genes cuyos productos se encargan de determinar si el profago se mantiene en el ciclo lisogénico o entra en el ciclo lítico produciendo la lisis de la célula huésped (12,26–28).

Estos mecanismos han sido ampliamente estudiados. En el caso del bacteriófago lambda, cuando múltiples copias infectan simultáneamente una misma célula de *Escherichia coli*, se requiere una "decisión unánime" por parte de todos los fagos para establecer la lisogenia. En el fago phi3T en *Bacillus*, se ha identificado un sistema de comunicación basado en péptidos: durante la infección, el fago libera una pequeña molécula señalizadora (“*arbitrium*”), cuya acumulación progresiva tras varios ciclos de infección favorece la entrada al ciclo lisogénico. Por otro lado, en algunos

micobacteriófagos, se ha observado que el gen del represor contiene una etiqueta en su extremo C-terminal que promueve su degradación antes de la integración del genoma; una vez insertado, esta etiqueta se separa y el represor se estabiliza permitiendo mantener la lisogenia (29).

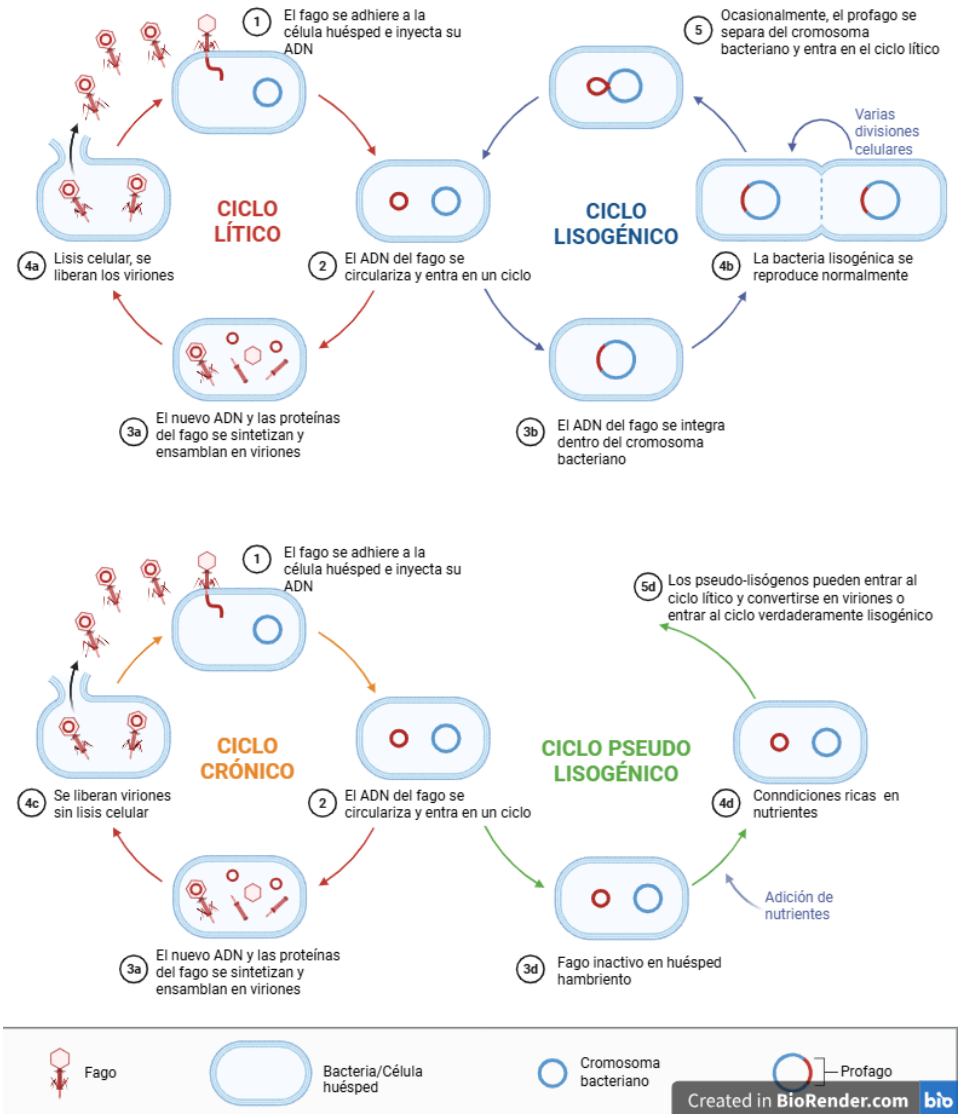

Figura 3: Diagrama de flujo que muestra las interacciones generales entre fagos y huéspedes: (a) ciclo lítico, (b) ciclo crónico, (c) ciclo lisogénico, (d) ciclo pseudo-lisogénico. Adaptado de Chung et. al, 2023 y realizado en https://BioRender.com

Una ventaja de esta clasificación es que no requiere análisis in silico, ya que puede determinarse en el momento del aislamiento observando el tipo de placas (halos) que forman en el hospedador bacteriano sembrado en medio de cultivo sólido. Se describe un tercer tipo de ciclo, el crónico, similar al lisogénico con la diferencia de que puede secretar los viriones por gemación en vez de

causar la lisis de la célula hospedadora, lo que dificulta su detección en placas, ya que, si bien disminuye la tasa de crecimiento de las bacterias, no genera un halo de lisis en la placa. Por último, se describe un ciclo pseudolisogénico, que representa una variante del ciclo lisogénico en la que el fago permanece inactivo dentro de la célula hospedadora, sin integrarse al genoma ni replicarse activamente, lo cual suele ocurrir en condiciones de estrés, como la escasez de nutrientes. (26,30,31).

Es importante remarcar que, entre las cualidades más importantes de los fagos temperados de *S. aureus*, se encuentra su capacidad de movilizar las islas de patogenicidad de esta bacteria (*S. aureus* Pathogenic Islands, SaPIs), fragmentos cromosómicos de aproximadamente 15kb que se integran al genoma y codifican factores de virulencia como superantígenos y toxinas (32,33).

Clasificación morfológica

Esta clasificación se basa en la observación directa mediante microscopía electrónica de transmisión (TEM). De esta forma, los fagos se agrupan en dos grandes tipos: sin cola (pleomórficos, filamentosos o poliédricos) o con cola. Estos últimos, pertenecen al orden de los Caudovirus, el cual incluye a la mayoría de los fagos conocidos y a todos los fagos activos en *S. aureus*. En este orden se reconocen tres familias: (i) los Podovirus con colas cortas no contráctiles, (ii) los Myovirus con colas largas y contráctiles y, por último, (iii) los Siphovirus con colas largas no contráctiles. Existe una correlación entre la morfología y el tamaño de los genomas. Aunque esta clasificación es útil para describir fagos, no es válida para su clasificación taxonómica, la cual ha sido refinada gracias a las técnicas de secuenciación de genomas (26,31).

Clasificación taxonómica

La clasificación taxonómica actual se basa en la homología de las secuencias nucleotídicas y es determinada por el Comité Internacional para la Taxonomía de Virus (ICTV). Según sus criterios, en un alineamiento de comparación por pares, dos fagos pertenecen a la misma especie si comparten al menos el 95% de su genoma y, al mismo género si este porcentaje es del 70% (25). En la siguiente tabla se resumen dichas clasificaciones las cuales exceden los objetivos de esta tesis.

Tabla 1: Taxonomía de bacteriófagos

Clasificación	Tipo ácido nucleico	Tipo hebra	Forma del ácido nucleico	Morfología	Ciclo de vida
Uroviricota	DNA	ds	Lineal	Con cola	Lítico o lisogénico
Kalamavirales	DNA	ds	Lineal	Icosaédrico	Lítico
Autolykiviridae	DNA	ds	Lineal	Icosaédrico	Lítico

Vinavirales	DNA	ds	Circular	Icosaédrico	Lítico
Matshushitaviridae	DNA	ds	Circular	Icosaédrico	Lisogénico
Plasmaviridae	DNA	ds	Circular	Esférico/ pleomorfo	Lisogénico
Finnlakeviridae	DNA	ss	Circular	Icosaédrico	Lítico
Hofneiviricota	DNA	ss	Circular, lineal	Filamentoso	Atípico
Phixiviridae	DNA	ss	Lineal	Icosaédrico	Lítico
Vidaverviricetes	RNA	ds	Tri-segmentado	Icosaédrico	Lítico
Leviviricetes	RNA	ss	Lineal	Icosaédrico	Lítico
Picobirnavirus	RNA	ds	Bisegmentado/unisegmentado	Esférico	Lítico

Clasificación según supergrupos

Los supergrupos, también llamados clústeres o familias, no forman parte de la taxonomía oficial del ICTV, pero su utilizan ampliamente para el análisis funcional y evolutivo. En este esquema genomas sin alta homología nucleotídica y, por ende, cercanía evolutiva, pueden pertenecer a un mismo supergrupo si comparten un conjunto de genes, organización (sintenia) o vías regulatorias comunes (26).

Bacteriófagos activos sobre *Staphylococcus aureus*

Infección de las bacterias blanco

El proceso de infección fágica consta de cinco fases: unión a receptores bacterianos –en primer momento reversible y luego irreversible-, inyección del material genético, replicación y biosíntesis macromolecular, ensamblaje y lisis. Tres de estas etapas son particularmente relevantes, ya que S. *aureus* puede interferir con ellas limitando la acción de los fagos y reduciendo su rango de hospedador (es decir, el número de cepas susceptibles a ser infectadas y eliminadas).

En primer lugar, se encuentra la unión de los fagos a la pared celular de *S. aureus*, una estructura compleja y dinámica, que en estas bacterias Gram-positivas está compuesta principalmente por una capa gruesa de peptidoglicano (20-40 nm de espesor), formado por cadenas de disacáridos (N-acetilglucosamina [GlcNAc] y ácido N-acetilmurámico [MurNAc]) unidas por puentes peptídicos de pentaglicina que conectan residuos específicos entre péptidos. Esta red altamente entrecruzada confiere rigidez y protección mecánica, actuando como barrera contra el estrés osmótico y componentes antimicrobianos (Fig. 4). Es importante señalar que los péptidos precursores del peptidoglicano consisten en L-alanina, D-isoglutamina, L-lisina con una pentaglicina unida al grupo amino épsilon y una D-alanina-D-alanina terminal cuya síntesis es blanco de algunas terapias antibióticas (34,35).

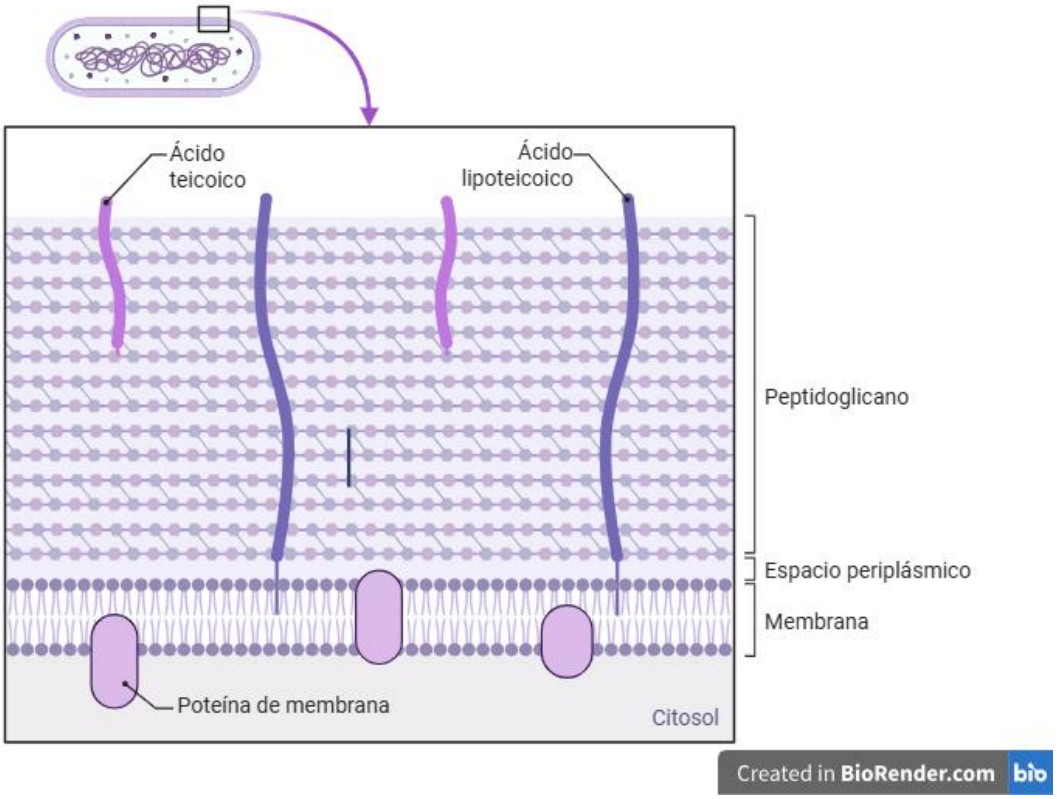

Figura 4: Estructura de la envoltura celular de estafilococos.

Integrados en esta matriz se encuentran los ácidos teicoicos (WTA) de tipo ribitol, polímeros aniónicos [poly-RboP] unidos covalentemente a residuos de ácido N-acetilmurámico mediante enlaces fosfato de glicerol$_3$-N-acetilmanosaminil β (1→4) GlcNAc. Estos ácidos teicoicos forman una matriz superficial esencial para la morfogénesis celular, la adhesión a tejidos y la interacción con componentes del sistema inmune. También actúan como zonas de anclaje para proteínas de superficie y receptores de fagos, donde la variabilidad en el grado de entrecruzamiento del peptidoglicano y la disponibilidad de puentes pentaglicina no unidos podría modular la accesibilidad de los fagos a sus receptores (33).

Entonces, el primer paso en el proceso de infección es la unión de los fagos a estos receptores de la pared celular externa de *S. aureus*, pudiendo ser también el primer motivo de resistencia, ya sea porque los receptores estén ausentes, bloqueados o no sean capaces de reconocer al fago. Se han identificado mutaciones en la síntesis de los ácidos teicoicos de pared que afectan tanto el polímero poly-RboP como la cadena de GlcNAc en los monómeros de ribitol, explicando fenómenos de resistencia al reconocimiento fágico (36).

En la fase de replicación y biosíntesis, existen tres mecanismos por los cuales se puede perturbar la infección. Uno es la inmunidad a la superinfección: un fago integrado en el cromosoma bacteriano

(profago) produce una proteína (represor) que impide que el mismo active su escisión del cromosoma y se replique llevando a la lisis celular. Este estado, conocido como lisogenia, beneficia a la bacteria (al adquirir genes fágicos) y al profago (que puede ser movilizado desde el cromosoma a otras células de *S. aureus*). El represor producido puede bloquear eficazmente la replicación de los fagos entrantes si están relacionados genéticamente con el profago (37–39). El segundo proceso consiste en la degradación del ADN del fago antes de que este pueda replicarse e iniciar el ciclo lítico o lisogénico mediante alguno de los sistemas de restricción/modificación (sistemas R-M). Este sistema permite a las bacterias reconocer secuencias cortas de su ADN modificado por enzimas metilantes; también, contiene enzimas con acción endonucleasa que reconocen la misma secuencia, pero no modificada por metilación; es decir, si el ADN es extraño (fago o plásmido) la endonucleasa lo corta iniciando su degradación. Por último, se encuentran los sistemas CRISPR (repeticiones palindrómicas cortas agrupadas y regularmente espaciadas) que incorporan de forma adaptativa segmentos cortos del ADN de los fagos que ya han destruido en el genoma bacteriano (espaciadores) que luego se expresan y procesan generando crRNAs maduros que forman complejos con proteínas Cas (enzimas con función endonucleasa) que reconocen y se unen a secuencias complementarias del ADN del fago invasor y lo cortan para así neutralizarlo.

La última barrera descripta se ubica en el momento de ensamblaje, las islas de patogenicidad de *S. aureus* (SaPI) parasitan fagos auxiliares mediante interferencia en su ensamblaje: remodelan cápsides haciéndolas más pequeñas (y por ende específicas para el tamaño de la SaPI), inhiben el empaquetamiento del ADN fágico mediante proteínas de interferencia (Ppi) y bloquean genes tardíos del fago, propagando así factores de virulencia sin lisar la célula hospedadora, en un escenario de co-evolución constante (33,40).

Características genómicas

Los bacteriófagos presentan una organización genómica compacta, con escasas y pequeñas regiones intergénicas, lo que da lugar, en el caso de fagos de *S. aureus*, a una densidad génica alta que ronda en promedio los 1,67 genes/kb. Sin embargo, la mayoría de estos genes hipotéticos tienen funciones desconocidas, y muchos no presentan dominios funcionales reportados (11). Otra característica de estos fagos es que su contenido de guanina y citosina (%GC) es similar al de su hospedador de (~ 33%) (7).

Como se mencionó anteriormente, los fagos de *S. aureus* pertenecen al orden de los Caudovirus. Morfológicamente, pueden diferenciarse según la clasificación propuesta por Ackermann (41) en tres familias, las cuales tienen íntima relación con su tamaño genómico y su ciclo de vida. La mayoría son fagos temperados pertenecientes a la familia de Siphovirus, con cabeza icosaédrica y

cola no contráctil finalizada en una estructura de base de plato, y un tamaño genómico de aproximadamente 40 kbp. Los Podovirus tienen genomas más pequeños (menor a 20 kbp) y colas cortas, no flexibles ni contráctiles, siendo los menos frecuentemente descriptos. En cambio, los Myovirus presentan colas contráctiles y genomas mayores a 125 kb. (42). Dentro de esta última familia se encuentran los denominados fagos "jumbo" cuyos genomas superan las 200 kbp debido a la presencia de abundantes intrones (43).

Cada familia tiene una organización modular característica, donde los genes se agrupan según funciones relacionadas (Fig. 5). Estos genes tienden a transferirse juntos durante los eventos de recombinación o transferencia horizontal, lo que refuerza la importancia de estudiar estos genomas también a nivel modular (44). Existen cuatro módulos comunes a todas las familias: (i) metabolismo del ADN, (ii) empaquetamiento del ADN y morfogénesis de cápside, (iii) morfogénesis de cola y (iv) lisis. En los Siphovirus, además, hay un quinto módulo de lisogenia presente únicamente en los fagos temperados. Entre dicho módulo y el de lisis suelen localizarse los factores de virulencia, agrupados formando clústeres

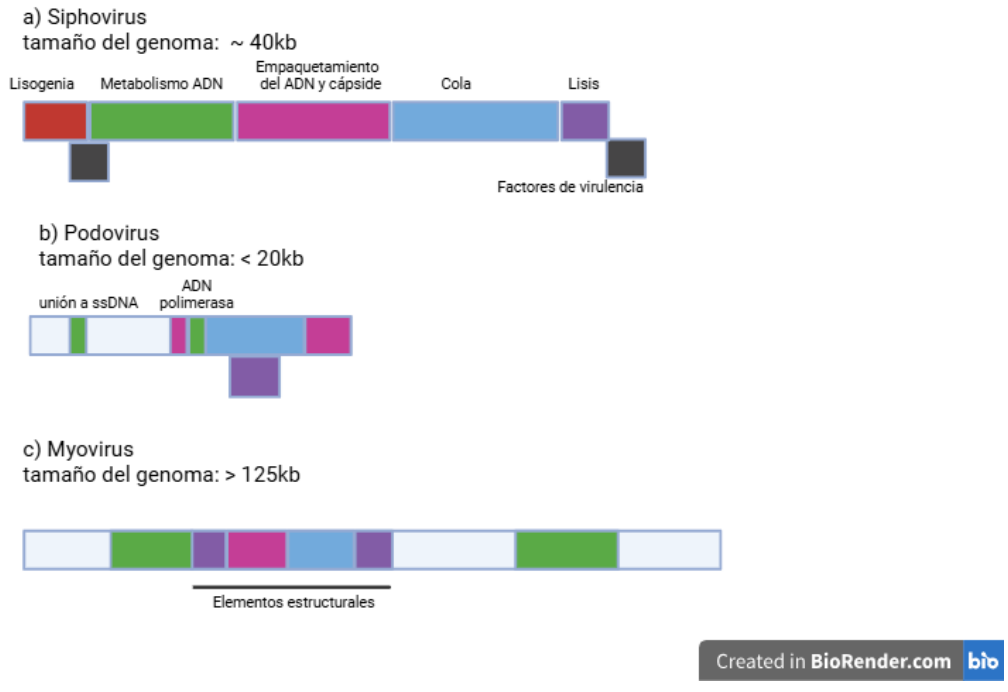

Figura 5: Organización modular de los genomas de los fagos de estafilococo. Adaptado de Deghorain et. al, 2012.

Dos características importantes de los genomas fágicos son: (i) la alta proporción de genes sin función asignada, que pueden estar dispersos por todo el genoma -como en el caso de los

Siphovirus- o agrupados en regiones específicas, como en las restantes familias de fagos (Fig. 5 en color gris); y (ii) la tendencia a maximizar el espacio codificante mediante superposición de genes. Estas particularidades abren nuevas posibilidades de investigación bioinformática, como la asignación de funciones a proteínas hipotéticas o la identificación de genes codificantes en regiones intergénicas, incluidos aquellos que producen péptidos cortos de aún no caracterizados.

Casete de lisogenia

Este módulo, presente exclusivamente en Siphovirus, es el responsable de la evolución bacteriana a corto plazo debido a que permite la integración en el genoma bacteriano y la transferencia de factores de virulencia. En *S. aureus,* destacan: la leucocidina de Panton-Valentine (*lukSF-PV*), la toxina exfoliativa A (*eta*), la proteína anclada a la pared celular (*SasX*) y el grupo de evasión inmunitaria (IEC) compuesto por la enterotoxina S (*sea*), la estafiloquinasa (*sak*), la proteína inhibidora de la quimiotaxis (*chp*) y el inhibidor del complemento estafilocócico (*scn*). Además, estos fagos temperados pueden alterar la expresión de genes del hospedador y protegerlos frente a infecciones de otros fagos (28,45).

La integración de los fagos al genoma bacteriano (Fig. 6) se da gracias a enzimas altamente específicas que catalizan la recombinación de forma unidireccional en sitios específicos. Estas proteínas, denominadas integrasas, también participan de la escisión del profago para luego continuar con el ciclo lítico, proceso que requiere proteínas accesorias como la excisionasas (*Xis*) o los factores de direccionalidad de recombinación (RDF) (46).

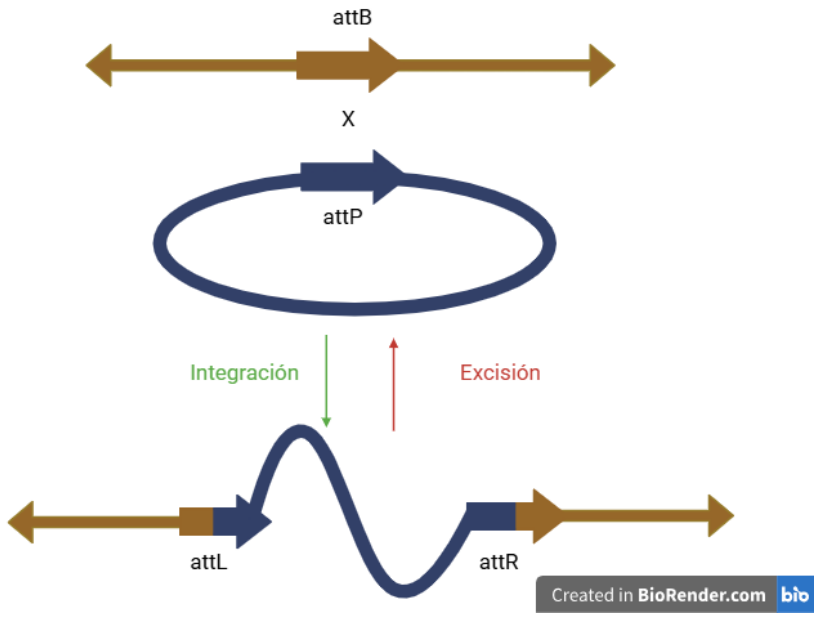

Figura 6: Sitios de integración y escisión de fagos.

Las integrasas se clasifican, en primera instancia, según el aminoácido en sus sitios catalíticos: en integrasas de tirosina (T-int) o serina (S-int) siendo las primeras las más frecuentes en fagos (45). Otra clasificación, propuesta por Goerke y su equipo (47), agrupa las integrasas en doce grupos (Sa1int a Sa12int) según su polimorfismo génico. Existe una asociación entre el tipo de integrasa, los factores de virulencia que introducen y los sitios de integración, aunque también se han reportado eventos donde la S-int3 se integra a un sitio de unión ilegítimo.

Una última relación corresponde a este módulo y el de lisis, donde se ha descripto una correlación entre los genes que codifican holinas y endolisinas, lo que sugiere un beneficio evolutivo al intercambiar ambos módulos en conjunto (36,45).

Metabolismo del ADN

La ubicación del módulo de metabolismo del ADN, así como su posible subdivisión en dos submódulos, varía entre familias fágicas (Fig. 6). Los productos génicos presentes en este módulo tienen funciones de replicación y de regulación del material genético incluyendo enzimas como ADN polimerasas, primasas, proteínas de unión al ADN monocatenario y helicasas (11,48).

Los mecanismos moleculares de la replicación en fagos (es decir, la duplicación de su material genético) han sido ampliamente estudiados y, en muchos casos, sirven como modelos para comprender procesos equivalentes en organismos más complejos, como bacterias y eucariotas.

En organismos con genomas de ADN de doble cadena (dsDNA), esta molécula forma una estructura helicoidal antiparalela, con las bases nitrogenadas orientadas hacia el centro de la hélice. Para que ocurra la replicación, es necesario separar las hebras, un proceso catalizado generalmente por helicasas que consumen ATP. A continuación, las ADN polimerasas sintetizan nuevas hebras utilizando las cadenas originales como molde, siempre en dirección 5’ a 3.

Debido a la orientación antiparalela, solo una hebra (la hebra líder) puede copiarse de manera continua. La otra, denominada hebra rezagada, se sintetiza de manera discontinua mediante la formación de fragmentos cortos conocidos como fragmentos de Okazaki.

Tanto las helicasas como las polimerasas son componentes centrales del replisoma: un complejo multi-proteico que coordina el desenrollamiento y la síntesis del ADN. Sin embargo, algunos estudios han revelado que ciertos fagos pueden replicar su ADN en ausencia de helicasas, lo que indica la existencia de mecanismos alternativos de replicación (49,50).

Empaquetamiento del ADN y morfogénesis de cápside

El proceso de ensamblaje de las cápsides en los Caudovirus comienza en el vértice portal mediante la co-polimerización de las proteínas de andamiaje y la proteína principal de la cápside, lo que

conduce a la formación de los precursores de la misma llamados pro-cápside. Estas estructuras están compuestas por la proteína portal, el núcleo interno de andamiaje y la cubierta externa de proteína principal de la cápsula que rodea a este núcleo. En muchos fagos, existe además una proteína de andamiaje secundaria localizada en el núcleo interno, que asegura una geometría adecuada durante la construcción de la cápside.

Las pro-cápsides de numerosos fagos contienen proteasas de maduración de la cabeza, activadas tras el ensamblaje, que degradan parcial o completamente el núcleo interno, liberando los productos de dicha digestión para dar lugar al ADN genómico. En otros casos, este núcleo proteico no se degrada y la proteína de andamiaje es expulsada intacta de la cápsula durante el empaquetado del ADN, probablemente empujada por el ADN entrante y saliendo a través de aberturas en la cubierta de la pro-cápside (51,52).

El genoma es empaquetado en la pro-cápside mediante translocación activa de ADN a través del vértice portal, impulsada por la hidrólisis de ATP, en un proceso mediado por el complejo de la terminasa, formado por dos subunidades: (i) TerL (subunidad larga), con dominios funcionales ATPasa y nucleasa responsables de la unión y translocación del ADN a la pro-cápside, y (ii) TerS (subunidad pequeña), encargada de reconocer el sitio de inicio del empaquetamiento y estimular la actividad ATPasa TerL. Durante este proceso, la pro-cápside sufre cambios estructurales para convertirse en la cápsula madura. En muchos casos, los sitios donde comienza y finaliza el proceso de empaquetamiento se denominan sitios *cos* y *pac* respectivamente (53,54).

Sin embargo, no todos los fagos utilizan estas señales específicas, ya que existen al menos seis estrategias diferentes de empaquetamiento del ADN, cada una asociada con tipos particulares de extremos genómicos y mecanismos de corte, siendo las cinco restantes: (i) extremos cohesivos (*cos ends*), fragmentos de cadena sencilla en 5’ o 3’ que se pueden reunirse tras la inyección del ADN; (ii) repeticiones terminales directas circularmente permutadas donde el empaquetamiento es en serie; (iii) repeticiones terminales exactas (cortas o largas); (iv) secuencias terminales derivadas del ADN del hospedador; (v) proteínas unidas covalentemente a los extremos del ADN (54). La naturaleza de la subunidad TerL y sus dominios funcionales permite predecir el tipo de estrategia de empaquetamiento que es empleada por cada fago, remarcando la importancia de los análisis bioinformáticos (55).

Durante la etapa final de maduración, las cápsulas de muchos fagos adhieren proteínas de estabilización y/o decoración en su superficie externa. Finalmente, tras el empaquetamiento del genoma, el motor de translocación de ADN se disocia del vértice portal y se adhieren las proteínas de finalización de la cápside, sellando la puerta del portal y proporcionando un núcleo para el

ensamblaje de la cola, en los podofagos, o una plataforma para la unión de la cola pre-ensamblada en los sipho- y myofagos (51,52).

Morfogénesis de cola

La cola de los fagos es especialmente importante ya que permite reconocer al hospedador, penetrar su barrera celular y transportar el ADN al citoplasma. La unión de los fagos a las moléculas de la superficie celular de las bacterias formada por ácidos teicoicos (WTA), marca el inicio del ciclo de infección siendo un proceso extremadamente específico mediado por las proteínas de unión a los receptores (RBPs), proteínas ubicadas en el extremo más distal de la cola.

No todos los receptores presentes en la superficie de las bacterias son candidatos para unirse a los fagos, deben ser: estructuralmente estables (para permitir la adaptación evolutiva del fago a los distintos cambios en la apariencia del receptor), físicamente accesibles y estar presentes en cantidades suficientes para que se pueda producir el contacto entre la RBP y el receptor con alta frecuencia. En el caso de *S. aureus,* todas estas características las ofrecen los ácidos teicoicos de la pared celular. La especificidad entre distintas variantes de los ácidos teicoicos y RBPs modela la transferencia de distintos factores de virulencia y mecanismos de adaptación de las bacterias que, en última instancia, diferencian los distintos clones bacterianos (45).

Las principales proteínas que componen este casete se muestran en la figura 7. Existen diferencias en las proteínas que podemos hallar en las tres familias que le otorgan la morfología característica a cada una de ellas. En el caso de los Siphovirus, las proteínas de vaina de la cola que brindan la capacidad contráctil a los Myovirus están ausentes. Los Podovirus, si bien tienen una organización similar a estas dos familias a pesar de tener un tamaño menor, muestran distintas proteínas debido a que el ensamble de la cola a la cápside se realiza por una vía distinta a las otras familias (52). Las características particulares de cada clase determinan también el mecanismo de unión al hospedador (56). En concordancia con esto, los fagos líticos son los reportados como de mayor rango de hospedador, actividad anti-biofilm y mayor efectividad para combatir las infecciones (36).

La complementariedad entre RBPs y los receptores bacterianos determina el rango de hospedador de los fagos; por lo tanto, su estudio es fundamental al planificar estrategias de uso de fagos en terapias contra cepas resistentes a antimicrobianos. Dependiendo del caso, pueden diseñarse estrategias con fagos polivalentes –fagos capaces de infectar dos o más géneros bacterianos-, con aplicación en el caso de las co-infecciones, o utilizar fagos monovalentes, específicos para un solo género (31,57).

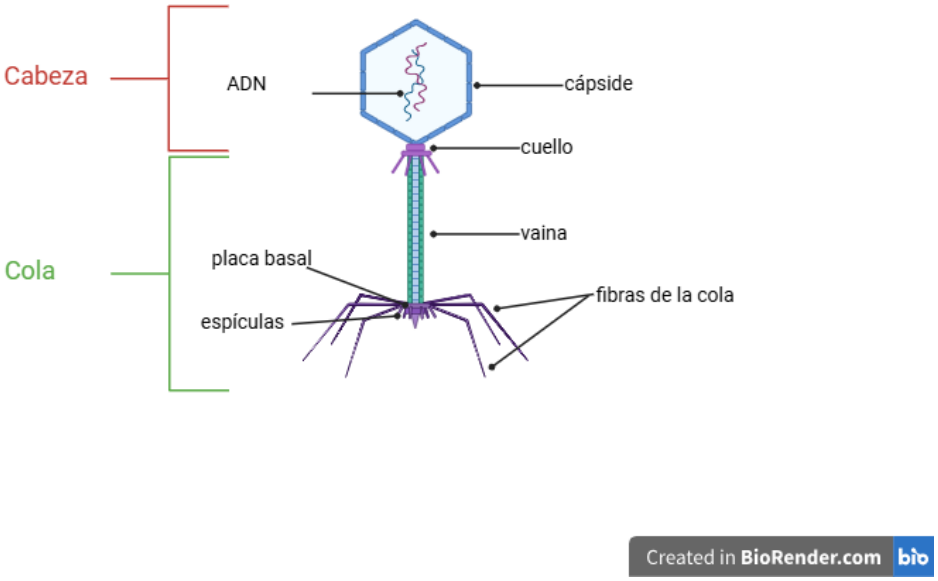

Figura 7: Representación esquemática de fagos Myovirides

Casete de lisis

El casete de lisis es el responsable de la destrucción de la integridad de la envoltura del hospedador por los fagos y la liberación de los nuevos viriones. Se caracteriza por dos tipos de proteínas que se encuentran en genes cercanos: holinas y endolisinas. Las primeras, son las encargadas de formar un poro en la membrana celular permitiendo que el peptidoglicano este accesible para las segundas (57,58).

Las endolisinas son una clase de enzimas pertenecientes a las lisinas, encargadas de unirse específicamente al peptidoglicano de la pared celular de la bacteria huésped comprometiendo su integridad y causando la lisis osmótica de la célula con el fin de liberar la progenie de fagos. Aunque, su función natural es desde el interior de la célula, también pueden realizar la lisis de afuera hacia adentro en bacterias Gram positivas, lo que, sumado a su especificidad a nivel del género, especie y/o cepa las convierte en candidatos prometedores para su uso como nuevos antimicrobianos. Por consiguiente, su análisis mediante herramientas bioinformáticas resulta imprescindible a la hora de comparar endolisinas o combinar sus dominios (58,59).

El sistema holina-endolisina permite la lisis bacteriana mediante una coordinación temporal y precisa luego de la replicación viral. Las holinas se acumulan como depósitos citoplasmáticos hasta alcanzar un umbral crítico, momento en que se insertan en la membrana citoplasmática produciendo poros y la disipación de potencial de membrana permitiendo el pasaje de las

endolisinas desde el citoplasma a la cara interna de la pared celular. Estas endolisinas alcanzan al peptidoglicano e hidrolizan enlaces específicos del mismo, debilitando la pared y provocando la explosión celular por presión osmótica interna (60) (Fig. 8).

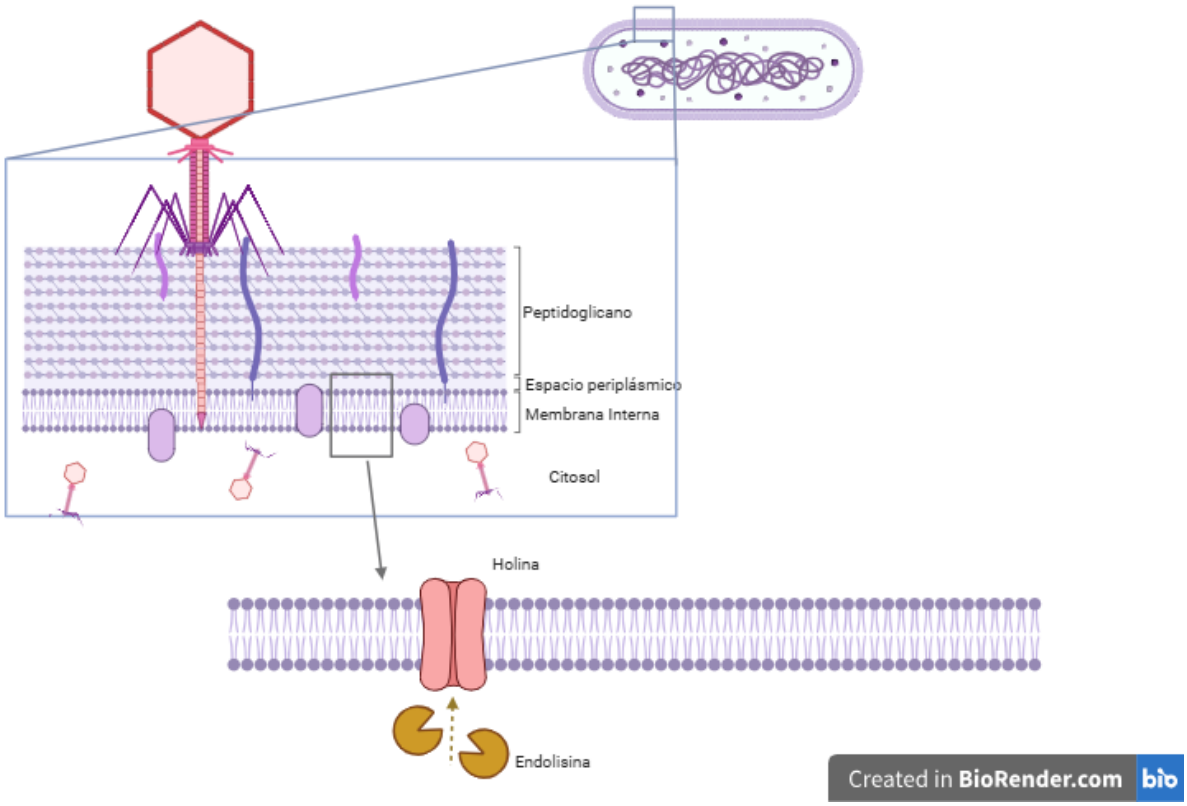

Figura 8: Representación esquemática del sistema lítico holina-endolisina

La clasificación de endolisinas se realiza en base a su arquitectura de dominios: típicamente poseen un dominio de reconocimiento y unión a la célula (CBD, por su traducción del inglés de Cell Binding Domain) -que confiere alta especificidad a nivel de género e inclusive de especie- y uno o más dominios catalíticos (EAD, enzymatically active domain) responsable de la lisis celular mediante su acción sobre los diversos enlaces químicos del peptidoglicano (61,62). Estos dominios se separan por un conector flexible compuesto por aminoácidos con carga y polares (63).

Se ha observado una correlación entre los CBDs y los EADs que lleva a suponer que existe una co-evolución de los mismos maximizando la eficiencia lítica (64). Hasta el momento, las endolisinas de fagos activos en *S. aureus* se clasifican en seis grupos distintos dependiendo de los dominios funcionales que poseen (Tabla 2) (65).

Tabla 2: Clasificación de endolisinas de Staphylococcus aureus basada en la composición de dominios funcionales

Grupo	Dominios funcionales	Dominios Pfam
Tipo_I	EAD (CHAP) – CBD (sin homología 1)	PF05257
Tipo_II	EAD (CHAP) – CBD (SH3)	PF05257 – PF08460
Tipo_III	EAD (CHAP – amidasa2) – CBD (SH3)	PF05257 – PF01510 – PF08460
Tipo_IV	EAD (CHAP – amidasa3) – CBD (SH3)	PF05257 – PF01520 – PF08460
Tipo_V	EAD (CHAP – amidasa3) – CBD (sin homología2)	PF05257 – PF01520
Tipo_VI	EAD (peptidasa_M23, amidasa2) – CBD (SH3)	PF01551 – PF01510 – PF08460

Estado del Arte: análisis bioinformáticos de fagos

Aumento masivo de secuencias de fagos

En los últimos años, el desarrollo de tecnologías de secuenciación de nueva generación (NGS) ha revolucionado la capacidad para secuenciar genomas completos de manera rápida y económica permitiendo a los investigadores identificar y caracterizar una vasta cantidad de fagos y proporcionar información crucial sobre su diversidad genética, estructura genómica y funciones biológicas. Gracias a esto y a nuevas herramientas para detectar profagos en genomas bacterianos, se ha producido un notable incremento en la disponibilidad de secuencias genómicas de fagos en bases de datos públicas. GenBank (66), PhageDB (https://phagesdb.org/) y otras plataformas especializadas contienen actualmente miles de secuencias de fagos, proporcionando una valiosa fuente de información para el análisis y diseño de nuevas terapias fágicas (29).

Este crecimiento de datos ha hecho que los análisis bioinformáticos sean cada vez más imprescindibles. La bioinformática, entendiéndola como la disciplina que busca soluciones a problemas biológicos mediante el uso de tecnologías computacionales, permite gestionar, analizar e interpretar grandes volúmenes de datos, facilitando la identificación de genes clave, la comparación de secuencias y el estudio de las interacciones fago-hospedador.

Por otra parte, dado que los fagos líticos no solo son menos frecuentes sino también presentan genomas de mayor tamaño que los temperados (>= 150 kb. vs. 50 kb.), una estrategia explorada consiste en eliminar las regiones que codifican los genes para la lisogenia en los fagos temperados, transformándolos en fagos líticos. Este desafío de detectar regiones prescindibles del fago puede ser abordado convenientemente gracias a la bioinformática.

Adicionalmente, la disponibilidad masiva de secuencias fágicas ha abierto nuevas posibilidades en el campo de la biología sintética, permitiendo el diseño racional de fagos modificados con propiedades mejoradas para aplicaciones industriales o terapéuticas. A partir del conocimiento

detallado del repertorio genético, es posible identificar y ensamblar módulos funcionales -como genes de lisis, adhesinas o inhibidores de mecanismos de defensa bacteriana- para producir fagos sintéticos personalizados contra cepas específicas de *S. aureus* (67,68). La ingeniería de fagos representa un enfoque prometedor para superar las limitaciones de los fagos naturales y optimizar su eficacia y seguridad en aplicaciones clínicas (69).

Análisis bioinformáticos para fagos

El crecimiento exponencial del número de genomas fágicos disponibles en bases de datos públicas ha transformado el estudio de bacteriófagos en un campo fuertemente impulsado por la bioinformática. Previamente al auge de las tecnologías de secuenciación masiva y los enfoques computacionales, la mayoría de los métodos descriptivos y de clasificación que se utilizaban se basaban en resultados obtenidos mediante la reacción de la cadena polimerasa (PCR) o en características morfológicas determinadas por microscopia electrónica.

En lo últimos años, se ha observado un avance significativo en el desarrollo y la democratización de técnicas de *machine learning* lo que ha ampliado el acceso a análisis computacionales complejos sin requerir necesariamente conocimientos avanzados de programación. No obstante, en el estudio de fagos persisten metodologías fundamentales resultan indispensables y que son susceptibles de ser automatizadas en pipelines integrales, tales como la anotación genómica, la descripción de características básicas y la clasificación de proteínas claves como integrasas y endolisinas.

Si bien los enfoques comparativos clásicos basados en alineamientos genómicos completos y redes de similitud proteica han demostrado ser robustos, su aplicación se vuelve cada vez más costosa y difícil de escalar a medida que aumenta la cantidad de genomas disponibles. En este contexto, enfoques alternativos basados en la arquitectura funcional de los genomas, el análisis modular y el uso de técnicas de aprendizaje automático emergen como herramientas prometedoras para la clasificación, caracterización y predicción funcional de nuevos fagos.

Análisis de Secuencias Genómicas

Una vez aislado, secuenciado y ensamblado el fago de interés el primer paso para su caracterización es la anotación de su genoma. Este proceso comprende la predicción de los marcos abiertos de lectura (ORFs por *Open Reading Frames*) -anotación estructural- y la asignación de su función hipotética, -anotación funcional-, así como la predicción de ARN, promotores y terminadores (70). La anotación puede realizarse de forma automática y/o manual. Hasta el momento, la anotación manual sigue siendo indispensable, dado que los métodos automáticos, pese a que los algoritmos han evolucionado radicalmente en los últimos años, aún presentan falsos positivos o negativos que deben corregirse. Para estas tareas se utilizan herramientas como Artemis (71) para facilitar la

visualización de los genomas; BLAST (72), para comparar secuencias genómicas de fagos con bases de datos existentes facilitando la identificación de genes y regiones funcionales; e InterPro (73), que integra distintos recursos en un solo sitio para la clasificación de proteínas en familias y para predecir los dominios y motivos funcionales.

Tradicionalmente, las anotaciones automáticas se realizaron utilizando herramientas bioinformáticas diseñadas para genomas bacterianos, como GeneMark (74,75), Glimmer (76) y Prodigal (77), causando una baja performance en la identificación presuntiva de los genes y una pobre anotación de las funciones de los mismos pese a haber lanzado en la última década versiones adaptadas a genomas fágicos. Otros programas, como DNA Master (78), PHAST (79) y PROKKA (80), integran uno o varios de estos predictores en sus flujos de trabajo, pero comparten el mismo enfoque de homología a genes bacterianos. Es importante señalar que la automatización de las anotaciones en procariotas es posible gracias a la co-linealidad del ADN que –salvo pocas excepciones- permite que los transcriptos de ARN sean un espejo de la secuencia de ADN del que fueron transcriptos.

En 2019 se lanzó PHANOTATE (81), que introdujo un nuevo algoritmo diseñado exclusivamente para fagos mejorando así las predicciones. A diferencia de otros programas, PHANOTATE no requiere entrenamiento previo con genes conservados ni perfiles de homología, lo que lo convierte en una herramienta libre de referencias. Esto es especialmente útil en fagos, cuyos genes suelen mostrar baja similitud con secuencias conocidas.

En sus evaluaciones, PHANOTATE fue comparado con otras herramientas ampliamente utilizadas como GeneMarkS, Glimmer y Prodigal. En un análisis realizado sobre más de 2.000 genomas de fagos completos disponibles en GenBank, PHANOTATE mostró un alto grado de coincidencia con los genes anotados por estas herramientas (alrededor del 82%). También fue capaz de predecir aproximadamente un 6% más de ORFs que los otros programas. Muchos de estos ORFs adicionales mostraron evidencia de funcionalidad, ya sea por su similitud con proteínas conocidas en bases de datos o por su presencia recurrente en estudios metagenómicos. Además, está disponible como software de código abierto, fácil de instalar, y permite exportar resultados en diversos formatos estándar como GenBank, GFF o FASTA, lo cual facilita su integración con otros flujos de trabajo.

Para finalizar esta etapa, distintas métricas descriptivas son utilizadas para la caracterización de nuevos genomas de fagos como por ejemplo el contenido de guanina y citosina y la cantidad de secuencias codificantes (CDs).

Se estima que entre el 60% y el 95% de las secuencias virales carecen de una función asignada, constituyendo la “materia negra” viral. Su estudio podría revelar nuevas proteínas con potencial

anti-bacteriano (82). Dentro de este conjunto destacan los péptidos cortos –secuencias de menos de 50 aminoácidos-, las cuales han cobrado relevancia en los últimos años por su capacidad para modular el ciclo de vida del fago, la expresión génica bacteriana y ejercer funciones regulatorias e incluso terapéuticas (83–85). Actualmente, gracias a los avances en inteligencia artificial (IA), específicamente en redes neuronales profundas y todo lo necesario para su funcionamiento, este tópico está siendo abordado mediante Modelos de Lenguaje Grande (LLMs) (86)

Análisis filogenéticos y evolutivos

Los análisis filogenéticos son esenciales para estudiar las relaciones evolutivas entre diferentes fagos y sus hospedadores. En fagos temperados, una misma bacteria puede contener varios profagos, favoreciendo la adquisición de factores de adaptación (45).

Programas como MEGA (87) y RAxML (88) permiten construir árboles filogenéticos basados en secuencias completas, proporcionando información sobre la evolución y diversificación de los fagos. Estos estudios son cruciales para identificar patrones de co-evolución y adaptación entre fagos y bacterias, favoreciendo el diseño de terapias fágicas más efectivas.

Hasta el momento se utilizaron diferentes enfoques para clasificar a los fagos, no solo a través de las distancias en árboles filogenéticos de los genomas completos sino bajo el estudio de módulos o proteínas específicas. Por ejemplo, se estudiaron los genes del empaquetamiento de ADN y los de cola mediante la caracterización de los sitios *cos* y *pac*. También, se ha estudiado la relación de polimorfismos de genes de distintos módulos como los que codifican las integrasas, holinas, endolisinas, represores y factores de virulencia con el fin de mejorar la clasificación y comprensión de la diversidad fágica (45,47,89).

Estudios recientes han enfatizado el rol emergente de la terapia fágica frente a infecciones causadas por bacterias multirresistentes y han examinado su seguridad, eficacia, y barreras regulatorias en contextos clínicos actuales. Por ejemplo, Subramanian (17) discute estrategias de formulación, sinergias con antibióticos y desafíos de implementación clínica de fagos como agentes terapéuticos frente a patógenos resistentes. Además, revisiones contemporáneas destacan que, a pesar de los avances, la evidencia clínica sigue siendo limitada, lo cual subraya la importancia de enfoques integrativos que combinen herramientas bioinformáticas y biología fágica para avanzar hacia aplicaciones seguras y eficaces (14,90,91).

Hipótesis de trabajo

Se plantea como hipótesis que el proceso de anotación, caracterización y análisis bioinformático de genomas de bacteriófagos puede automatizarse mediante la integración de herramientas

bioinformáticas y algoritmos específicos dentro de un pipeline computacional. Dicha automatización permitiría obtener resultados comparables a los alcanzados mediante análisis manuales, reduciendo significativamente el tiempo de procesamiento, minimizando la intervención del operador y favoreciendo la estandarización y reproducibilidad de los análisis.

OBJETIVOS

Objetivo General

Staphylococcus aureus es un patógeno prioritario a nivel global por su creciente resistencia a antibióticos convencionales, lo que ha renovado el interés en los bacteriófagos como alternativas terapéuticas y herramientas biotecnológicas.

En este contexto, el objetivo general de esta tesis es automatizar e implementar nuevas estrategias para la anotación, caracterización y análisis evolutivos de los genomas de fagos activos sobre *S. aureus* aislados en el Laboratorio de Microbiología Molecular (Fac. de Cs. Médicas, UNR) donde, desde hace más de una década, se trabaja con fagos líticos y temperados. Este abordaje busca identificar regiones genómicas susceptibles de modificación o reemplazo para el desarrollo de biosensores y/o para incrementar el rango de hospedador.

Objetivos específicos

a) Desarrollar herramientas bioinformáticas para el laboratorio que permitan la anotación genómica de los fagos secuenciados, su clasificación y agrupamiento.
b) Identificar genes no esenciales para la replicación de fagos temperados (tales como módulos de lisogenia) y genes no deseables (codificantes) para su eliminación en la construcción sintética de biosensores.
c) Realizar la búsqueda in silico en nuestros fagos de homólogos a RBP reportadas, tanto para fagos de la familia Siphovirus, Myovirus como Podovirus. Analizar variaciones de secuencia y presencia de dominios y correlacionar estos datos con los de rango de hospedador de dichos fagos.

MATERIALES Y MÉTODOS

Métodos computacionales

En este apartado se describen las herramientas y enfoques computacionales empleados en la anotación y caracterización de genomas fágicos, así como las técnicas de aprendizaje automático (*Machine Learning*, ML) utilizadas para su clasificación.

Bases de datos biológicas

Las bases de datos biológicas son repositorios organizados que almacenan, gestionan y permiten el acceso a información relacionada con las ciencias de la vida, como secuencias nucleotídicas, secuencias y estructuras de proteínas, metadatos experimentales y literatura científica relacionada. Estas bases de datos pueden ser locales, alojadas en los servidores de una institución, o remotas, accesibles en la nube a través de plataformas mantenidas por organizaciones internacionales.

Existen bases de datos especializadas por dominio taxonómico -como reino, género o especie- así como bases de datos generalistas que abarcan una gran variedad de organismos. La elección de una base de datos depende del tipo de información requerida, pero también de factores como su frecuencia de actualización, la calidad de los datos y el respaldo institucional. Idealmente, se priorizan aquellas mantenidas por entidades de prestigio como el National Center for Biotechnology Information (NCBI) (92) y el European Bioinformatics Institute (EBI) (93).

En el estudio de bacteriófagos, la calidad y curación de las bases de datos resulta particularmente crítica debido a la alta proporción de proteínas hipotéticas y a la naturaleza “en mosaico” de sus genomas.

Entre las bases de datos más relevantes destacan:

- GenBank (NCBI): una de las bases de datos más utilizadas y completas para secuencias nucleotídicas. Contiene datos de secuencias de genes, genomas completos, elementos regulatorios y anotaciones asociadas. GenBank (66) es parte del consorcio International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org) (94), lo que garantiza interoperabilidad con otras bases como el European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/) y el DNA DataBank of Japan (DDBJ, https://www.ddbj.nig.ac.jp/).
- InterPro (EBI): una base de datos que integra información de múltiples fuentes para identificar dominios, familias y sitios funcionales en proteínas. Utiliza modelos de alineamiento y perfiles de secuencia para ofrecer anotaciones funcionales precisas y estandarizadas (73).

- Protein Data Bank (PDB): base de datos central para el almacenamiento y consulta de estructuras tridimensionales de proteínas, ácidos nucleicos y complejos macromoleculares. Incluye datos experimentales obtenidos por cristalografía de rayos X, RMN y microscopía electrónica, siendo esencial para estudios estructurales y de diseño de fármacos (95,96).

Alineamiento de secuencias

Los métodos de alineamiento de secuencias se basan en algoritmos diseñados para comparar secuencias de ADN, ARN o proteínas con el objeto de identificar similitudes que indiquen relaciones evolutivas, funcionales o estructurales. Estos análisis consideran eventos como mutaciones, inserciones, deleciones y reordenamientos bajo ciertos modelos de penalización y puntuación. Son herramientas fundamentales en los estudios biológicos in silico, ya que constituyen la base para otros enfoques analíticos como los análisis filogenéticos, la identificación de dominios y la detección de motivos conservados (97).

Existen distintos modelos matemáticos para realizar alineamientos, entre los que se destacan el alineamiento global, que compara secuencias en toda su extensión (por ejemplo, con el algoritmo de Needleman-Wunsch), y el alineamiento local, que identifica regiones de alta similitud dentro de secuencias más extensas (como el algoritmo de Smith-Waterman). De acuerdo al número de secuencias a alinear, estos métodos se clasifican en alineamientos por pares o múltiples (MSA) que permiten comparar dos secuencias o varias secuencias en simultáneo, respectivamente, revelando patrones de conservación evolutiva o funcional (98,99).

Dado que los algoritmos exactos, como Needleman-Wunsch y Smith-Waterman, son computacionalmente costosos, especialmente en el contexto de grandes bases de datos, se han desarrollado métodos heurísticos que permiten realizar alineamientos de forma mucho más rápida, aunque con una leve pérdida de precisión. Estos métodos no exploran todas las posibles soluciones, sino que utilizan estrategias aproximadas para encontrar alineamientos óptimos o cercanos al óptimo en tiempos razonables. Herramientas como BLAST y FASTA son ejemplos clásicos de enfoques heurísticos ampliamente utilizados en bioinformática para búsquedas rápidas y eficientes en grandes conjuntos de secuencias (100).

Entre las herramientas más utilizadas se encuentran BLAST (para alineamientos locales rápidos), Clustal Omega y MUSCLE (para alineamientos múltiples), y MAFFT, que ofrece una alta precisión en el alineamiento de grandes conjuntos de secuencias. Para aplicaciones estructurales, herramientas como T-Coffee y HHPred permiten integrar datos funcionales y estructurales en los alineamientos (97,101,102).

ClustalW implementa una estrategia de alineamiento progresivo basada en un árbol guía, incorporando esquemas de ponderación de secuencias y matrices de sustitución específicas para proteínas, lo que lo hace especialmente adecuado para conjuntos de secuencias con niveles moderados a altos de homología. Además, ha sido extensamente utilizado en estudios filogenéticos y comparativos, lo que facilita la comparación con trabajos previos y garantiza resultados robustos y reproducibles. MAFFT suele ofrecer un mejor rendimiento computacional y mayor precisión en conjuntos de secuencias grandes o altamente divergentes, sin embargo, su implementación no pudo ser ejecutada en el entorno de R utilizado en este trabajo debido a limitaciones de compatibilidad y dependencias del sistema.

Anotación genómica

La anotación genómica es el paso posterior al ensamblado del genoma y consiste en interpretar biológicamente la secuencia obtenida. Este proceso incluye la predicción de genes, la localización de regiones codificantes (CDS), y la asignación de funciones probables a través del reconocimiento de dominios funcionales, motivos conservados y similitudes con secuencias previamente caracterizadas.

La predicción de genes se llevó a cabo utilizando PHANOTATE (81), que introduce un enfoque innovador para la anotación de genes en genomas fágicos, basado en una representación del genoma como un grafo dirigido y ponderado. Resumidamente, en esta estructura los nodos corresponden a inicios y terminaciones de codón, mientras que las aristas representan posibles marcos abiertos de lectura (ORFs), así como los solapamientos y espacios que pueden ocurrir entre ellos. A cada tipo de arista se le asigna un peso o penalización, que refleja su plausibilidad biológica. Estos pesos se calculan considerando factores como la longitud del ORF, la probabilidad de que no aparezca un codón de parada (“stop codon”) en esa región y la penalización adicional si el ORF implica un cambio de hebra.

El modelo de puntuación utilizado por PHANOTATE tiene en cuenta diferentes elementos. Para los ORFs, se calcula una probabilidad de no encontrar un codón de parada en la secuencia, que se basa en la frecuencia observada de los codones en el genoma analizado. Este valor se ajusta además mediante un análisis del contenido de GC en los tres marcos de lectura, lo que permite detectar cuál de los marcos es más probable que sea funcional. En el caso de las regiones sin codificación entre ORFs, la penalización depende de su longitud y de si ocurre un cambio de hebra. Para los solapamientos (“overlaps”), se calcula un valor combinado que incluye la media de los pesos de los ORFs implicados y una penalización adicional si hay cambio de hebra.

Una vez construida esta representación del genoma como grafo, PHANOTATE aplica el algoritmo de Bellman-Ford, que permite encontrar el camino con la menor suma total de penalizaciones a lo largo del genoma. El resultado es un conjunto óptimo de ORFs que maximiza la probabilidad de representar

genes reales en el genoma, aún en casos donde estos se solapan o están anidados, como es común en fagos. La asignación de funciones hipotéticas se realizó mediante la búsqueda de dominios Pfam (103) en la base de datos de InterProScan y la anotación según el consenso tal como se explica en el apartado Descarga y procesamiento de archivos asociados.

La predicción de ARNt (Ácidos Ribonucleicos de transferencia) se llevó a cabo mediante el programa tRNAscan-SE (104) (versión 2.0.9), programa que emplea modelos de covarianza para capturar la información de las secuencias primarias y las estructuras secundarias en los datos de entrenamiento para luego encontrar los ARNt en las secuencias consulta.

Técnicas de Machine Learning

Las técnicas de aprendizaje automático empleadas en este trabajo corresponden a enfoques supervisados, en los cuales los modelos son entrenados a partir de datos previamente etiquetados. Estas metodologías permiten identificar patrones complejos en datos biológicos de alta dimensionalidad, como secuencias de proteínas o perfiles funcionales genómicos.

Clasificación a través de SVM

El clasificador de Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) es un método de aprendizaje supervisado utilizado principalmente para problemas de clasificación y regresión. La idea central de un SVM es encontrar el hiperplano que mejor separa las clases en un espacio de características, maximizando el margen entre las clases más cercanas (vectores de soporte) y el hiperplano. Se habla de aprendizaje supervisado cuando un modelo es entrenado utilizando un conjunto de datos etiquetados.

Cuando se trabaja con secuencias de aminoácidos, el principal desafío es que los SVM tradicionales requieren vectores numéricos como entradas, pero las secuencias biológicas son cadenas de caracteres. Aquí es donde entra en juego el paquete "kebabs", que está diseñado para manejar secuencias biológicas como datos de entrada en un SVM. Este paquete permite representar secuencias de aminoácidos utilizando kernels específicos para secuencias. Un kernel es una función que mide la similitud entre dos secuencias, permitiendo así que el SVM opere en el espacio de características implícito sin necesidad de convertir directamente las secuencias en vectores de características. En este trabajo, uno de los kernels utilizados para el clasificador de endolisinas, fue el Gappy Pair Kernel que considera pares de k-mers con un cierto número de posiciones intermedias (gaps) entre ellos.

Una vez entrenado el modelo, se pueden clasificar nuevas secuencias de aminoácidos utilizando el modelo SVM entrenado. El modelo calculará la similitud entre las nuevas secuencias y las secuencias

de entrenamiento a través del kernel, y luego determinará en qué lado del hiperplano se encuentran las secuencias nuevas para hacer la predicción de clase.

Clasificación a través de kNN

El clasificador *k*-Nearest Neighbors (kNN) es un método de aprendizaje supervisado no paramétrico que asigna una clase a una observación desconocida en función de las clases de sus *k* vecinos más cercanos en el espacio de características. La similitud entre observaciones se define mediante una métrica de distancia, de modo que una nueva instancia se clasifica según la clase mayoritaria entre sus vecinos más próximos. Debido a su simplicidad conceptual y a que no requiere un entrenamiento explícito del modelo, kNN resulta especialmente adecuado para conjuntos de datos de tamaño moderado y con relaciones no lineales, como ocurre en problemas de clasificación biológica basados en medidas de similitud entre secuencias. Este método fue utilizado de manera exploratoria como referencia comparativa frente a enfoques más complejos.

Random Forest (árboles de decisión)

Random Forest es un método de aprendizaje supervisado basado en conjuntos de árboles de decisión ampliamente utilizado tanto para problemas de clasificación como de regresión. El algoritmo construye un gran número de árboles de decisión de manera independiente, utilizando diferentes subconjuntos aleatorios de los datos de entrenamiento y de las variables predictoras.

Durante el entrenamiento, cada árbol se genera a partir de una muestra aleatoria del conjunto de datos y en cada nodo de división se considera solo un subconjunto aleatorio de variables. Esta estrategia introduce diversidad entre los árboles y reduce el sobreajuste característico de los modelos basados en un único árbol de decisión. La predicción final del modelo se obtiene mediante un esquema de votación mayoritaria entre los árboles individuales.

Una ventaja central de Random Forest es su robustez frente a datos ruidosos y de alta dimensionalidad, así como su capacidad para manejar conjuntos de datos desbalanceados cuando se incorporan pesos de clase. Además, el método permite estimar de manera interna el desempeño predictivo mediante el error out-of-bag (OOB), sin necesidad de un conjunto de validación independiente.

Otro aspecto relevante es la posibilidad de evaluar la importancia relativa de las variables, comúnmente estimada a través de métricas como el Mean Decrease in Accuracy y el Mean Decrease in Gini. Estas medidas permiten identificar qué características contribuyen de manera más significativa a la capacidad predictiva del modelo, proporcionando una interpretación biológica adicional en estudios genómicos.

Debido a estas propiedades, Random Forest resulta especialmente adecuado para el análisis de datos genómicos complejos, donde las relaciones entre variables pueden ser no lineales y altamente interdependientes. Este método fue empleado para la clasificación de fagos en subclusters a partir de matrices binarias de presencia/ausencia de dominios funcionales.

Colección de datos

Los datos para la realización de esta tesis provinieron tanto de secuencias genómicas propias como de aquellas alojadas en bases de datos públicas y su información asociada.

Bacteriófagos secuenciados en el laboratorio

Se utilizaron 14 secuencias de genomas de bacteriófagos temperados activos contra *S. aureus* aislados por miembros del Laboratorio de Microbiología Molecular. El método de aislamiento para todos los fagos fue mediante la inducción con mitomicina C de cepas de *S. aureus* provenientes de muestras humanas y veterinarias, de distintos lugares de nuestro país, (ver Tabla 3) según la metodología descripta en "Bioinformatic Analysis of a set of 14 temperate bacteriophages isolated from *Staphylococcus aureus* strains highlights their massive genetic diversity"(105).

Tabla 3: Origen de los aislados

Bacteriófago	Fuente	N° de fagos obtenidos / N° de muestras analizadas	Método de aislamiento	Ubicación geográfica
277g, 287, 308, 320, 321c, 690, 713, 760, 775, 832	Aislados humanos de manos de portadores sanos (carniceros)	37/75	Inducción con mitomicina C	La Plata, Bs.As., Argentina
I73	Aislados de vacas con mastitis	4/23		Rafaela, Santa Fe, Argentina
Mh1, Mh_4, Mh15	Aislados clínicos (hemocultivos, muestras blandas de infecciones tisulares	10/18		Rosario, Santa Fe, Argentina

Las secuencias fueron secuenciadas en INDEAR (Instituto de Agrobiotecnología de Rosario) mediante un equipo Hiseq/Illumina y ensamblados en el mismo lugar utilizando el pipeline A5 (106). El extremo izquierdo de los mismos fue establecido en el gen que codifica la integrasa con el fin de permitir comparaciones colineales entre genomas.

Los genomas fueron renombrados según las directrices de nomenclatura propuestas por Kropinski et al. (107) luego de analizar el tamaño genómico a través de electroforesis en gel de campo pulsado (PFGE) y de confirmar la existencia del gen de la integrasa en los mismos. Asi, los nombres fueron armados como "vB_SauS_**x**" donde **x** es el nombre original del bacteriófago.

Estas secuencias ensambladas previamente por otros miembros del laboratorio fueron los materiales de inicio para este trabajo.

Creación de base de datos local de bacteriófagos de *Staphylococcus spp*.

Extracción, transformación y limpieza de datos

Se creó una base de datos local con las secuencias de bacteriófagos aislados en *Staphylococcus spp.* disponibles en NCBI en la colección de nucleótidos y los metadatos de las mismas (Fecha de actualización: 26/02/2025). En primer lugar, se realizó la búsqueda por términos para recuperar todos los elementos que fuesen bacteriófagos de *Staphylococcus spp*. y correspondan a la especie virus evitando así todas las secuencias de bacterias. El término de búsqueda fue formulado de forma tal que permita la devolución del mayor número de secuencias completas posibles:

```
Search_term = '(Staphylococcus[Organism] OR Staphylococcus[Title]) AND phage[All Fields] AND (complete genome [All Fields] OR complete sequence [All Fields]) AND viruses[filter]'
```

Es importante aclarar que la búsqueda se realizó mediante la automatización del proceso utilizando la librería "`rentrez`" que permite la conexión a NCBI a través de una API (Interfaz de Programación de Aplicaciones) y aloja los resultados en una lista que posee los identificadores GenInfo Identifier (Gi) permitiendo luego, con los mismos, la búsqueda de toda la información asociados mediante el entrecruzamiento de bases de datos de NCBI. De esta forma, se obtiene una tabla con los datos crudos que luego sufrirá un proceso de limpieza y transformación cuyo destino final es un archivo en formato Excel.

El proceso de limpieza de datos consiste en la aplicación de distintos filtros para que, en última instancia, la base de datos este formada por ejemplares únicos. El término de búsqueda utilizado aseguró que todos los individuos pertenezcan al género *Staphylococcus* y sean genomas completos. Sin embargo, algunos eran plásmidos y otros eran organismos sin verificar (Título = "UNVERIFIED_ORG"). Por esta razón, se procede al primer filtrado eliminando todos individuos no deseables. En una segunda instancia se desagregan las columnas "SubType" y "SubName" creando nuevas columnas cuyos nombres de las variables son aquellas de "SubType" y cuyo registro es el dato

de “SubName” en orden de aparición. Es importante aclarar que no todos los registros tienen el mismo número de variables, pero, al hacer esta transformación se igualan las mismas para todos los individuos con campos “N/A” (del inglés not available) en los casos que no había información.

Luego, se eliminaron duplicados (igual título u organismo) priorizando aquellos que estaban en la base de datos de Reference Sequence (RefSeq)(108) sobre los que estaban en International Nucleotide Sequence Database Collaboration (INSDC)(109); esta elección se basa en que RefSeq es una base de datos con conjuntos de secuencias completas, no redundantes y bien anotadas: las secuencias primero se suben a INSDC y, una vez revisadas, pasan a RefSeq con distinto Gi impidiendo el rastreo por el mismo identificador, por lo tanto, no es posible detectar duplicados por Gi.

También, se transformaron los datos para homogeneizar las fuentes de donde fueron aislados y se agregó una nueva variable que indique de que especie se trata para luego filtrar según el análisis lo requiera. Aquellas especies que tienen menos de 20 individuos en la base de datos se agruparon bajo el término ‘otro’.

Los metadatos, además de guardarse en un archivo Excel como se mencionó anteriormente, fueron guardados en un RDATA junto con los Gi únicos para *Staphylococcus aureus* y no *aureus* para facilitar trabajos posteriores en el entorno de R.

Hasta este punto todos los procesos se realizan de forma automática y están orientados a bacteriófagos de *Staphylococcus spp.* con posterior discriminación por especie. A continuación, se realizó el análisis de la composición de la base de datos mediante la realización de un tablero utilizando Microsoft Power BI (110), herramienta que permite la inspección dinámica de los datos y permite compartir informes a distintos usuarios. Así, la base de datos sufrió una última curación manual.

Se comparó mediante BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) a aquellos fagos que poseían idéntica longitud de secuencia asumiendo que la probabilidad de que al azar dos secuencias posean la misma longitud y sean distintas es baja; si poseían una cobertura alta y un alto porcentaje de identidad (>95% en ambos casos) fueron asumidos como fagos similares y, por lo tanto, solo una de las secuencias permaneció en la base de datos. Otro filtro fue realizado por la longitud, los fagos tienden a tener similar longitud según la familia a la que pertenecen, cualquier valor atípico lleva a suponer que hay algún error en la secuencia y, por ende, a revisarla (ver Gi de fagos eliminados en Anexo Tabla 15). Por último, debido a que el momento de la actualización de la base de datos los genomas trabajados en esta tesis ya fueron depositados en la base de datos del NCBI, se procedió a eliminar los mismos. De esta forma, los Gi que se quieren eliminar se guardan en un Excel que ingresa nuevamente a R para eliminar las secuencias no deseables al inicio de la próxima etapa.

Descarga y procesamiento de archivos asociados

Las secuencias genómicas se encuentran típicamente disponibles en dos formatos principales: FASTA y GBF. Mientras que el primero contiene únicamente la secuencia nucleotídica, los archivos GBF incluyen, además, la información de la anotación asociada a cada gen. Ambos formatos resultan necesarios para los análisis posteriores desarrollados en este trabajo.

Por este motivo, se realizó una nueva consulta a la base de datos de nucleótidos del NCBI para extraer los archivos GBF correspondientes a cada fago. Para esta tarea se utilizó nuevamente la librería "`rentrez`" en una función que recibe como entrada ("input") cada identificador Gi de la tabla de metadatos y devuelve los archivos GBF, los cuales fueron almacenados en una carpeta específica.

A partir de estos archivos, se extrajeron las secuencias nucleotídicas de los genomas utilizando la función "`gb2fasta`" de la librería "`seqinr`", generándose archivos individuales en formato FASTA. Asimismo, se construyó un archivo multifasta que reúne todas las secuencias genómicas de la base de datos.

Dado que varios de los análisis realizados en este trabajo se desarrollan a nivel proteico, fue necesario extraer todas las secuencias de proteínas codificadas por las regiones CDS presentes en los archivos GBF. Para esto, se utilizó el paquete "`read.gb`", que permite recuperar la información asociada a cada CDS (posición genómica, producto, identificadores, entre otros atributos), la cual fue posteriormente organizada en una tabla para su limpieza y curación.

Con el objetivo de lograr un consenso en la anotación de las funciones de dichas proteínas se extrajeron las secuencias únicas y se transformaron en una matriz utilizando herramientas del ecosistema "`Bioconductor`", obteniendo el formato requerido para utilizar los servicios de InterProScan mediante su API empleando el paquete "`httr`". La identificación de dominios funcionales se realizó considerando las bases de datos Pfam, GENE3D (111), SUPERFAMILY (112) o NCBIFAM, priorizando un único dominio por proteína y siguiendo el orden de aparición mencionado.

En función de los dominios funcionales detectados y de las anotaciones previas, se generó una anotación consensuada siguiendo las normativas propuestas por SEA-PHAGES (https://seaphages.org/). Entre las cuales recomiendan no usar abreviaturas o unifican las anotaciones por ejemplo llamar "tyrosine integrase" a todas las "tyrosine homologous recombinase" o "T-Int". Una vez finalizado este proceso, la tabla definitiva con todas las secuencias proteicas asociadas a los fagos de la base de datos fue incorporada al archivo RDATA generado previamente, y las secuencias fueron almacenadas adicionalmente en formato FASTA para su uso en análisis posteriores.

Anotación de genomas, caracterización y clasificación

En este apartado se describen los procesos necesarios para anotar, caracterizar y clasificar los catorce fagos aislados en el Laboratorio de Microbiología Molecular que fueron previamente publicados (105).

El análisis incluye procedimientos considerados esenciales en el estudio genómico de fagos, tales como la anotación de los genomas, la descripción de características generales (%GC, cantidad de CDS, etc.), la categorización según el clúster genómico y el análisis de genes particulares. Todos los análisis se dividieron en dos enfoques metodológicos complementarios:

1. Método 1: es un enfoque basado en múltiples tareas manuales realizadas por el operador y el uso de diversos softwares independientes. Los resultados fueron publicados en "Bioinformatic Analysis of a Set of 14 Temperate Bacteriophages Isolated from *Staphylococcus aureus* Strains Highlights Their Massive Genetic Diversity" (105) y se utilizaron como referencia para evaluar la concordancia y veracidad de los resultados obtenidos mediante el pipeline semi-automático. Aquellos análisis realizados exclusivamente por otros miembros del laboratorio serán debidamente indicados.
2. Método 2: corresponde el desarrollo e implementación de un pipeline semi-automático que integra los análisis necesarios para la anotación y clasificación genómica de bacteriófagos. Este pipeline fue diseñado de manera tal que, ante la incorporación de nuevas secuencias genómicas de fagos activos contra *S. aureus*, estas puedan ser alojadas en un directorio específico y procesadas de forma encadenada y reproducible (Fig.9).

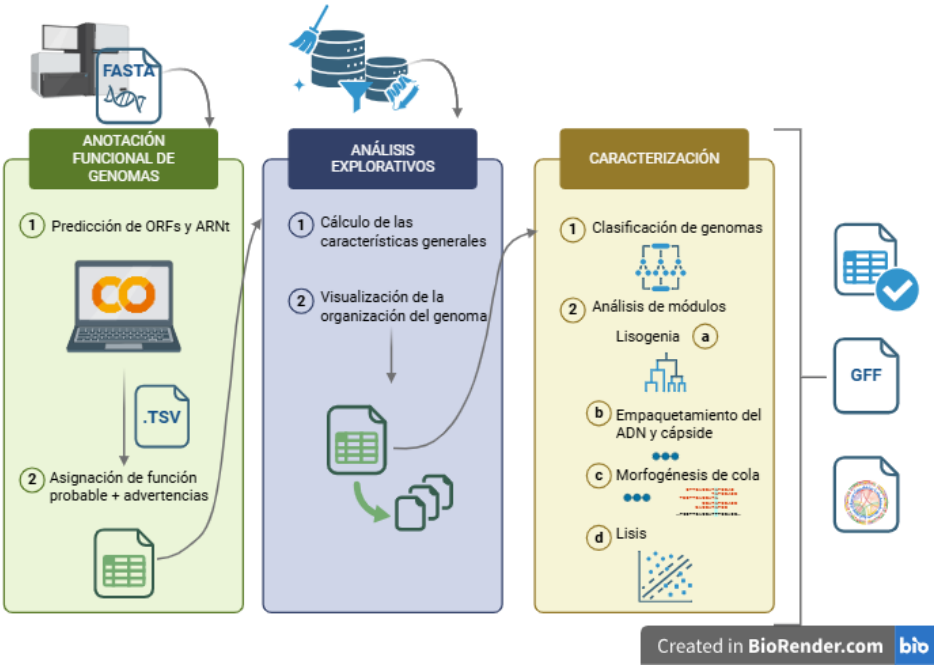

Figura 9: Flujo de trabajo del pipeline semi-automático

Es importante remarcar que, si bien el flujo de trabajo contempla distintas etapas en las que podría realizarse curación manual, en esta tesis no se efectuaron dichas curaciones, con el objetivo de comparar ambos métodos con la mayor fidelidad posible.

Procesamiento de datos propios: anotación de genomas

Método 1:

Para la anotación de los genomas se utilizó DNAMaster (74)(http://Phagesdb.Org/DNAMaster/) mediante el cual se realizó la predicción de los marcos abiertos de lectura (ORFs) utilizando Glimmer (versión 3.02) (76) y GeneMarkS (versión 4.28) (74) las cuales luego fueron curadas manualmente. Para la asignación de funciones probables se utilizó BLASTP internamente desde DNAMaster, y externamente se utilizaron HHPred (113), Pfam (114), y tRNAscan-SE (104) para la búsqueda de ARNt. Por otra parte, se utilizó IslandViewer (115) para la predicción de islas genómicas. Las características de los genomas fueron calculadas en R Software (versión 3.6.3) (ver descripción en Método 2). Los mapas genómicos fueron creados en la misma plataforma utilizando el paquete `genoPlotR` (versión 0.8.11). Por último se detectaron los factores de virulencia con VFanalyzer (116).

Método 2:

Predicción de ORFs

```
Entrada: secuencias genómicas en formato FASTA.
Salida: archivos en formato .TSV con los datos de cada ORF predicho.
```

Para la predicción de marcos abiertos de lectura (ORFs) se utilizó Google Colaboratory (https://colab.research.google.com/) donde se instaló y ejecutó PHANOTATE, un programa desarrollado en Python diseñado específicamente para genomas fágicos (81).

El primer paso fue integrar Google Drive a la notebook de Colab, alojando allí una carpeta con todos los archivos FASTA. Luego, estos archivos se procesan uno por uno generando dos archivos de salida en formato TSV para cada secuencia, que son almacenados en una carpeta separada y, al finalizar el proceso, se descarga a una carpeta local.

El primer archivo TSV contiene para cada ORF: la posición de inicio y fin de cada ORF, si la cadena es la codificante (+) o la complementaria (-), y el puntaje asignado por el programa. El segundo archivo, que será utilizado como entrada en análisis posteriores, incluye los encabezados ajustados y agrega las secuencias nucleotídicas y proteicas correspondientes a cada ORF.

Dicho script también contiene las sentencias necesarias para correr tRNAScan-SE. Los archivos de entrada son los mismos que se utilizan para la predicción de ORFs con PHANOTATE, pero la salida es

un solo archivo TSV con los ARNt encontrados en todos los fagos que son descargados automáticamente al finalizar la corrida. Si no se encuentran ARNt no se descarga ningún archivo.

Tanto el alojamiento de los archivos FASTA en Google Drive como la ejecución del script se realizan manualmente. El último paso manual, posterior a la descarga automática de los archivos, es descomprimir la carpeta que aloja los archivos TSV de PHANOTATE y trasladarla, junto con el archivo de salida con la predicción de ARNt (de encontrarse) al entorno local de R Studio para continuar con los siguientes análisis.

Asignación de función probable

```
Entrada: archivos en formato TSV con los datos de cada ORF predicho.
Salida: libro de Excel (archivo .xlsx) con las anotaciones con
advertencias.
```

Se procesó cada archivo con las predicciones de los ORFs para generar una tabla única con la información de origen más otras variables como los cálculos de longitud de los ORFs, overlaps o gaps. Para la asignación de las funciones probables se utilizó el servidor de InterProScan de la misma forma que se menciona en el apartado Descarga y procesamiento de archivos asociados homologando la nomenclatura de las funciones.

Se siguieron los siguientes principios para crear advertencias y que faciliten la decisión de eliminar o no algún gen o no:

- Overlaps mayores a 30 pb.
- Gaps mayores a 100 pb.
- Longitud de genes menor a 150 pb.
- Cambios de orientación de solo un gen desde la hebra líder a rezagada (forward to reverse) o viceversa.
- Al menos 50 pb entre los sitios de inicio de dos genes contiguos transcriptos en direcciones opuestas para dejar lugar a los promotores.
- Los genes que codifican proteínas deber finalizar con los codones de finalización “TAG”, “TGA” o “TAA”.
- Los genes que codifican proteínas deber iniciar con los codones “ATG”, “GTG” o “TTG”.

El archivo con las anotaciones y las advertencias se guardan en formato .xlsx para su revisión y modificación manual de corresponder.

Análisis exploratorios de las secuencias

Características generales y organización de los genomas

```
Entrada: archivo .xlsx.
Salida: (1) archivo .xlsx con las características generales, (2) archivo
.png con las imágenes de los genomas, (3) archivo .GFF con las anotaciones
finales.
```

El archivo de ingreso es el Excel de salida del punto anterior. Se realiza la limpieza de los datos eliminando automáticamente solamente los genes que tienen errores en los codones de inicio o finalización y/o aquellos con longitudes demasiado cortas de 50 pb.

En el caso de las secuencias que poseían intrones se preservó dicha anotación y, además, se guardó otra solo con el exón para poder realizar los análisis a nivel de proteínas.

Las características generales de los genomas fueron resumidas y exportadas a un archivo `.xlsx`. Para cada fago se describió su tamaño genómico, %GC, número de genes y de CDs, la densidad genómica, la longitud promedio de los CDs y el porcentaje codificante del genoma. A continuación, se citan los cálculos accesorios:

1. Porcentaje de G+C:

$$\%GC = \frac{G + C}{A + T + G + C} \times 100$$

2. Porcentaje de secuencia codificante:

$$\%Codificante = \frac{\sum_{i=1}^{n} LCDs_i}{LG} \times 100$$

3. Densidad génica (genes/kb):

$$Densidad\ génica\ = \frac{N_{genes}}{LG \div 1000}$$

4. Longitud promedio de los genes codificantes:

$$Longitud\ promedio = \frac{\sum_{i=1}^{n} LCDs_i}{n}$$

Donde:

- $G,\ C,\ A,\ T$ → número de nucleótidos de Guanina, Citosina, Adenina y Timina, respectivamente, en el genoma.
- $\mathrm{LCDs_i}$ → longitud de los CDs en i pares de bases
- n → número total de CDs en el genoma
- LG → longitud del genoma
- N_{genes} → número total de genes codificantes anotados

Se utilizó el paquete `Gviz` para graficar la organización de los genomas donde cada módulo fue coloreado de forma distinta para su fácil visualización. Además, las anotaciones finales fueron exportadas a un archivo GFF con la ayuda de la librería "`Bioc.gff`".

A partir de este punto los datos necesarios para la continuación del pipeline ya se encuentran en el entorno de R y se exportan los resultados a un único libro de Excel y las imágenes en formato png.

Clasificación de fagos

Genómica comparativa

Método 1

La clasificación taxonómica de los bacteriófagos analizados se realizó utilizando vConTACT2, ejecutado en la plataforma CyVerse (https://cyverse.org/). Como datos de entrada se empleó un archivo multifasta que contenía todas las proteínas codificadas por los genomas de los fagos estudiados, junto con un archivo de mapeo gen–genoma. La similitud proteína–proteína fue calculada mediante DIAMOND, utilizando un umbral de valor E de 1×10^{-5}. La construcción de los clústeres virales (virus clusters, VC) se llevó a cabo empleando el algoritmo ClusterONE. Para esta comparación se utilizó como base de referencia la colección NCBI Bacterial and Archaeal Viral RefSeq, versión 85.

La asignación de los fagos a subclusters, tanto para los genomas analizados en este trabajo como para aquellos que no habían sido incluidos en el análisis de Oliveira et al. (117), se realizó siguiendo un enfoque comparativo basado en similitud genómica global. En primer lugar, se utilizó Gegenees para realizar un alineamiento fragmentado de un conjunto de 187 genomas de fagos. A partir de la matriz de similitud obtenida, se construyó una red filogenética mediante el método Neighbor-Net, implementado en el programa SplitsTree versión 4.0. Esta aproximación permitió definir relaciones filogenéticas entre los fagos y asignarlos a los subclusters correspondientes de acuerdo con su proximidad genómica.

Método 2

Como estrategia complementaria al enfoque comparativo clásico, se desarrolló un método automatizado para la asignación de bacteriófagos a los subclusters utilizando aprendizaje supervisado. Este método se basó en la composición de dominios funcionales codificados en los genomas de los fagos, integrando información estructural y funcional de las proteínas virales.

Para el entrenamiento del modelo se utilizó un conjunto de fagos previamente clasificados en subclusters, obtenidos a partir del análisis de Oliveira et al. (117) y curados manualmente para asegurar la consistencia entre genomas y anotaciones funcionales. La unidad de clasificación fue el genoma del fago, identificado por su número de acceso, mientras que las variables predictoras

correspondieron a la presencia o ausencia de dominios proteicos anotados en cada genoma. Los dominios funcionales fueron definidos a partir de anotaciones InterPro/PFAM y, cuando correspondió, se agruparon en arquitecturas de dominios para evitar redundancias debidas a anotaciones parciales o solapadas. La elección de las mismas se realizó en base a lo reportado previamente por el equipo de Oliveira sin tener en cuenta proteínas hipotéticas.

A partir de estas anotaciones se construyó una matriz binaria de características (feature matrix), donde cada fila representó un fago y cada columna un dominio funcional, codificado como 1 (presente) o 0 (ausente). Adicionalmente, se incorporó información sobre la especie hospedadora como variables categóricas binarias, con el objetivo de evaluar su contribución a la discriminación entre subclusters.

El modelo de clasificación se entrenó utilizando el algoritmo de árboles de decisión (Random Forest), implementado en el paquete `randomForest` de R. Dado el marcado desbalance entre subclusters —incluyendo subclusters representados por un único genoma— se aplicaron pesos de clase inversamente proporcionales a la frecuencia de cada subcluster, con el fin de mitigar el sesgo del modelo hacia las clases mayoritarias. El entrenamiento se realizó utilizando 2.000 árboles y se habilitó el cálculo de medidas de importancia de variables.

El desempeño del modelo se evaluó mediante el error fuera de bolsa (out-of-bag, OOB), así como a través de la inspección de la matriz de confusión, lo que permitió identificar subclusters correctamente clasificados y aquellos con mayor tasa de error. La importancia de los dominios funcionales en la clasificación se analizó a partir de las métricas de Disminución Media de la Precisión (Mean Decrease in Accuracy) y Disminución Media de Gini (Mean Decrease in Gini), permitiendo identificar dominios informativos a nivel global y específicos de determinados subclusters.

Finalmente, el modelo entrenado fue utilizado para predecir la asignación a subclusters de fagos no incluidos en el conjunto de entrenamiento, generando una clasificación automática basada exclusivamente en su composición funcional.

Análisis de genes esenciales

Casete de Lisogenia

Método 1

Se extrajeron las secuencias nucleotídicas de 44 integrasas utilizadas por Goerke y su equipo (47) de la base datos del NCBI y se analizaron junto con las 14 integrasas de nuestra colección. Las secuencias se alinearon con ClustalW y se construyó un árbol filogenético por Neighbor-Joining usando SeqinR y visualizado con iTOL para clasificar los diferentes tipos de integrasas. Adicionalmente, otros miembros

del equipo de nuestro laboratorio alinearon las secuencias proteicas con MAFFT y se generó una visualización en Jalview. Además, estimaron las relaciones evolutivas entre integrasas de serina mediante máxima verosimilitud en PhyML, con la representación final del árbol en iTOL.

Método 2

Con el fin de replicar dicho trabajo e implementar la automatización de la clasificación de las integrasas se procede en tres fases:

1. Fase I: Análisis no supervisado para definir los grupos SaInt.
2. Fase II: Clasificación supervisada de nuevas integrasas.
3. Fase III: Clasificación de nuevas secuencias (fase que se incluye en el pipeline automático).

Fase I: Análisis no supervisado para definir los grupos SaInt

Se recolectaron las secuencias de aminoácidos de integrasas disponibles en la base de datos de referencia junto con sus metadatos asociados, incluyendo identificadores de proteína, arquitectura de dominios Pfam, tipo de integrasa (serina [S-Int] o tirosina [T-Int]). Cuando estuvo disponible, se incorporó además la clasificación propuesta por Goerke et. al., asignando a las integrasas los grupos SaInt1- SaInt12.

Las secuencias proteicas fueron alineadas mediante el alineamiento múltiple de secuencias (MSA) utilizando el algoritmo ClustalW, implementado en el paquete msa, un método ampliamente validado para la comparación de secuencias proteicas homólogas. A partir del alineamiento se calculó una matriz de distancias entre secuencias, basada en la divergencia de aminoácidos. Esta matriz se utilizó para dos propósitos complementarios: (i) la visualización de la similitud global entre las integrasas mediante un árbol Neighbor-Joining (NJ) y (ii) la definición objetiva de clústeres mediante clustering jerárquico aglomerativo.

El árbol NJ se construyó utilizando el método Neighbor-Joining, con el objetivo de representar la similitud global entre las secuencias y facilitar la exploración de las relaciones evolutivas entre integrasas. Cabe destacar que este árbol se empleó exclusivamente con fines descriptivos y exploratorios, y no como una reconstrucción filogenética formal.

En paralelo, la matriz de distancias se utilizó como entrada para un análisis de clustering jerárquico aglomerativo (función `hclust`), empleando el método de enlace promedio (average linkage). El dendrograma resultante fue evaluado sistemáticamente explorando distintos puntos de corte, analizando la relación entre la altura de corte, el número de clústeres resultantes y el tamaño mínimo de cada clúster. A partir de este análisis se seleccionó empíricamente un corte que definió 13 clústeres evolutivamente coherentes.

Cada clúster fue evaluado en función de: (1) la presencia de integrasas previamente clasificadas según Goerke et al., (2) la coherencia del tipo de integrasa (S-Int o T-Int) y (3) la consistencia de la arquitectura de dominios Pfam. El resultado de esta fase fue un conjunto de integrasas de referencia etiquetadas y validadas, que se utilizó como conjunto de entrenamiento para la fase de clasificación supervisada.

Fase II: Clasificación supervisada de nuevas integrasas

Con el objetivo de asignar nuevas integrasas a los grupos SaInt definidos en la Fase I, se desarrolló un enfoque de clasificación supervisada basado en distancias. A partir de la matriz de distancias par-a-par derivada del alineamiento múltiple, se construyó una tabla de características (features) donde cada integrasa fue representada por un vector de distancias medias a cada uno de los clústeres de referencia. Para cada integrasa, se calculó la distancia promedio a las secuencias pertenecientes a cada grupo, generando un espacio de características de dimensión igual al número de clústeres. Este enfoque permite capturar la proximidad relativa de una integrasa a los distintos grupos evolutivos definidos previamente.

Así, se utilizó un clasificador k-Nearest Neighbors (kNN) entrenado sobre este espacio de distancias, donde la clase asociada a cada instancia correspondió al grupo asignado en la Fase I. La performance del clasificador se evaluó mediante validación cruzada Leave-One-Out (LOOCV), explorando distintos valores del parámetro k. El valor óptimo de k se seleccionó en función de la precisión global del modelo y del equilibrio entre exactitud y robustez frente al sobreajuste. La exactitud del clasificador fue evaluada para distintos valores del parámetro k. Se observó una exactitud muy elevada para valores bajos de k, con un desempeño máximo para k = 1 (exactitud = 1.00) y una meseta de alta exactitud para k = 3 y k = 5 (exactitud ≈ 0.99). Para valores mayores de k, la exactitud disminuyó levemente y se estabilizó alrededor de 0.97. En base a estos resultados, se seleccionó k = 3 como valor óptimo, al combinar una elevada exactitud con una mayor robustez frente a ruido local.

Una vez optimizado, el clasificador final se entrenó utilizando la totalidad del conjunto de referencia y se almacenó junto con la matriz de distancias y los metadatos mínimos necesarios para su reutilización.

Fase III: Clasificación de nuevas secuencias

Esta fase corresponde a la implementación del clasificador dentro de un pipeline automático para la clasificación de nuevas integrasas. En esta etapa, las secuencias nuevas se alinean junto con el conjunto de referencia, permitiendo el cálculo de las distancias entre las nuevas integrasas y las secuencias de entrenamiento previamente definidas.

Las nuevas integrasas, en este caso las correspondiente a los 14 fagos temperados, no se incorporan al clustering jerárquico original, con el fin de preservar la estructura de referencia de los grupos. En cambio, se calculan las distancias entre cada nueva integrasa y los grupos de referencia, generando el vector de características requerido por el clasificador kNN entrenado en la Fase II.

Cada integrasa nueva es entonces asignada a uno de los 13 grupos más probable, de acuerdo con la clase mayoritaria entre sus k vecinos más cercanos en el espacio de distancias. Este enfoque permite la incorporación escalable y reproducible de nuevas secuencias sin necesidad de recalcular el clustering original, asegurando la consistencia del sistema a lo largo del tiempo.

Casete de Empaquetamiento del ADN y morfogénesis de cápside

Método 1

El mecanismo de empaquetamiento del ADN fue inferido mediante el análisis filogenético de las subunidades grandes de la terminasa (TerL). Para ello, las secuencias aminoacídicas de TerL correspondientes a los fagos analizados fueron comparadas con un conjunto de secuencias de referencia para las cuales el mecanismo de empaquetamiento ha sido determinado experimentalmente (52).

Un total de 87 secuencias de TerL fueron alineadas utilizando el programa MAFFT, empleando el método iterativo G-INS-i, adecuado para alineamientos precisos de proteínas con regiones conservadas. A partir de este alineamiento, las regiones filogenéticamente informativas fueron seleccionadas mediante el programa BMGE (versión 1.2), utilizando un umbral de entropía de 0.8 para eliminar posiciones poco informativas o altamente variables.

Sobre la base del alineamiento depurado, se construyó un árbol filogenético utilizando el algoritmo FastTree. La inferencia del mecanismo de empaquetamiento se realizó asumiendo que los fagos que agrupan filogenéticamente con secuencias de referencia previamente caracterizadas y que comparten proteínas TerL homólogas utilizan un mecanismo de empaquetamiento de ADN similar.

Método 2

Se identificó el módulo de empaquetamiento del ADN de las anotaciones previas de los fagos a utilizar. Para esto se extrajeron, para cada contig, los genes comprendidos desde el último ORF asignado al casete de metabolismo de ADN, ARN y nucleótidos hasta el ORF inmediatamente anterior al casete de morfogénesis de cola. Este criterio se aplicó de manera consistente a todos los genomas, permitiendo aislar un bloque génico conservado correspondiente al módulo de empaquetamiento y región estructural temprana.

La organización génica del módulo fue analizada gráficamente mediante primero, la generación de objetos GRanges, en los que las coordenadas génicas fueron normalizadas por contig, estableciendo el inicio del módulo en la posición cero. Para facilitar la comparación entre fagos, cada módulo fue representado como un track independiente utilizando la librería Gviz, manteniendo la orientación génica y la anotación funcional de cada ORF. Las funciones proteicas se representaron mediante un esquema de colores consistente, facilitando la identificación visual de genes homólogos entre genomas.

Posteriormente, se identificó la proteína terminasa de subunidad grande (TerL) a partir de su anotación funcional. Para cada TerL se registró su longitud proteica estimada, calculada a partir de la longitud nucleotídica del ORF correspondiente, y su arquitectura de dominios, obtenida a partir de las anotaciones de dominios conservados. Adicionalmente, se evaluó la presencia de una proteasa de pro-cabeza (prohead protease) codificada en el mismo módulo, dado su rol funcional en ciertos mecanismos de empaquetamiento del ADN.

La estrategia de empaquetamiento del ADN fue inferida integrando la arquitectura de dominios de la TerL y la presencia o ausencia de una proteasa asociada al módulo. Las TerL que presentaron arquitecturas de dominios compatibles con proteínas tipo P22 fueron clasificadas como asociadas a un mecanismo de empaquetamiento tipo *pac*. En contraste, aquellas TerL con arquitectura compatible con proteínas tipo HK97 fueron clasificadas como asociadas a un mecanismo tipo *cos*. Los casos que no se ajustaron claramente a estos criterios fueron clasificados como ambiguos. Esta inferencia se utilizó exclusivamente como una categorización funcional basada en características estructurales y organizacionales del módulo, y no implica la identificación directa de sitios de corte o señales de empaquetamiento en las secuencias genómicas.

Los resultados del análisis de TerL, incluyendo la arquitectura de dominios, longitud proteica estimada y la estrategia de empaquetamiento inferida, fueron integrados en una tabla resumen y exportados para su posterior uso en análisis comparativos y visualización conjunta con otros módulos funcionales.

Casete de Morfogénesis de cola

Método 1

El análisis del módulo de interacción con el hospedador se realizó con el objetivo de identificar y caracterizar las proteínas estructurales involucradas en el reconocimiento y la adhesión del bacteriófago a la célula bacteriana. La función putativa de los componentes de este módulo se evaluó mediante la comparación de estructuras proteicas y dominios conservados utilizando las bases de datos HHPred, Protein Data Bank (PDB) y Pfam.

Otros miembros del laboratorio continuaron con el análisis reconstruyendo primero la organización génica del módulo de interacción a partir de los genomas completos de los fagos mediante el uso del software EasyFig (versión 2.2.2) (118), lo que permitió la visualización comparativa de la disposición de los genes involucrados en la estructura de la cola y la placa basal.

Las proteínas medidoras de la cola (tail measure proteins, TMPs) fueron analizadas en mayor profundidad debido a su relevancia estructural y funcional dentro del módulo. Para ello, se construyó una base de datos local compuesta por secuencias de aminoácidos de TMPs. Estas secuencias fueron alineadas utilizando el programa MAFFT (versión 7), empleando el método iterativo G-INS-i, con ponderación por WSP (weighted sum-of-pairs) y evaluación de consistencia del alineamiento seguido de la identificación de las regiones filogenéticamente informativas del alineamiento mediante el programa BMGE (versión 1.2) (119) , utilizando un umbral de entropía de 0,8 para la eliminación de posiciones poco informativas. A partir del alineamiento filtrado se construyó un árbol filogenético empleando el método de máxima verosimilitud implementado en el software PhyML.

La visualización y edición final de los árboles filogenéticos se realizaron utilizando la plataforma Interactive Tree Of Life (iTOL), lo que permitió una representación clara de las relaciones evolutivas entre las proteínas analizadas.

Método 2

Se diseñó el siguiente flujo de trabajo para caracterizar el módulo de interacción con el hospedador de los bacteriófagos analizados mediante un enfoque no supervisado, centrado en la organización génica y en el análisis estructural y funcional de proteínas clave de la cola. Este abordaje permite evaluar de manera comparativa los componentes involucrados en el reconocimiento del hospedador sin asumir clasificaciones previas. El análisis se estructuró en tres fases complementarias: (i) análisis de las Tape Measure Proteins (TMPs), (ii) identificación y caracterización de proteínas candidatas a Receptor Binding Proteins (RBPs), y (iii) análisis de la organización génica del módulo de cola.

Fase I: Análisis de Tape Measure Proteins (TMPs)

Se realizó un análisis descriptivo focalizado en las TMPs, proteínas cuya longitud se ha asociado con la morfología del virión y con la organización del módulo de reconocimiento del hospedador cuya función es facilitar el tránsito del ADN al citoplasma de la célula durante la infección.

Estos genes fueron identificados a partir de las anotaciones funcionales del conjunto de genomas, seleccionando aquellas proteínas anotadas como *tape measure protein*. Para cada fago, se verificó la presencia de una única TMP dentro del módulo de morfogénesis de cola, condición esperable dada la organización conservada de este módulo en fagos caudovirales.

La longitud de las TMPs fue calculada directamente a partir de la secuencia nucleotídica y utilizada como variable cuantitativa para comparar los fagos analizados. Paralelamente, la arquitectura de dominios se definió como la combinación de dominios detectados en cada proteína, permitiendo agrupar las TMPs según su organización modular independientemente de su longitud absoluta. Los datos fueron organizados en una tabla resumen que incluyó el identificador del fago, el identificador de la proteína, la longitud de la TMP (recuento de nucleótidos) y los dominios detectados.

La variabilidad en la longitud de las TMPs fue visualizada mediante gráficos de tipo lollipop (cada TMP fue representada como un segmento vertical desde el origen hasta su longitud total, acompañado por un punto terminal), ordenando los fagos según la longitud proteica y coloreando las proteínas de acuerdo con su arquitectura de dominios. Este enfoque permitió evaluar simultáneamente la variabilidad de tamaños y la diversidad estructural de estos genes entre los genomas analizados.

Este enfoque permite identificar patrones de conservación y divergencia en las TMPs entre los distintos fagos analizados, así como explorar la relación entre la organización del módulo de cola y la arquitectura proteica, sin recurrir a análisis filogenéticos ni a modelos supervisados. Los resultados obtenidos en esta fase fueron posteriormente integrados con la organización genómica del módulo de cola para su interpretación funcional.

Fase II: Análisis de proteínas candidatas a proteínas de unión a receptores (RBPs)

Para cada genoma, se delimitó el casete de morfogénesis de cola como la región comprendida entre la primera proteína anotada dentro de dicho casete y la proteína inmediatamente anterior a la holina, la cual constituye de forma consistente el primer gen del módulo de lisis. Esta estrategia permitió excluir proteínas claramente asociadas a lisis, evitando al mismo tiempo la pérdida de proteínas líticas ocasionales anotadas dentro del casete de cola, y conservar proteínas hipotéticas o con funciones adicionales potencialmente relevantes para el reconocimiento del hospedador (como las glucosaminidasas).

Dado que en fagos caudovirales las RBPs suelen encontrarse en posiciones relativas posteriores a las TMPs, las proteínas candidatas a RBP se definieron como aquellas localizadas corriente abajo de este gen y se extrajeron sus secuencias proteicas. En paralelo, se recopiló un conjunto de RBPs de referencia previamente caracterizadas en fagos de *Staphylococcus aureus*, seleccionadas a partir de la literatura y de bases de datos públicas (Tabla 4).

Las secuencias candidatas y de referencia se alinearon conjuntamente mediante alineamiento múltiple utilizando el algoritmo MUSCLE, seleccionado por su desempeño robusto en alineamientos de proteínas con baja identidad de secuencia y longitudes variables. El alineamiento resultante fue

convertido en una matriz de caracteres para su análisis cuantitativo y se calculó, así, el porcentaje de identidad par-a-par (PID) entre todas las secuencias, ignorando posiciones con gaps en ambas secuencias comparadas. El resultado fue una matriz de identidad aminoacídica all-versus-all que incluyó tanto proteínas candidatas como RBPs de referencia.

Tabla 4: Proteínas de unión a receptores (RBPs) de referencia

Fago	ORF	Id en base de datos	Publicación
Φ11	gp45	NP_803298.1	Kizziah JL, Manning KA, Dearborn AD, Dokland T (2020) (120)
80α	gp61	YP_001285375.1	
ΦSA012	Orf103	YP_009006774.1	Takeuchi et. al. (2016) (121)
ΦSA012	Orf105	YP_009006776.1	
812	Orf123	YP_009224533.1	Botka et. al. (2019) (122)
812	Orf125	YP_009224535.1	

La matriz de similitud fue visualizada mediante mapas de calor (heatmap), permitiendo inspeccionar patrones globales de similitud y agrupamiento entre las proteínas analizadas. Adicionalmente, se realizó una comparación dirigida entre cada proteína candidata y todas las RBPs de referencia, conservando todos los valores de similitud mayores a 20% para cada par candidato–referencia. Este enfoque permitió detectar similitudes múltiples potenciales, en concordancia con la evidencia de que algunos fagos pueden codificar más de una RBP funcional.

Los resultados fueron integrados con información contextual, incluyendo la distancia relativa de cada proteína candidata a la TMP, facilitando la posterior definición de RBPs candidatas finales.

Fase III

Con el fin de comparar la organización génica del módulo de cola entre los distintos fagos, se representó gráficamente el módulo de morfogénesis de cola para cada genoma. Para permitir una comparación directa entre fagos de diferente tamaño, las coordenadas genómicas fueron transformadas a posiciones relativas, estableciendo el inicio en la posición cero para cada contig.

Las regiones génicas fueron representadas mediante objetos GRanges, incorporando información sobre orientación, función anotada y tipo de característica. Para la visualización se generaron pistas genómicas individuales para cada fago (`GeneRegionTrack`), utilizando un cromosoma ficticio común, y se representaron las proteínas como flechas orientadas según el marco de lectura utilizando la librería Gviz.

Las proteínas fueron coloreadas de acuerdo con su función anotada, permitiendo identificar patrones conservados de organización génica y variaciones entre fagos. Este análisis permitió evaluar la estructura modular del casete de cola entre los 14 fagos analizados y contextualizar la posición de las TMPs y RBPs candidatas dentro de la arquitectura global del módulo.

Casete de Lisis

Método 1

La composición de dominios proteicos de las endolisinas codificadas por los fagos fue analizada utilizando la base de datos Pfam, empleando un umbral de significancia de E-value < 1. A partir de los dominios identificados, las endolisinas fueron clasificadas en distintos tipos de acuerdo con el esquema propuesto por Chang y Ryu (65) (Tabla 5), el cual se basa en la arquitectura modular de estas proteínas y permite distinguir variantes funcionales y estructurales.

En paralelo, el polimorfismo de longitud de las holinas fue determinado de manera manual a partir de las secuencias nucleotídicas obtenidas de los genomas de fagos anotados. Este análisis consistió en la comparación directa de la longitud (recuento de nucleótidos) de las holinas, criterio utilizado previamente para su clasificación en diferentes grupos, dado que la variabilidad en el tamaño de estas proteínas se podría asociar con diferencias en su función y mecanismo de acción durante el proceso de lisis celular.

Tabla 5: Arquitectura de dominios funcionales de endolisinas según el tipo

Tipo	EAD	EAD	CBD
I	PF05257	-	**PF24246*
	CHAP domain	-	SH3b_T
II	PF05257	-	PF08460
	CHAP domain	-	SH3_5
III	PF05257	PF01510	PF08460
	CHAP domain	Amidase 2	SH3_5
IV	PF05257	PF01520	PF08460
	CHAP domain	Amidase 3	SH3_5
V	PF05257	PF01520	**PF25606*
	CHAP domain	Amidase 3	SH3b_P2
VI	PF01551	PF01510	PF08460
	Peptidase family M23	Amidase 2	SH3_5

Método 2

Con el objetivo de reproducir y sistematizar el análisis manual del casete de lisis, se desarrolló un pipeline automático que permitió analizar de forma estandarizada las holinas y endolisinas codificadas por los fagos.

Fase I: Caracterización automática de holinas y endolisinas

En el caso de las holinas, se estudió tanto la arquitectura de dominios Pfam como el polimorfismo de longitud mediante el recuento del número de nucleótidos correspondiente a cada gen. Este enfoque permite clasificar las holinas en función de su tamaño génico, replicando el criterio utilizado en el análisis manual, pero de manera reproducible y escalable.

A continuación, las endolisinas fueron clasificadas de acuerdo con su tipo funcional. Para ello, se extrajeron las secuencias proteicas correspondientes desde la base de datos con su anotación de dominios derivadas de Pfam. Cada endolisina fue etiquetada en función de la combinación de dominios detectados, siguiendo la clasificación previamente establecida (Tipos I a V).

Fase II: Curación del dataset y entrenamiento del clasificador de endolisinas

Durante la anotación automática de dominios se observó que, dependiendo del momento de la búsqueda en la base de datos Pfam, algunas endolisinas Tipo I y Tipo V cuyos dominios de unión a la pared celular (CBD) no presentan homología a dominios SH3 reportados podían aparecer erróneamente como portadoras de dichos dominios. Esta ambigüedad dificultó la discriminación directa entre endolisinas Tipo I vs Tipo II y Tipo V vs Tipo IV basándose exclusivamente en la anotación de dominios.

Con el fin de resolver esta limitación, se desarrolló un clasificador supervisado capaz de distinguir entre estos tipos de endolisinas. Previo al entrenamiento, se llevó a cabo una curación estricta del conjunto de datos, excluyendo aquellas secuencias que cumplían al menos uno de los siguientes criterios:

- Secuencias cuya arquitectura no correspondía a ninguno de los tipos de endolisinas considerados o que presentaban dominios inesperados.
- Secuencias con endolisinas tipo VI debido al bajo número de representantes disponibles (N = 1), lo cual impide un análisis estadístico y computacional robusto.
- Secuencias con longitudes proteicas atípicas para su grupo (outliers) (<200nt.).
- Secuencias incorrectamente anotadas.

Como resultado de este proceso de curación se generó un dataset final compuesto por 295 secuencias de endolisinas, cada una asociada a una etiqueta de clase numérica (1 a 5) correspondiente a su tipo

funcional. Este conjunto fue vinculado con las secuencias aminoacídicas mediante identificadores únicos, permitiendo su utilización directa en el entrenamiento del modelo.

Sobre este conjunto se entrenó un modelo de Support Vector Machine (SVM) utilizando un kernel Gappy Pair, implementado mediante los paquetes kebabs y e1071. Se emplearon los parámetros k = 1 y m = 2, adecuados para capturar patrones locales de secuencia relevantes para la discriminación entre tipos de endolisinas. Dado el marcado desbalance entre las clases, se incorporaron pesos de clase inversamente proporcionales a su frecuencia, con el fin de evitar el sesgo del clasificador hacia los grupos mayoritarios.

El conjunto de datos fue dividido aleatoriamente en un 70% para entrenamiento (N = 206) y un 30% para prueba (N = 89). El desempeño del modelo fue evaluado mediante el cálculo explícito de métricas de clasificación, incluyendo la precisión (accuracy) global y balanceada, sensibilidad y precisión por clase, así como la construcción de una matriz de confusión. Una vez validado su desempeño, el modelo entrenado fue almacenado para su utilización posterior en la clasificación de nuevas secuencias de endolisinas.

Fase III: Clasificación de nuevas endolisinas

El clasificador SVM entrenado fue finalmente aplicado para la clasificación automática de las endolisinas correspondientes a los 14 fagos del laboratorio. Las predicciones obtenidas fueron comparadas con las clasificaciones reportadas previamente (Método I), las cuales habían sido realizadas de manera manual mediante la identificación de dominios en la plataforma web de InterPro.

Este enfoque permitió evaluar la concordancia entre el método automático y el análisis manual, así como validar la capacidad del pipeline propuesto para reproducir y extender la clasificación funcional de endolisinas de manera reproducible y escalable.

Análisis Complementarios

Ausencia de CBDs en endolisinas

Con el objetivo de explorar la posible presencia de dominios de unión a carbohidratos (CBDs) no detectables mediante enfoques basados en anotación automática de dominios funcionales, se realizó un análisis estructural complementario de endolisinas seleccionadas. Este análisis se enfocó en aquellas endolisinas que, de acuerdo con la caracterización basada en dominios Pfam, no presentaban CBDs conocidos.

Se modelaron las estructuras tridimensionales de las endolisinas correspondientes a los fagos vB_SauS_308 (endolisina Tipo I) y vB_SauS_287 (endolisina Tipo V, al igual que vB_SauS_713), ya que

ninguna de estas proteínas presentó dominios funcionales compatibles con CBDs previamente reportados. Adicionalmente, se modeló la endolisina de vB_SauS_287 clasificada como Tipo V, con el objetivo de establecer comparaciones estructurales, dado que las endolisinas Tipo IV y Tipo V se diferencian principalmente por la ausencia de un CBD reconocido en estas últimas. Asimismo, se incluyó en el análisis la comparación con la endolisina Tipo I, la cual carece no solo de CBDs identificables sino también del segundo dominio catalítico (dominio amidasa).

Las predicciones estructurales se realizaron utilizando AlphaFold Server (https://alphafoldserver.com/) (123), empleando los parámetros por defecto. Las estructuras obtenidas fueron posteriormente superpuestas y alineadas utilizando ChimeraX (124). Este enfoque permitió evaluar la conservación estructural y la posible presencia de regiones estructurales compatibles con funciones de unión a la pared celular, aun en ausencia de dominios funcionales anotados.

Proteínas de unión a receptores y rango de hospedador.

El rango de hospedador de los fagos fue determinado fenotípicamente mediante ensayos de spot para dieciséis estafilococos coagulasa negativos (CNS) (4 *S. epidermidis*, 4 *S. hominis*, 2 *S. caprae*, 2 *S. haemolyticus*, 2 *S. capitis*, 1 *S. lugdunensis*, and 1 *S. sciuri*) por otros miembros del laboratorio. Brevemente, la técnica consiste en preparar placas indicadoras con medio TS y los cultivos de las cepas a ensayar. Luego se realizan dos spots por cada fago con distinta concentración y se incuban las placas a 30° por 24 horas para luego ser evaluadas. El detalle de dicha técnica puede consultarse en Suárez et. al (105). Además, algunos de los fagos se testearon en diferentes cepas de *S. aureus* para caracterizar su especificidad en el marco de la Tesis de Doctorado de Boncompain C.

RESULTADOS Y DISCUSIONES

Se presentan los resultados del pipeline semi-automático y se omiten los resultados del método manual, cuyos resultados ya han sido publicados, (105) referenciándose solo con fines comparativos y de validación.

Análisis de la base de datos

Número de individuos en la base de datos

Se creó una base de datos local con las secuencias en formato FASTA, anotaciones en formato GBF y metadatos (como un Excel para uso externo y tabla en el entorno de R) de fagos de *Staphylococcus spp.* alojadas en la base de datos de nucleótidos del NCBI. El término de búsqueda halló información de 901 secuencias.

A través de los Gi se accedió a los metadatos de la misma, cuyos datos crudos poseían 24 variables para caracterizarlos. En primera instancia, se filtró teniendo en cuenta que los individuos posean los criterios específicos de búsqueda: ser secuencias completas, cuyo material genético sea ADN, no sean plásmidos, sean organismos verificados y, por supuesto, pertenecer al género de *Staphylococcus*, provocando la disminución a 686 identificadores únicos. El último proceso de limpieza consistió en: (1) eliminar duplicados (por título, organismo o longitud de la secuencia con alto porcentaje de identidad) que estuviesen tanto en la base de datos de RefSeq como de INSDC, priorizando los primeros y (2) aplicar un filtro a nivel de la longitud de la secuencia eliminando, por un lado, aquellos individuos con una longitud menor a 5000 pb y, por otro lado, analizando aquellas de igual longitud. Esto último se basa en la premisa que, debido a la variabilidad de las secuencias, es poco probable encontrar dos que posean la misma longitud por azar. Por lo tanto, las secuencias fueron alineadas mediante BLAST NCBI y se observó que la mayoría de ellas eran posiblemente las mismas secuencias, teniendo en cuenta la cobertura de la secuencia y el porcentaje de identidad (poniendo un corte en el 95% para ambos casos, pero siendo en la mayoría igual o mayor al 99%). Solo tres pares de secuencias mostraron tener la misma longitud, pero baja cobertura o poca identidad y se mantuvieron en la base de datos (Gi: 2904591305 y Gi: 2618404572; Gi:2046360946 y Gi:2320100154; Gi:2260453719 y Gi:2905010333). Así, como resultado final, la base de datos fue conformada por 524 secuencias genómicas (Material Suplementario: 01_metadata_db_st.xlsx).

Mediante los Gi de los 524 individuos se obtuvieron las secuencias en formato FASTA y los archivos GBF con las anotaciones de cada fago. Gracias al proceso de limpieza de los datos la cantidad de identificadores (y por ende secuencias) únicas y confiables para utilizar representan aproximadamente 59% del número de individuos inicial.

Análisis de la composición de la base de datos

Se analizó la información proveniente a los metadatos de los fagos disponibles en la base de datos mediante un tablero en Power BI (Fig. 10) (Material Suplementario: 02_analisis_db.pbix). Al tratarse de un tablero dinámico, no solo logra una eficiente visualización de los datos para su presentación, sino que, además, permite la rápida inspección de los mismos al momento de realizar la limpieza y transformación. Es importante remarcar, en este punto, que la información puede ser parcial ya que no es obligación al subir las anotaciones a la base de datos del NCBI que todos los campos estén completos. Además, aun aquellos que están completos, tienen falta de criterio de homologación (por ejemplo, la fuente de aislamientos); razón por la cual, se unificaron las anotaciones en todos los campos que fue posible.

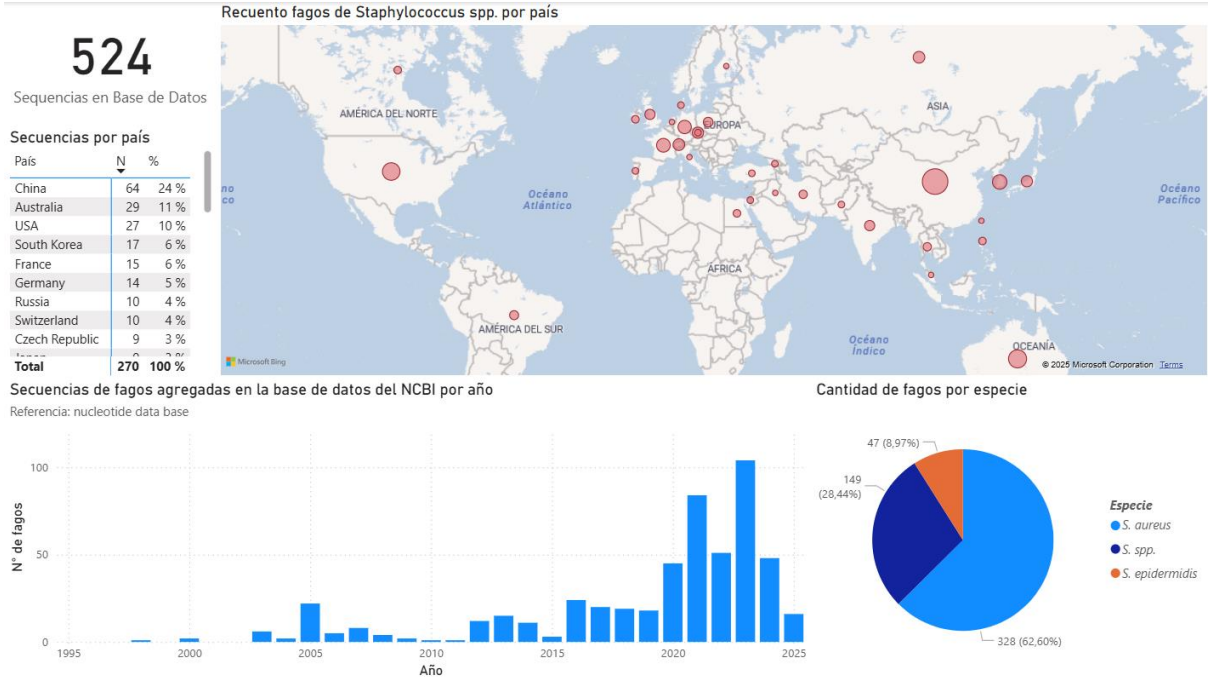

Figura 10: Tablero con información de las secuencias genómicas de fagos contra Staphylococcus spp. disponibles en la base de datos local

Todo esto puede visualizarse en la cantidad de individuos que tienen el campo “País” (región geográfica del aislamiento) con datos (270, 52% del total). Entre los países con mayor cantidad de aislamientos se encuentra China (24%) seguido por Australia (11%) y Estados Unidos (10%), aportando entre los tres el 75% de las secuencias con origen conocido. Esta tesis suma 14 secuencias a la base de datos, posicionándonos entre los diez países con más secuencias publicadas y demostrando el impacto de la misma.

Como se puede observar en la figura 10, del total de fagos del género *Staphylococcus* el 62.60% corresponden a la especie *aureus*, el 8.97% a *S. epidermidis* y el 28.44% restante a otras especies (*S. pseudintermedius, S. haemolyticus, S. warneri*, entre otros).

Otro dato evidente es el aumento de la cantidad de secuencias a partir de 2020, suponiendo al menos dos motivos para este incremento: la disminución del costo de secuenciación y la pandemia COVID-19. Esta última, a pesar de las distintas medidas de contención de la enfermedad llevada adelante por cada país, llevó a un aislamiento general de la población imposibilitando los trabajos de wet-lab y dirigiéndose los grupos de investigación a realizar mayor cantidad de análisis in silico.

Fagos aislados de *Staphylococcus aureus* en la base de datos

De las 524 secuencias en la base de datos local, 328 corresponden a fagos activos sobre *S. aureus* de los cuales menos del 50% se especifica la fuente de la cual fueron aislados (Fig. 11) siendo el ~72% aislados de aguas residuales o de muestras clínicas humanas. Las muestras restantes provienen en su mayoría de fuentes asociadas a animales, ya sean ambientales de granjas, leche de ganado o del animal per se. Se puede asumir, a priori, que a mayor cantidad de fagos aislados en un nicho mayor interés de buscar la solución a alguno problema basado en fagos, por ejemplo, la necesidad de aislar fagos para su posterior uso clínico.

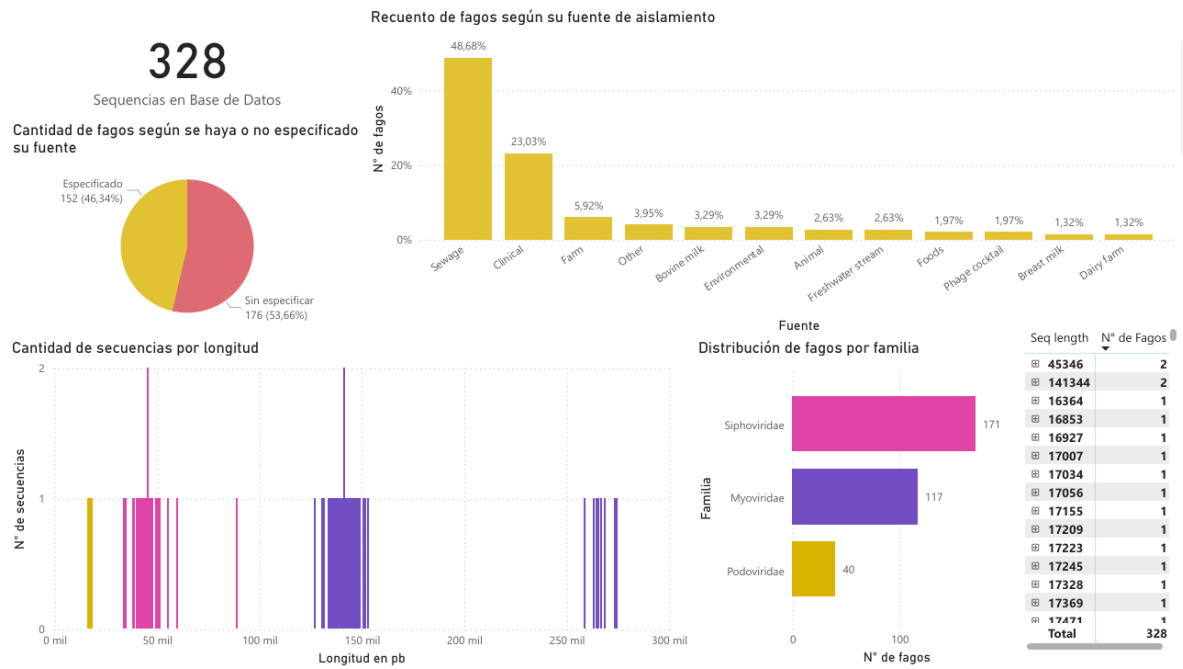

Figura 11: Tablero de fagos de Staphylococcus aureus disponibles en la base de datos

Por último, es importante ver como se distribuyen las secuencias según la longitud de las mismas (en pares de bases) observándose cuatro grupos bien definidos según su longitud. Clásicamente, los fagos de *S. aureus* se pueden agrupar según su longitud y su morfología en tres categorías distintas según su familia: Podovirus (el primer grupo con una longitud menor a 20kb), Siphovirus (el grupo central, más abundante, con una longitud promedio de ~40kb) o Myovirus (~125kb) (11) representados en el gráfico en color amarillo, magenta y violeta respectivamente. A su vez, el grupo de los Myovirus reportó en el último tiempo un nuevo subgrupo (con solo ocho representantes en nuestra base de datos) denominado "jumbo" con un tamaño de ~269kb, más del doble de kb que lo esperado (43) para este grupo. Por otro lado, podemos observar que la familia más abundante es la de los Siphovirus, coincidente a los fagos aislados en nuestro laboratorio, presuntamente por ser la técnica más frecuentemente utilizada para su aislamiento la inducción de profagos.

Proteínas fágicas en la base de datos

A partir de los 524 genomas de *Staphylococcus* disponibles en la base de datos, se lograron extraer 62440 secuencias proteicas correspondientes a 519 genomas. La extracción no fue posible en cinco genomas debido a que los archivos en formato GBF contenían únicamente la secuencia nucleotídica sin anotaciones y/o los archivos tenían regiones codificantes (CDs) no definidas. Dado que el proceso de extracción fue automatizado, la función encargada de procesar los archivos simplemente omite aquellos que presentan errores y continúa con el próximo, asegurando así el flujo de trabajo.

Las secuencias obtenidas fueron limpiadas para eliminar caracteres problemáticos como espacios o comas. De este filtrado resultaron 25587 secuencias únicas sin errores (aproximadamente el 41% del total). Estas secuencias fueron analizadas con InterProScan para identificar dominios funcionales para su caracterización, primero mediante Pfam y, de no poseer ninguno, se prosiguió con la búsqueda en GENE3D, SUPERFAMILY y NCBIFAM en ese orden.

Como resultado, se identificaron 597 arquitecturas modulares únicas, es decir, combinaciones específicas y ordenes únicos de dominios funcionales (Material Suplementario: 03_dominios_proteinas.xlsx). A partir de esta información y de las anotaciones originales, se definió una nomenclatura consensuada para clasificar funcionalmente las proteínas generando una lista con 210 funciones diferentes. Del total de secuencias (62440):

- Al 28,5% (17810) se les pudo asignar una función específica
- El 71,5% (44630) fue clasificada como hipotéticas. De estas, el 85% no contenían dominios funcionales detectables, y el 15% restante presentó dominios que no permitieron inferir una función clara.

Finalmente, se analizó cómo se agrupan estas proteínas dentro de distintos casetes funcionales. Excluyendo las proteínas sin función asignada (Fig. 12B), se observó que la mayoría está involucrada en procesos de metabolismo de ADN, ARN y nucleótidos (34 %), mientras que el casete con menor número de proteínas es el de lisogenia (3,5 %) – no teniendo en cuenta los factores de virulencia que no son *per se* un módulo específico.

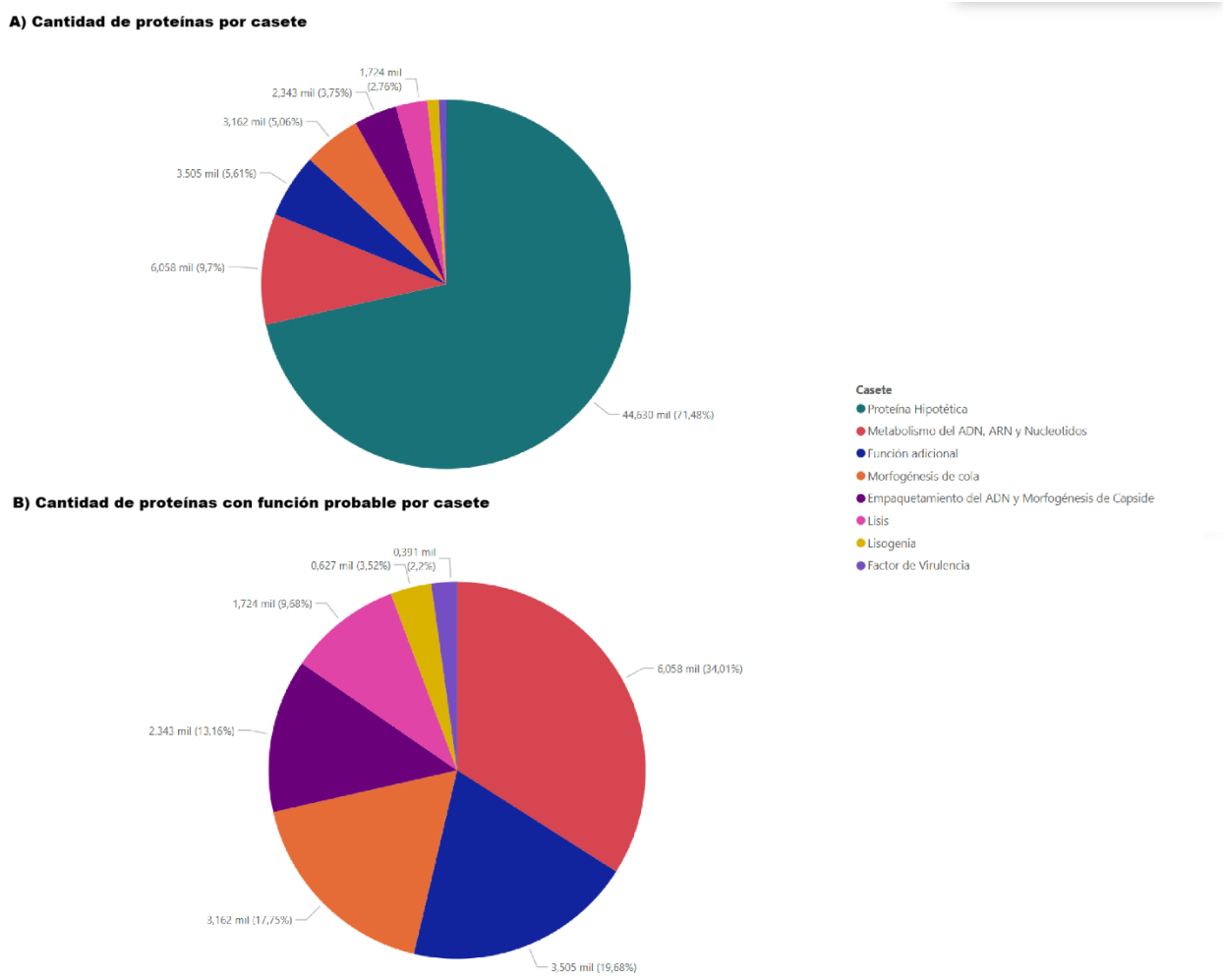

Figura 12: Recuento de Proteínas por módulo

A) Cantidad de proteínas teniendo en cuenta las proteínas hipotéticas. B) Cantidad de proteínas con función probable por casete

Anotación de fagos temperados

Las secuencias genómicas de 14 fagos temperados fueron anotadas utilizando un pipeline desarrollado con ese fin que utiliza, entre otros programas, PHANOTATE para el llamado de los ORFs e InterProScan para la asignación de funciones. Con el fin de comparar los resultados con y sin la utilización de este pipeline (105) deliberadamente se prescindió de la curación manual, los resultados se observan en la tabla 6, mostrándose entre paréntesis aquellas diferencias con la anotación manual. Los tamaños genómicos coinciden con los esperados para la familia de Siphovirus.

Los genomas tienen un promedio de contenido G+C de 34,39% (variando de 33,27% a 35,35%) y, como es esperable, se encuentran altamente empaquetados, con pequeñas regiones intergénicas (promedio de secuencia codificante de 95,17%) y una alta densidad génica (en promedio de 1,74 genes/kb). En promedio se identificaron 7 genes más por fago y disminuyó, a su vez, la longitud promedio de los genes debido a que la mayoría de los nuevos genes predichos son genes cortos (menores a 150pb) pero cuyas funciones pueden resultar imprescindibles pudiendo ser, por ejemplo, genes regulatorios. Las diferencias entre la anotación manual y la semi-automática se encuentra en el material suplementario: 04_diferencias_anotaciones.xlsx

Tabla 6: Características de los catorce fagos temperados

Bacteriófago	Tamaño genómico (pb)	%GC	N° genes	N° CDs	Densidad génica	Longitud Promedio CDs	%Codificante
vB_SauS_277g	44016	34,06	71 (*67*)	71 (*67*)	1,61 (*1,52*)	599 *(627)*	96,25 *(95,45)*
vB_SauS_287	43134	34,75	76 (*69*)	76 (*69*)	1,76 (*1,60*)	545 *(588)*	95,32 *(94,07)*
vB_SauS_308	39968	34,11	75 (*68*)	75 (*68*)	1,88 (*1,72*)	510 *(548)*	95,15 *(93,25)*
vB_SauS_320	45809	33,54	70 (*65*)	70 (*65*)	1,53 (*1,42*)	619 *(664)*	94,16 *(94,25)*
vB_SauS_321c	42397	34,08	74 (*67*)	74 (*67*)	1,75 (*1,58*)	546 *(598)*	94,76 *(94,54)*
vB_SauS_690	45771	33,27	74 (*64*)	74 (*64*)	1,62 (*1,40*)	585 *(664)*	94,09 *(92,90)*
vB_SauS_713	42961	34,26	69 (*62*)	69 (*62*)	1,61 (*1,44*)	596 *(654)*	95,28 *(94,43)*
vB_SauS_760	42788	35,35	76 (*66*)	76 (*66*)	1,78 (*1,54*)	538 *(605)*	94,94 *(93,29)*
vB_SauS_775	43321	33,91	74 (*69*)	74 (*69*)	1,71 (*1,59*)	564 *(600)*	95,75 *(95,61)*
vB_SauS_832	43499	34,07	77 (*70*)	77 (*70*)	1,77 (*1,61*)	540 *(586)*	95,17 *(94,31)*
vB_SauS_I73	43575	35,09	79 (*69*)	79 (*69*)	1,81 (*1,58*)	529 *(593)*	95,35 *(93,87)*
vB_SauS_Mh1	43800	35,09	83 (*75*)	83 (*75*)	1,89 (*1,71*)	511 *(554)*	96,29 *(94,82)*
vB_SauS_Mh15	43090	35,26	80 (*73*)*	82 (*73*)*	1,86 (*1,69*)	501 *(546)*	94,40 *(92,55)*
vB_SauS_Mh4	44199	34,65	76 (*70*)	76 (*70*)	1,72 (*1,58*)	558 *(598)*	95,40 *(94,71)*

(*) vB_SauS_Mh15 posee un intrón.

Con fines comparativos los genomas tienen en su inicio el gen que codifica la integrasa. Así, podemos ver de izquierda a derecha la organización modular de los genomas siguiendo el siguiente orden: casete de lisogenia, metabolismo del ADN, ARN y nucleótidos, empaquetamiento del ADN y morfogénesis de cápside, morfogénesis de cola y lisis (Fig. 13). Más de la mitad de los genes de dichos fagos son proteínas sin función conocida (54,74%) y se ubican principalmente en los módulos de integración y metabolismo del ADN, ARN y nucleótidos. Se encontraron tres factores de virulencia en los fagos: leucocidina de Panton-Valentine (PVL), toxina del sistema toxina-antitoxina del tipo

PemK/MazF y la proteína E, presentes en todos los fagos (PVL), en cinco (PemK/MazF) y en dos (proteína E).

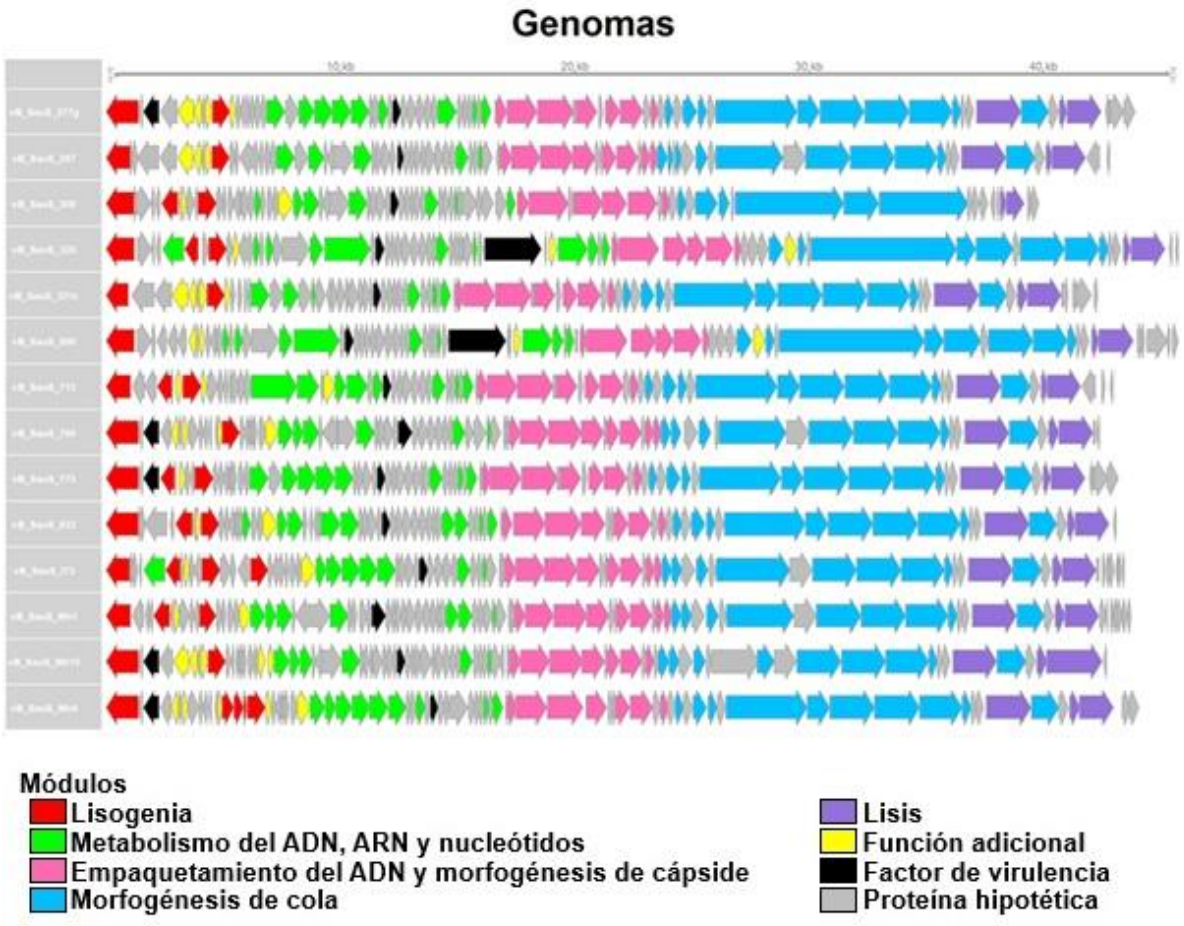

Figura 13: Organización modular de los fagos analizados en este trabajo coloreados según el casete al que pertenecen

Genómica comparativa

Clasificación supervisada de subclusters basada en dominios funcionales

Se desarrolló un modelo de clasificación supervisada basado en Random Forest con el objetivo de predecir la asignación de subclusters a partir de la presencia/ausencia de dominios funcionales anotados en los genomas de fagos. El conjunto de datos estuvo compuesto por 190 genomas, representando múltiples clases y subclases (clusters y subclusters), con un número variable de fagos por clase.

Debido al fuerte desbalance de clases/subclusters (Fig. 14) —incluyendo clases representadas por uno o muy pocos genomas— el entrenamiento y la evaluación del modelo se realizaron utilizando el mismo conjunto de datos, incorporando pesos de clase inversamente proporcionales a la frecuencia de cada subcluster. Esta estrategia permitió mitigar parcialmente el sesgo hacia las clases mayoritarias,

aunque limitó la posibilidad de una validación independiente. Por lo tanto, los resultados deben interpretarse como una evaluación exploratoria del potencial discriminante de las variables, más que como una estimación definitiva del desempeño predictivo.

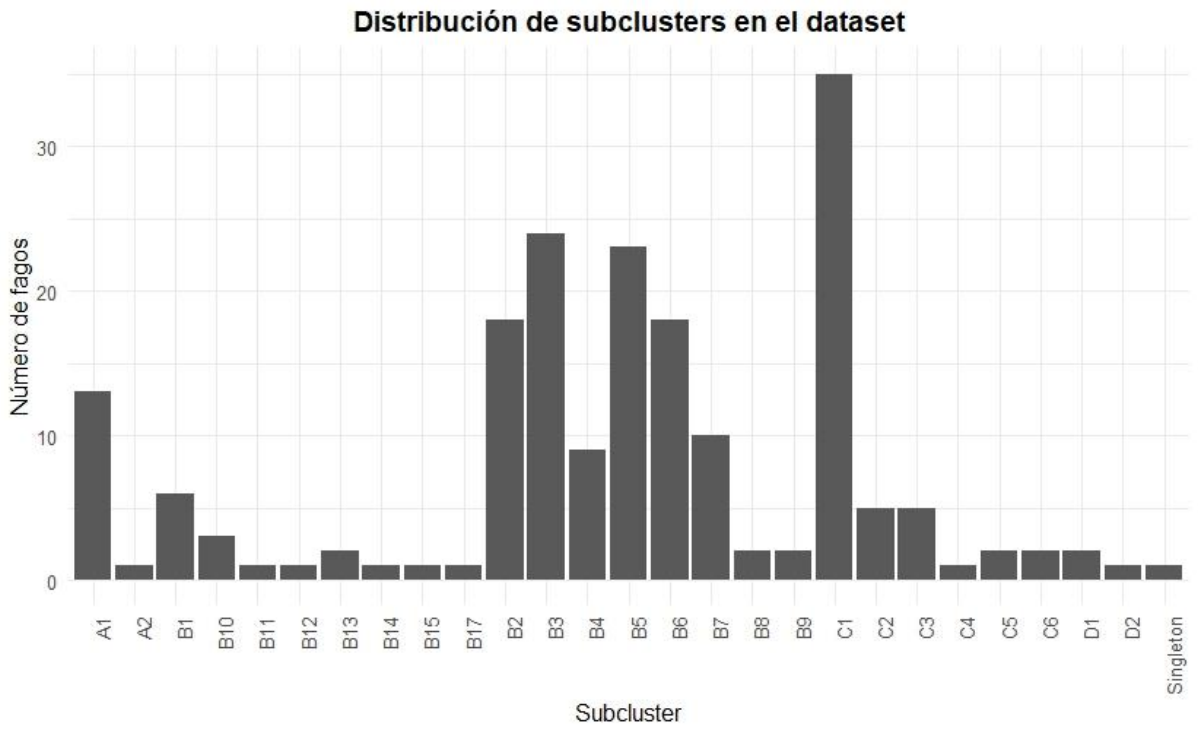

Figura 14: Número de fagos por subcluster en el dataset de entrenamiento

El desempeño global del modelo, evaluado mediante el error out-of-bag (OOB), fue moderado (OOB ≈ 0,21), indicando una capacidad predictiva parcial pero no suficiente para una clasificación robusta de todos los subclusters. Algunos grupos con mayor número de genomas (por ejemplo, varios subclusters del cluster B) fueron clasificados correctamente en la mayoría de los casos, mientras que otros subclusters minoritarios presentaron tasas de error elevadas o directamente no pudieron ser predichos correctamente. En la Tabla 7 se muestran los valores de error para cada clase, el mismo corresponde a la proporción de instancias incorrectamente clasificadas para cada subcluster, estimada a partir de las predicciones OOB del modelo.

Tabla 7: Tasa de error por clase del modelo RandomForest para la predicción de subclusters

Subcluster	A1	A2	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
class.error	0,00	1,00	0,00	0,05	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,33	1,00	1,00
Subcluster	**B13**	**B14**	**B15**	**B17**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**D1**	**D2**	**Singleton**	
class.error	1,00	1,00	1,00	1,00	0,63	0,00	0,20	1,00	0,00	1,00	0,50	1,00	1,00	

Las tasas de error por clase observadas reflejan, en gran medida, la estructura del conjunto de datos utilizado para el entrenamiento del modelo. Varios subclusters están representados por un número reducido de genomas, en algunos casos uno o dos individuos, lo que limita severamente la capacidad del clasificador para aprender patrones funcionales robustos asociados a dichas clases. En este contexto, valores elevados de error de clase no deben interpretarse necesariamente como un fracaso del modelo, sino como una consecuencia esperable del fuerte desbalance entre subclusters. Por el contrario, aquellos grupos con mayor número de representantes y arquitecturas funcionales más homogéneas, particularmente dentro del cluster B, presentan tasas de error bajas o nulas, lo que indica que el modelo logra capturar señales discriminantes consistentes cuando la información disponible es suficiente. Esta dependencia del desempeño respecto del tamaño de clase pone de manifiesto una limitación inherente a los enfoques supervisados aplicados a genomas fágicos, donde la disponibilidad de secuencias sigue siendo desigual entre linajes, y refuerza la necesidad de interpretar los resultados en clave exploratoria y comparativa más que predictiva estricta.

A pesar de estas limitaciones, el análisis de importancia de variables reveló patrones biológicamente informativos. Un número reducido de dominios concentró los valores más altos de Mean Decrease in Accuracy y Mean Decrease in Gini (ver Anexo Tabla 16), lo que sugiere que ciertas funciones genómicas aportan una señal discriminante relevante para la clasificación (Fig. 15).

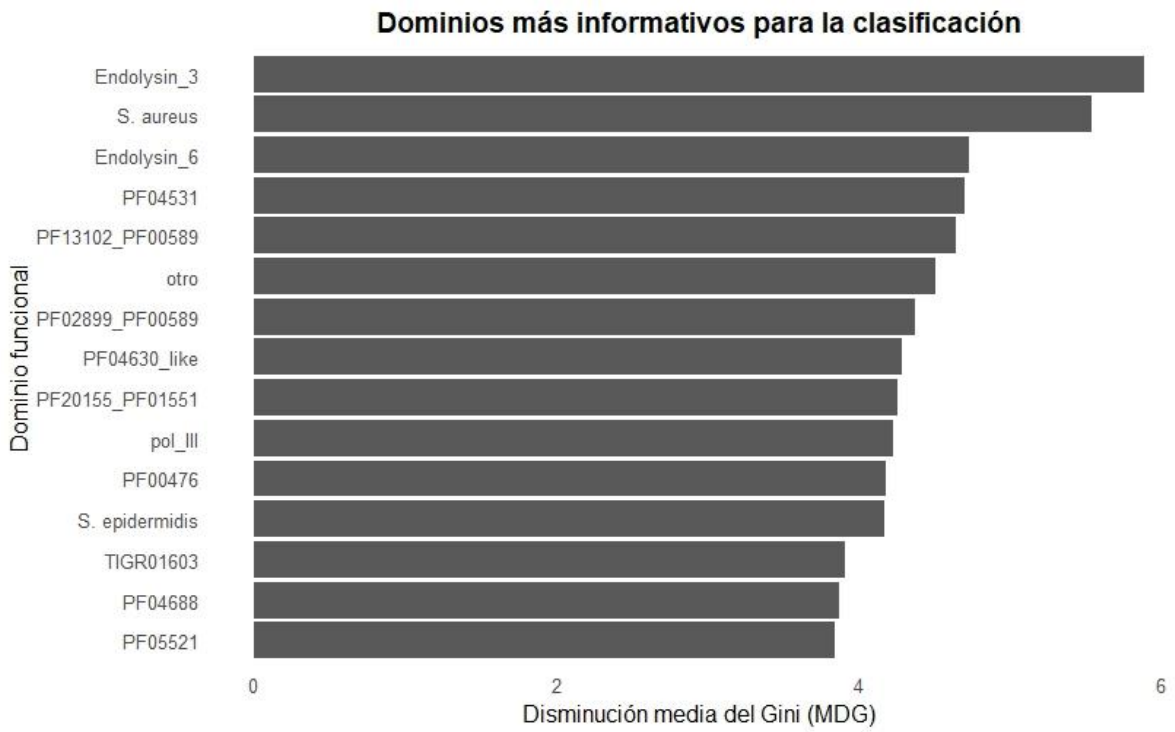

Figura 15: Esquema de los quince dominios más informativos para la clasificación de subclusters

Entre los dominios más importantes se destacaron aquellos asociados a enzimas líticas, particularmente dominios de endolisinas (por ejemplo, Endolysin_3 y Endolysin_6). Estas proteínas, responsables de la degradación del peptidoglicano bacteriano durante la fase lítica, son altamente modulares y presentan una notable variabilidad en su arquitectura. Su contribución al modelo sugiere que los mecanismos de lisis y la composición de los módulos catalíticos y de unión al sustrato podrían estar parcialmente conservados dentro de ciertos subclusters, actuando como marcadores funcionales.

Asimismo, dominios vinculados a replicación y mantenimiento del genoma, incluyendo proteínas relacionadas con la ADN polimerasa (por ejemplo, PF00476, pol_III), mostraron una alta importancia relativa. Estas funciones centrales del ciclo replicativo fágico podrían reflejar estrategias replicativas compartidas entre fagos filogenéticamente relacionados, contribuyendo a la señal capturada por el clasificador.

También se observaron dominios asociados a empaquetamiento del ADN y morfogénesis del virión, incluyendo componentes característicos de proteínas terminasa y proteínas estructurales del cápside y la cola. Este resultado es coherente con la organización modular de los genomas fágicos, donde los casetes estructurales suelen conservarse dentro de linajes específicos y han sido previamente utilizados como criterios taxonómicos en enfoques basados en redes y similitud proteica.

Las variables correspondientes a la especie hospedadora mostraron una alta contribución a la capacidad predictiva del modelo, en particular *Staphylococcus aureu*s, lo cual refleja la fuerte asociación entre ciertos subclusters y su rango de hospedadores predominante. A pesar de ello, estas variables fueron utilizadas como información auxiliar y no como criterio determinante de clasificación.

No obstante, el desempeño desigual del modelo entre subclusters pone de manifiesto limitaciones inherentes al enfoque. Subclusters con pocos representantes o con arquitecturas de dominios altamente variables presentaron tasas elevadas de error de clasificación, mientras que aquellos con mayor número de genomas y perfiles funcionales más homogéneos fueron correctamente clasificados. Este comportamiento refleja una característica bien conocida de los genomas de bacteriófagos: su naturaleza de mosaico, resultado de recombinación frecuente y transferencia horizontal de módulos funcionales, lo que dificulta la asignación taxonómica basada exclusivamente en la presencia de dominios individuales.

En conjunto, estos resultados indican que, si bien la predicción de subclusters basada en dominios funcionales no reemplaza a los enfoques taxonómicos establecidos, sí permite identificar funciones clave que contribuyen a la diferenciación entre los mismos. El análisis de importancia de variables

aporta además un marco interpretativo valioso para explorar la relación entre arquitectura funcional y organización taxonómica de los genomas fágicos, y sienta las bases para el desarrollo de enfoques alternativos que integren información de orden génico, contexto modular o similitud de secuencias completas.

Por otra parte, es importante mencionar que la división en clusters y subclusters propuesta por el equipo de Oliveira se relaciona directamente con la clasificación de proteínas en distintas familias (Phams) y, muchas de ellas, son proteínas sin ninguna función asignada, razón por la cual no fueron incluidas como variables, pero demuestran la importancia de las proteínas desconocidas en el estudio de bacteriófagos.

Clasificación de fagos nuevos mediante el enfoque automatizado

La capacidad predictiva del clasificador supervisado fue evaluada mediante la asignación de subclusters a un conjunto de 14 fagos adicionales que no habían sido utilizados durante el entrenamiento del modelo (Tabla 8). Estos fagos habían sido previamente clasificados mediante el enfoque comparativo tradicional reportado en la bibliografía, lo que permitió utilizar dicha clasificación como referencia para la validación de las predicciones.

El modelo Random Forest asignó a los 14 fagos analizados los subclusters B2, B3, B5 y B7, en concordancia total con la clasificación de referencia. En todos los casos, la predicción automatizada coincidió con el subcluster determinado mediante el análisis comparativo basado en alineamientos genómicos y redes filogenéticas, lo que indica una tasa de concordancia del 100% para este conjunto de validación.

Tabla 8: Predicción de clusters de fagos propios

Fago *vB_SauS_*	**277g**	**287**	**308**	**320**	**321c**	**690**	**713**	**760**	**775**	**832**	**I73**	**Mh1**	**Mh15**	**Mh4**
Predicción	B2	B3	B7	B5	B2	B5	B2	B3	B2	B2	B3	B3	B3	B2

Los fagos vB_SauS_277g, vB_SauS_321c, vB_SauS_713, vB_SauS_775 y vB_SauS_832 fueron asignados al subcluster B2, mientras que vB_SauS_287, vB_SauS_760, vB_SauS_I73, vB_SauS_Mh1 y vB_SauS_Mh15 fueron clasificados como miembros del subcluster B3. Por su parte, vB_SauS_320 y vB_SauS_690 fueron asignados al subcluster B5, y vB_SauS_308 al subcluster B7. Esta distribución reproduce fielmente la organización subcluster reportada previamente para estos fagos.

Desde un punto de vista metodológico, estos resultados demuestran que la clasificación basada en la presencia y arquitectura de dominios proteicos, complementada con información contextual del

hospedador, es suficiente para recuperar señales filogenéticas y funcionales relevantes a nivel de subcluster. En contraste con los enfoques tradicionales, que requieren alineamientos genómicos completos y análisis manuales extensos, el método automatizado permite asignar nuevos fagos de manera rápida, reproducible y escalable.

Asimismo, la concordancia observada refuerza la interpretación de que los subclusters definidos para fagos de *Staphylococcus* reflejan no solo relaciones evolutivas globales, sino también patrones conservados en la composición funcional de los genomas. En este sentido, el modelo no "memoriza" genomas específicos, sino que captura combinaciones de dominios asociadas a estrategias biológicas compartidas dentro de cada subcluster.

En conjunto, estos resultados validan el enfoque supervisado propuesto como una alternativa robusta y eficiente para la clasificación de fagos nuevos, particularmente útil en contextos donde el número de genomas disponibles continúa expandiéndose y los métodos comparativos clásicos resultan computacionalmente costosos o difíciles de escalar.

Análisis módulo a módulo

Con el objetivo de interpretar los resultados del clasificador supervisado y vincular la importancia de variables con funciones biológicas concretas, se realizó un análisis módulo por módulo de los genomas fágicos. Dado que los bacteriófagos presentan una organización genómica marcadamente modular, donde conjuntos de genes funcionalmente relacionados se agrupan en casetes relativamente conservados, este enfoque permite contextualizar la contribución de los dominios funcionales más informativos identificados por el modelo Random Forest.

El análisis se centró en los principales módulos funcionales previamente definidos: lisogenia, metabolismo del ADN, ARN y nucleótidos, empaquetamiento del ADN, morfogénesis de cápside, morfogénesis de cola y lisis. Para cada módulo se evaluó su grado de conservación, variabilidad funcional y potencial contribución a la discriminación entre subclusters, considerando tanto la presencia de dominios específicos como su recurrencia dentro de determinados linajes.

Casete de Lisogenia

Identificación y agrupamiento de integrasas

A partir de la base de datos de proteínas anotadas, se identificaron 265 secuencias de integrasas. Las secuencias aminoacídicas fueron alineadas mediante alineamiento múltiple y utilizadas para inferir una matriz de distancias aminoacídicas, a partir de la cual se construyó un dendrograma jerárquico que resume las relaciones de similitud global entre las integrasas analizadas.

Con el objetivo de definir un esquema de agrupamiento robusto y reproducible, se realizó una exploración sistemática del punto de corte del dendrograma, evaluando la relación entre la altura de corte, el número de clusters resultantes y el tamaño mínimo de los mismos. Este análisis permitió identificar un compromiso adecuado entre resolución evolutiva y representatividad de los grupos. En base a estos criterios, se seleccionó un esquema final de 13 clusters de integrasas, que permitió capturar la diversidad observada manteniendo tamaños de grupo comparables, aunque incluyendo algunos clusters representados por una única secuencia. El dendrograma resultante evidenció una clara estructuración jerárquica de las integrasas, con grupos bien definidos y separaciones consistentes con la divergencia evolutiva entre secuencias (Fig. 16). Estos clusters (Material Suplementario: 05_integrasas_grupos.xlsx) constituyen unidades evolutivamente coherentes y fueron utilizados como base para la clasificación supervisada desarrollada en las fases posteriores del análisis.

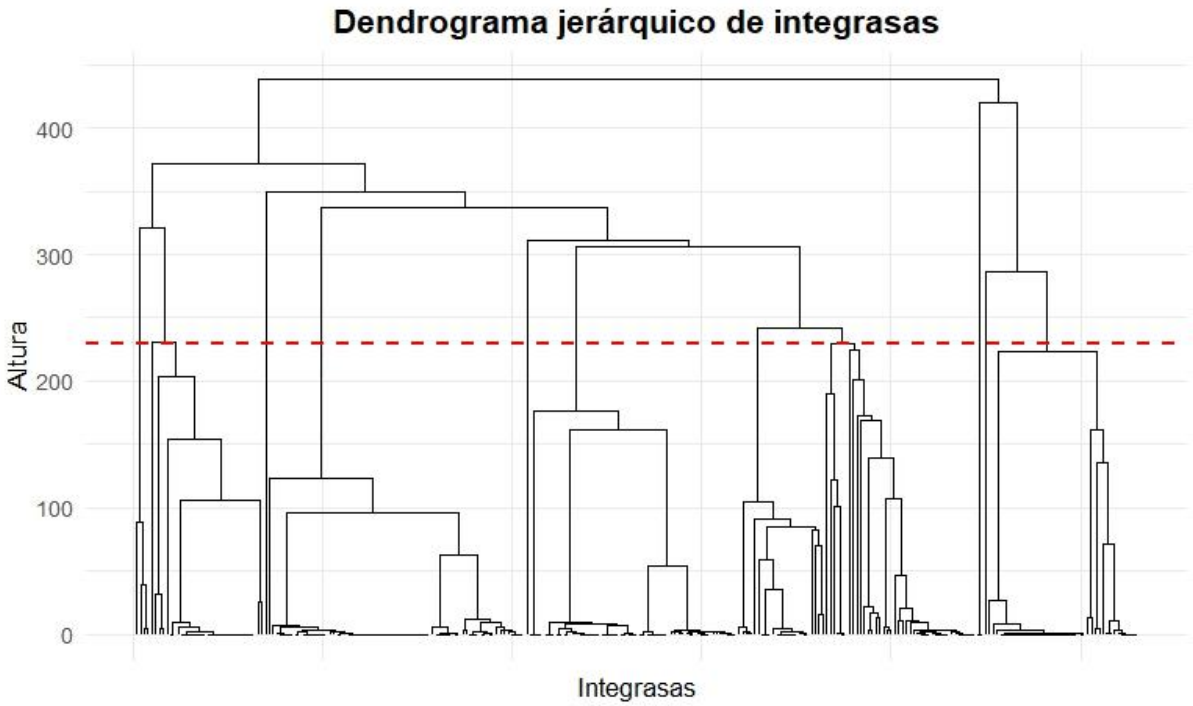

Figura 16: Dendrograma de clustering jerárquico de integrasas basado en distancias de aminoácidos. La línea roja discontinua indica la altura de corte seleccionada para la definición de los clústeres evolutivos.

Goerke et al. (47) propusieron la agrupación de integrasas de *Estafilococos* en distintos grupos SaInt, donde los grupos Sa1int a presentan un mayor número de representantes, mientras que los grupos Sa8Int a Sa12Int se encuentran representados por uno o muy pocos individuos. En dicho esquema, cada grupo SaInt se asocia de manera consistente a una arquitectura específica de dominios Pfam, lo que sugiere una fuerte correspondencia entre organización modular y clasificación funcional (Tabla 9).

Tabla 9: Presencia de tipo y grupo de integrasas en los fagos reportados por Goerke

Integrasa	Dominio Pfam	Grupo	Fagos reportados	Cluster asignado
S-Int	PF00239_PF07508	Sa7Int	53, 80-alpha, 85, 92, phiNM2, X2	7
		Sa12Int	phiMR11	7
T-Int	PF00589	Sa6Int	52a, 77, 80, phiNM4, ROSA, tp310-2	1
		Sa4Int	PhiSauS-IPLA35	1
		Sa2Int	47, phi2958PVL, PVL, PVL108, tp310-1	1
		Sa11Int	EW	5
	PF13102_PF00589	Sa9Int	96	3
		Sa8Int	phiSauS-IPLA88	4
		Sa10Int	37	5
	PF14657_PF14659_PF00589	Sa1Int	55, 71, phiETA, phiETA2, phiETA3	6
		Sa5Int	187, 29, 69, 88, phi11, phiMR25, phiNM1	6
	PF14659_PF00589	Sa3Int	42e, phi13, phiNM3, phiN315, tp310-3	5

Sin embargo, al incorporar el esquema de clustering definido en este trabajo (Tabla 10), se observa que no existe una correspondencia unívoca entre los clusters obtenidos y una única arquitectura de dominios Pfam. En particular, algunos clusters agrupan integrasas con arquitecturas de dominios diferentes. Por ejemplo, el cluster 5 incluye integrasas con las arquitecturas “PF00589” y “PF14659_PF00589”, a pesar de sus diferencias en organización modular y, el cluster 11 posee también la arquitectura “PF00589”. Estos resultados indican que, si bien la arquitectura de dominios es altamente informativa, no resulta suficiente por sí sola para reflejar la totalidad de las relaciones evolutivas entre integrasas. El enfoque basado en similitud global de secuencia permite capturar señales evolutivas adicionales, posiblemente asociadas a regiones conservadas fuera de los dominios anotados, o a eventos de recombinación y reordenamiento modular.

Tabla 10: Caracterización de los grupos de integrasas según el número de individuos, el tipo de integrasa y la arquitectura de dominios funcionales.

Cluster	N° de individuos	Integrasa	Dominio Pfam
1	68	T-Int	PF00589
			PF14659_PF00589
2	4	T-Int	PF00589
3	23	T-Int	PF13102_PF00589
			PF14657_PF13102_PF00589
4	6	T-Int	PF13102_PF00589
			PF14657_PF13102_PF00589
			PF14659_PF00589
5	34	T-Int	PF00589
			PF13102_PF00589
			PF14657_PF14659_PF00589
			PF14659_PF00589
6	55	T-Int	PF14657_PF14659_PF00589
7	40	T-Int	PF00239_PF07508
			PF00239_PF07508_PF13408
8	1	S-Int	PF00239_PF07508_PF13408
9	2	S-Int	PF00239_PF07508
10	29	T-Int	PF02899
			PF02899_PF00589
11	1	T-Int	PF00589
12	1	T-Int	PF14657_PF13102_PF00589
13	1	T-Int	PF14657_PF14659_PF00589

Esta interpretación se ve reforzada al comparar el clustering obtenido con el análisis filogenético presentado por Goerke et al., donde la asignación de los grupos SaInt depende del criterio del operador y del punto de corte seleccionado en el árbol filogenético (Figura 17). En contraste, el enfoque adoptado en este trabajo utiliza un criterio sistemático y reproducible para la definición de clusters. Como consecuencia, algunos grupos SaInt filogenéticamente cercanos —como Sa2Int, Sa4Int

y Sa6Int— se agrupan dentro de un mismo cluster en el esquema propuesto (cluster 1), reflejando su alta similitud de secuencia a escala global.

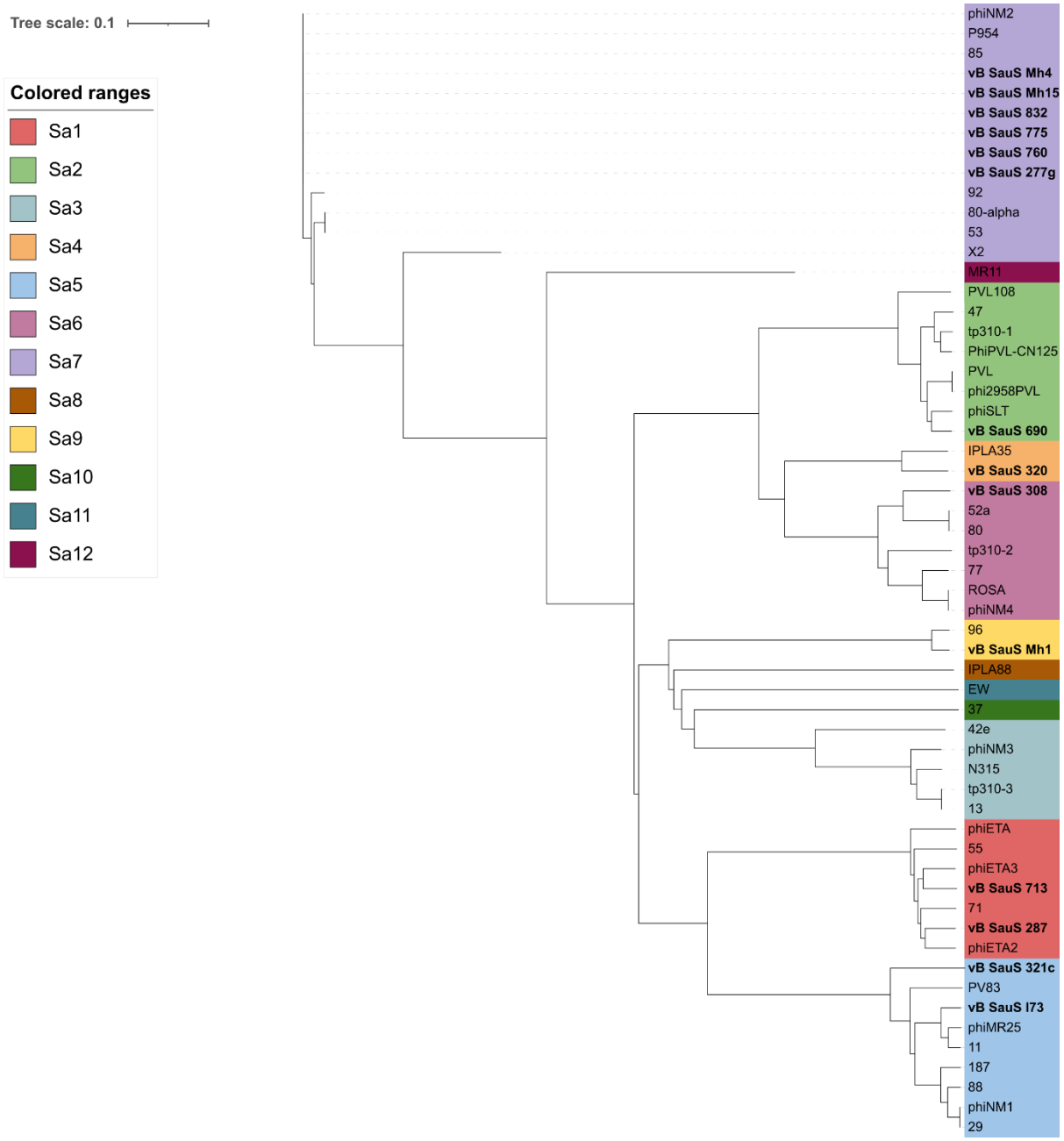

Figura 17: Análisis filogenético de integrasas

Los clusters de integrasas definidos en la Fase I constituyen grupos evolutivamente coherentes derivados de un análisis no supervisado basado exclusivamente en la similitud global de secuencia. Estos grupos proporcionan un marco de referencia objetivo que permite transformar el problema de clasificación de integrasas en un enfoque supervisado, donde cada secuencia puede ser representada cuantitativamente en función de su relación con los clusters previamente definidos. En este contexto,

los clusters obtenidos se utilizaron como clases de entrenamiento para el desarrollo de un clasificador automático destinado a la asignación de nuevas integrasas a los grupos definidos.

Análisis supervisado

Debido al tamaño limitado del conjunto de datos y a la elevada similitud existente entre muchas de las integrasas analizadas, no se realizó una partición fija en conjuntos de entrenamiento y prueba. En su lugar, la capacidad de las características definidas para discriminar entre los clusters fue evaluada mediante un enfoque supervisado utilizando un clasificador k-nearest neighbors (kNN) combinado con validación cruzada Leave-One-Out (LOOCV). Este esquema permite maximizar el uso de la información disponible y resulta particularmente adecuado para conjuntos de datos con clases de tamaño reducido o desbalanceadas.

Cada integrasa fue representada mediante un vector de características construido a partir de su distancia evolutiva promedio a cada uno de los clusters definidos en la Fase I, calculada a partir de la matriz de distancias aminoacídicas globales. En este espacio de características, la clase asociada a cada instancia correspondió al cluster asignado previamente en la Fase I, permitiendo evaluar en qué medida dicha representación captura información suficiente para discriminar entre los distintos grupos.

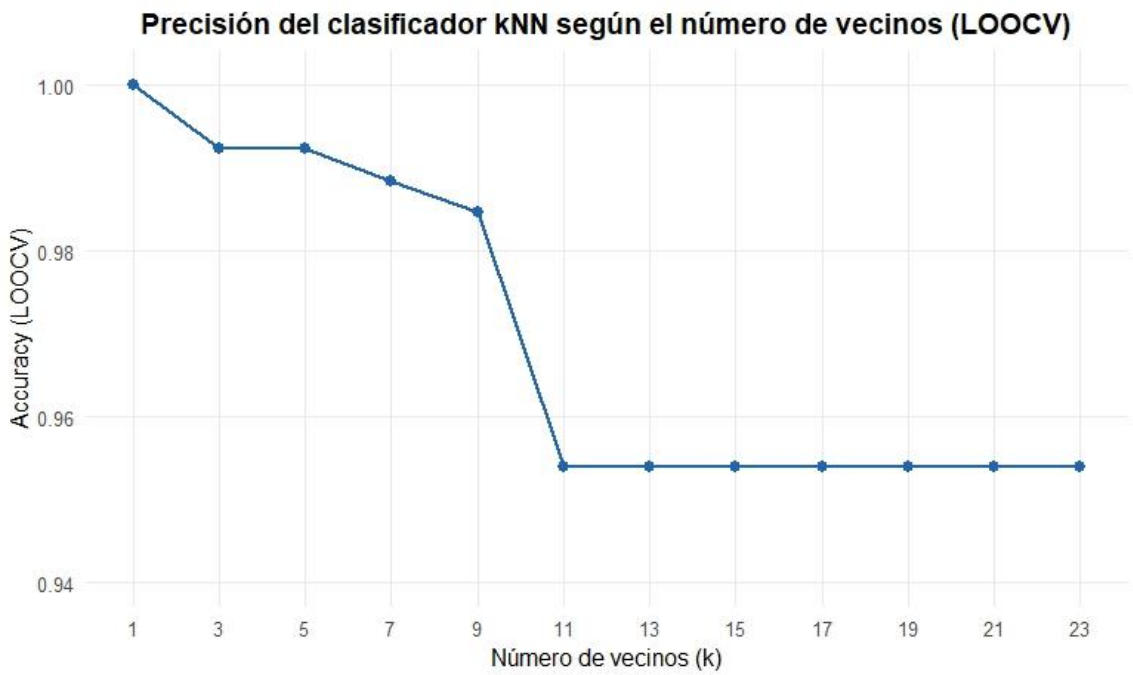

Figura 18: Evaluación del desempeño del clasificador k-Nearest Neighbors (kNN) mediante validación cruzada leave-one-out (LOOCV) para distintos valores del parámetro k.

La performance del clasificador kNN fue evaluada explorando distintos valores del parámetro k (Fig. 18). Para valores pequeños de k, el modelo mostró un desempeño elevado, con una precisión máxima

del 100 % para k = 1 y valores de precisión muy similares para k = 3 y k = 5 (accuracy = 0.992). A medida que k aumentó, se observó una disminución progresiva de la precisión global, alcanzando valores cercanos a 0.954 para k ≥ 11, lo que sugiere una pérdida de resolución asociada a la incorporación de vecinos más lejanos en el espacio de distancias.

Si bien el valor k = 1 presentó la mayor precisión, se seleccionó k = 3 como valor óptimo, ya que proporciona un mejor equilibrio entre sensibilidad a la estructura local del espacio de características y robustez frente a posibles efectos de ruido o asignaciones espurias. Este criterio permitió evitar el sobreajuste potencial asociado a valores extremadamente bajos de k, manteniendo al mismo tiempo una precisión global elevada.

La matriz de confusión obtenida a partir de la validación LOOCV (Fig. 19) mostró un desempeño casi perfecto del clasificador. La precisión global alcanzada fue del 99.23 % (IC 95 %: 97.26–99.91 %), con un valor de Kappa de 0.99, lo que indica un acuerdo casi perfecto entre las clases predichas y las clases de referencia. La mayoría de los clusters fueron clasificados sin errores, con valores de sensibilidad y especificidad iguales a 1.0.

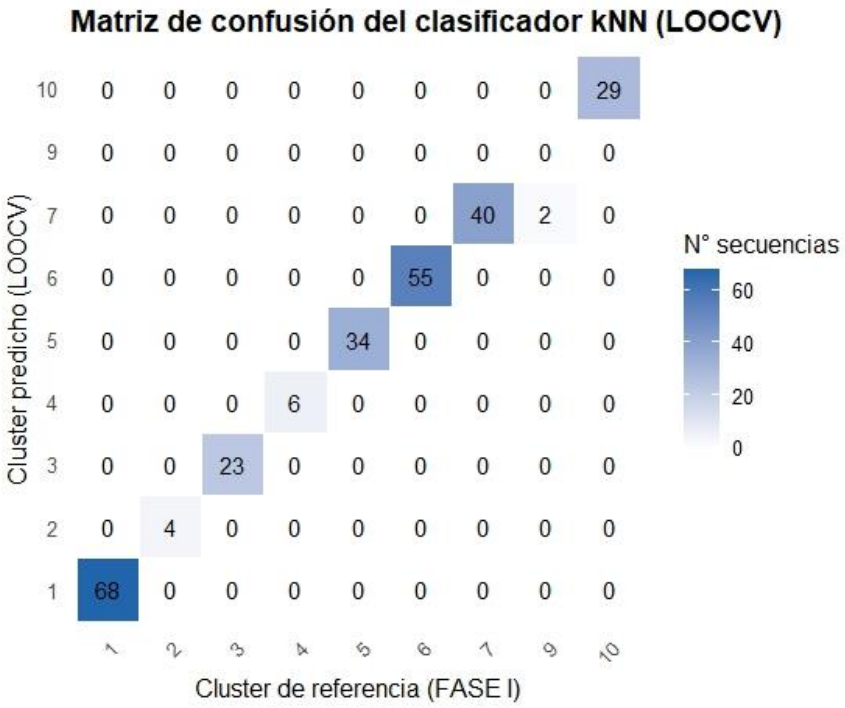

Figura 19: Matriz de confusión obtenida mediante validación cruzada leave-one-out (LOOCV) del clasificador k-Nearest Neighbors (k = 3), entrenado sobre el espacio de distancias aminoacídicas entre integrasas de referencia

El único error observado correspondió a ambas integrasas del cluster 9 que fueron clasificadas como integrantes del cluster 7. Este resultado es consistente con la elevada similitud evolutiva entre ambos grupos y con el reducido tamaño del cluster 9, lo que limita la capacidad del clasificador para identificar

vecinos cercanos representativos en el espacio de características. No obstante, la ausencia de errores sistemáticos y el alto desempeño general del modelo indican que las distancias promedio a clusters constituyen una representación adecuada para la clasificación de integrasas.

El modelo entrenado con k = 3 fue almacenado y utilizado como clasificador de referencia para la asignación automática de nuevas integrasas a los grupos ya definidos, y aplicado en la fase siguiente para el análisis de integrasas provenientes de genomas fágicos no caracterizados.

Predicción de clústeres en integrasas de nuevos fagos

Con el objetivo de asignar automáticamente grupos a integrasas no caracterizadas previamente, se aplicó el modelo supervisado kNN entrenado sobre el conjunto completo de integrasas agrupadas en 13 clusters (Fases I y II). El modelo final, validado mediante validación cruzada leave-one-out (LOOCV), utiliza como variables predictoras las distancias medias de cada integrasa a los clusters definidos a partir del análisis no supervisado.

El clasificador fue aplicado a las integrasas correspondientes a 14 fagos adicionales, que no habían sido utilizadas durante la etapa de entrenamiento del modelo. Para cada integrasa, se calculó su vector de características en el mismo espacio de distancias empleado para el entrenamiento y se procedió a su clasificación utilizando el valor óptimo del parámetro k (k = 3), seleccionado en base al desempeño del modelo durante la fase de validación.

El modelo permitió asignar cada una de las integrasas analizadas a uno de los grupos definidos, proporcionando una clasificación consistente con la estructura jerárquica observada en el dendrograma de referencia. Las asignaciones obtenidas reflejaron la proximidad evolutiva de las nuevas integrasas a los grupos previamente definidos, reforzando la validez biológica de los clusters y su capacidad para generalizar a secuencias no incluidas en el conjunto de entrenamiento.

En todos los casos analizados, las integrasas no anotadas presentaron una cercanía dominante a un único grupo SaInt, reflejada en valores de confianza máximos del clasificador (*confianza*= 1). Esto indica que los tres vecinos más próximos en el espacio de distancias (k=3) pertenecieron de manera consistente a un mismo clúster, resultando en predicciones inequívocas y sin evidencia de asignaciones ambiguas o intermedias entre grupos.

El conjunto de las integrasas identificadas en los fagos analizados correspondieron a integrasas del tipo serina (S-Int) o tirosina (T-Int), confirmando la naturaleza temperada de estos fagos. Todas las integrasas de nuestros fagos se encuentran en alguno de los grupos más numerosos (1, 6, 7 o 3) (Tabla 11) diferenciándose de la clasificación propuesta por Goerke et. al. (Fig. 17) donde las integrasas de vB_SauS_Mh1 y vB_Saus_320 se ubicaban en singletons (Sa9Int y Sa4Int respectivamente).

Tabla 11: Resultados de la predicción del grupo de integrasas y su relación con los grupos según Goerke.

Fago	Cluster predicho	Confianza	Grupo Goerke
vB_SauS_277g	7	1	Sa7Int
vB_SauS_287	6	1	Sa1Int
vB_SauS_308	1	1	Sa6Int
vB_SauS_320	1	1	Sa4Int
vB_SauS_321c	6	1	Sa5Int
vB_SauS_690	1	1	Sa2Int
vB_SauS_713	6	1	Sa1Int
vB_SauS_760	7	1	Sa7Int
vB_SauS_775	7	1	Sa7Int
vB_SauS_832	7	1	Sa7Int
vB_SauS_I73	6	1	Sa5Int
vB_SauS_Mh1	3	1	Sa9Int
vB_SauS_Mh15	7	1	Sa7Int
vB_SauS_Mh4	7	1	Sa7Int

El hecho de que todas las integrasas analizadas hayan sido clasificadas con valores máximos de confianza indica que estas secuencias no ocupan posiciones intermedias o ambiguas en el espacio de distancias, sino que se integran claramente dentro de linajes previamente definidos. Esto sugiere que los grupos capturan señales evolutivas robustas que trascienden diferencias menores de secuencia y que son conservadas incluso en integrasas provenientes de fagos no caracterizados previamente.

Asimismo, la alta homología observada entre las integrasas de los fagos analizados y las secuencias de referencia refuerza la hipótesis de que estos fagos comparten estrategias lisogénicas similares, coherentes con su clasificación como fagos temperados. En este contexto, la asignación a grupos específicos puede considerarse un proxy de propiedades biológicas conservadas, tales como la especificidad del sitio *att*, el tipo de integrasa (serina o tirosina) y la organización del módulo de lisogenia.

En conjunto, estos resultados indican que el enfoque basado en distancias promedio a clusters evolutivamente definidos no solo reproduce clasificaciones previamente reportadas, sino que además evita la fragmentación artificial en singletons cuando existen relaciones de similitud global robustas. Este comportamiento resulta particularmente relevante para la clasificación de nuevos fagos, donde

el tamaño reducido del conjunto de referencia puede sesgar enfoques puramente filogenéticos o basados en dominios individuales.

Empaquetamiento del ADN y morfogénesis de cápside

El módulo de empaquetamiento del ADN y de la cápside incluye las proteínas responsables del reconocimiento, procesamiento y translocación del genoma viral hacia el interior de las cápsides previamente ensambladas, constituyendo un paso esencial para la formación de partículas fágicas infectivas. En todos los genomas analizados, este módulo contiene al menos una terminasa de subunidad grande (TerL), acompañada por proteínas asociadas a la cápside y, en algunos casos, por una proteasa de procápside.

Análisis de proteínas específicas

El análisis de la organización génica y de la arquitectura de dominios de las proteínas TerL permitió inferir el mecanismo de empaquetamiento del ADN. En 11 de los 14 fagos estudiados, las TerL presentan dominios PF03237 o la combinación PF04466–PF17288, junto con longitudes relativamente homogéneas (≈408–448 aminoácidos) (Tabla 12). En estos fagos no se detectaron proteasas de procápside asociadas al módulo de empaquetamiento. Este conjunto de características es consistente con un mecanismo de empaquetamiento tipo *pac*.

Tabla 12: Características de las terminasas largas de los 14 fagos temperados

Fago	Longitud TerL (aa)	Dominios funcionales (TerL)	Dominios funcionales (proteasa)	Estrategia de empaquetamiento inferida
vB_SauS_287 **vB_SauS_760** **vB_SauS_I73** **vB_SauS_Mh1** **vB_SauS_Mh15**	408	PF03237	-	TerL P22-like, *pac*
vB_SauS_713	426	PF04466_PF17288		
vB_SauS_277g	430			
vB_SauS_321c **vB_SauS_775** **vB_SauS_832** **vB_SauS_Mh4**	448			
vB_SauS_308	565	PF04586	PF03354_PF20441	TerL HK97-like, cos

vB_SauS_320 **vB_SauS_690**	567	PF00574		

En contraste, los fagos vB_SauS_320, vB_SauS_690 y vB_SauS_308 presentan TerL de mayor longitud (≈565–567 aminoácidos), con dominios PF03354–PF20441, típicamente asociados a sistemas de empaquetamiento tipo *cos*. En estos genomas, el módulo de empaquetamiento incluye además una serina proteasa de la procápside, una proteína característica de fagos con extremos cohesivos. En el caso de vB_SauS_308, se identificó una serina proteasa de la procápside (gp50) dentro del módulo de empaquetamiento; este tipo de proteasa es característica de los fagos de tipo *cos*.

Organización general del módulo

La organización del módulo de empaquetamiento mostró un patrón conservado según pertenezcan a uno u otro mecanismo de empaquetamiento, mostrando dos subgrupos para cada uno (Fig. 20).

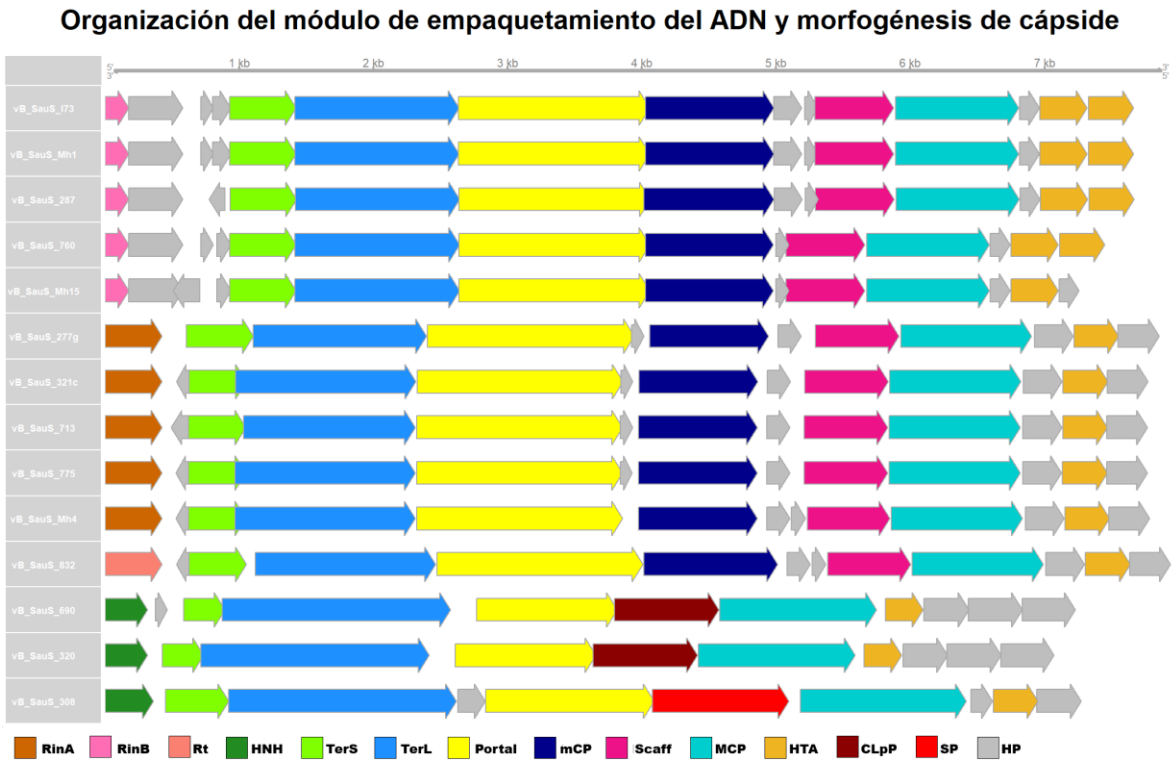

Figura 20: Organización del módulo de empaquetamiento del ADN y morfogénesis de cápside

Los fagos clasificados como tipo *pac* presentan una organización más módulos de empaquetamiento compactos y altamente conservados, mientras que los fagos tipo cos exhiben una arquitectura génica diferenciada, con la presencia adicional de la proteasa de procápside y terminasa de mayor tamaño.

Las inferencias obtenidas a partir del análisis de dominios, longitudes proteicas y organización génica del módulo de empaquetamiento (Método 2) resultaron concordantes con las predicciones basadas en el análisis filogenético de las proteínas TerL (Método 1), reforzando la asignación de los mecanismos de empaquetamiento dentro de los diferentes clústeres de fagos analizados.

Casete de morfogénesis de cola

Con el objetivo de caracterizar el módulo de reconocimiento al hospedador de los fagos analizados, se examinó la organización génica del módulo de morfogénesis de cola, haciendo especial énfasis en la posición, diversidad y relaciones de las proteínas de medición de cola (Tape Measure Proteins, TMPs) y de las Proteínas de Unión al Receptor (Receptor Binding Proteins, RBPs).

Organización general

El análisis comparativo de los genomas reveló una organización altamente conservada en todos los fagos estudiados. En cada genoma, este casete se extiende desde el gen que codifica para una proteína de cola no específica (dominio PF04883 o TIG1725 en vB_SauS_308) hasta inmediatamente antes del casete de lisis, manteniendo un orden génico similar entre fagos (Fig. 21).

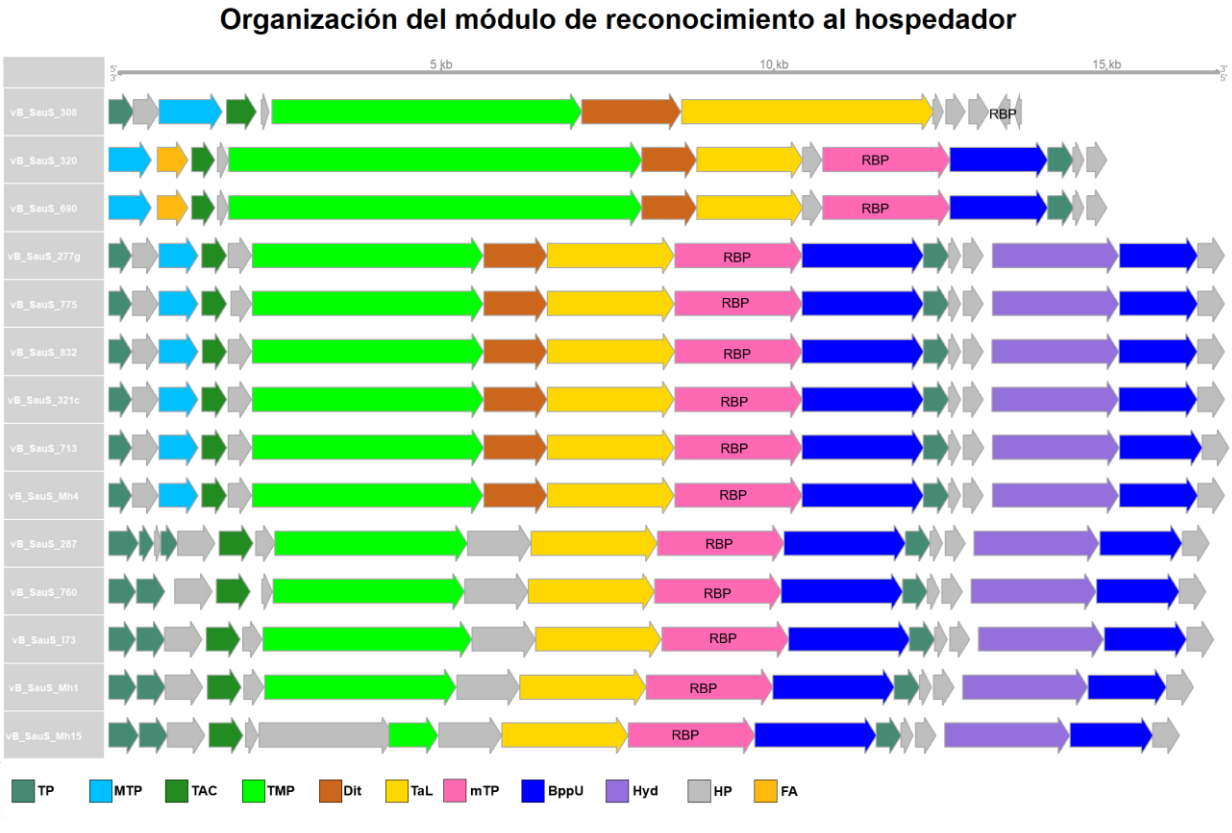

Figura 21: Organización del módulo reconocimiento del hospedador

Podemos observar 4 grupos bien definidos, en concordancia a los hallazgos del Método 1, donde: el fago vB_SauS_308 tiene una organización extremadamente disímil al resto de los grupos y, los fagos vB_SauS_320 y vB_SauS_690 tienen a su comienzo los genes que codifican la proteína mayor de cola (MTP) seguidas por una proteína con función adicional (domino Ig-like protein), no poseen ningún dominio con función hidrolasa y solo una proteína superior de la placa base (BppU). Siendo relevante, en ambos grupos, la diferencia en la longitud de la TMP coincidente con arquitecturas de dominios funcionales distintas a otros grupos (Tabla 13). Los dos grupos restantes, formados por los representantes de fagos vB_SauS_277g y vB_SauS_287, son más similares entre ellos pero con diferencias claves: tanto la MTP como la proteína distal (Dit) del grupo vB_SauS_287 no fueron reconocidas con la anotación en base a dominios funcionales y, por lo tanto, figuran como proteína hipotéticas, aunque por contexto se podría decir que se encuentran 3 genes corriente arriba de la TMP y un gen corriente abajo respectivamente; además, el segundo gen del módulo que codifica para un proteína de cola tiene un dominio PF11367 a diferencia del grupo de vB_SauS_277g que no se encontró ningún dominio conocido.

A pesar de pequeñas variaciones en longitud y contenido génico, el número total de proteínas codificadas dentro del casete de cola fue notablemente similar entre los distintos genomas, oscilando entre 13 y 18 proteínas por fago. En particular, la mayoría de los fagos presentó entre 15 y 16 proteínas, lo que sugiere una fuerte conservación estructural del módulo. El fago vB_SauS_308 es el que menos genes posee, careciendo de BppU, pero tiene, sin embargo, una proteína distal (DiT) y una lisina (TaL) particularmente largas en comparación al resto de los fagos.

Un rasgo recurrente fue que las últimas proteínas del casete, ubicadas inmediatamente antes del inicio del módulo de lisis, se anotaron sistemáticamente como proteínas hipotéticas, sin dominios funcionales conocidos. Esta observación fue consistente en todos los genomas analizados y sugiere un posible rol estructural o regulatorio aún no caracterizado dentro del módulo de cola.

Es relevante mencionar que, debido a la anotación no curada, la TMP del fago Mh15 parecería estar entrecortada, así como del segundo al cuarto gen de vB_SauS_287.

Tape Measure Proteins (TMPs)

En todos los fagos analizados se identificó una única TMP por genoma, localizada en una posición conservada dentro del casete. Sin embargo, estos genes mostraron una alta variabilidad en longitud, con tamaños que oscilaron entre aproximadamente 700 (probable error de anotación en vB_SauS_Mh15) y 6200 nucleótidos (Figura 22).

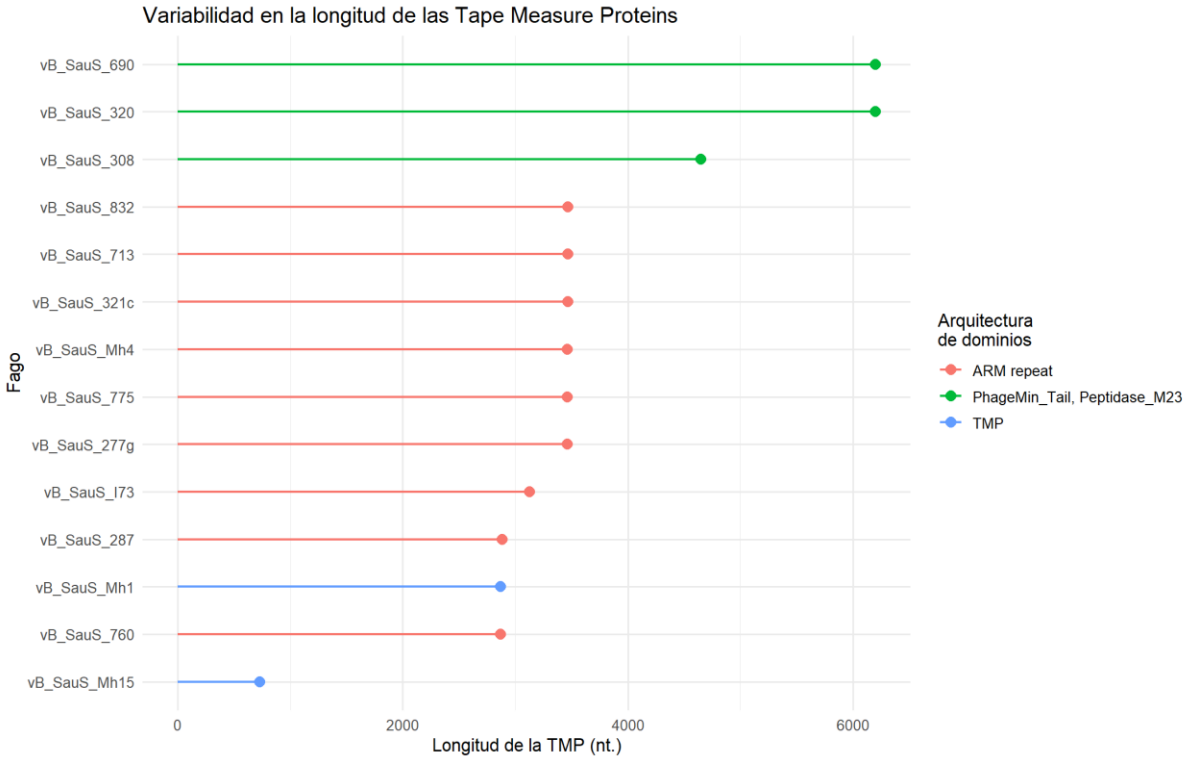

Figura 22: Variabilidad en la longitud de las TMPs por fago

El análisis de dominios Pfam reveló la presencia de dos arquitecturas principales de TMPs. Un subconjunto de fagos codificó TMPs largas que contienen los dominios PhageMin_Tail (PF10145) y Peptidase_M23 (PF01551) (fagos vB_SauS_308, vB_SauS_320 y vB_SauS_690), mientras que otro grupo presentó TMPs más cortas enriquecidas en repeticiones ARM (SSF48371). En dos casos particulares se identificaron TMPs cortas con el dominio PF05017, previamente asociado a proteínas de medición de cola en otros fagos (Tabla 13).

Tabla 13: Análisis de las TMPs de fagos del laboratorio

Fago	Id Proteína	Longitud (nt.)	Dominios	Nombres dominios
vB_SauS_320	58	6200	PF10145_PF01551	PhageMin_Tail, Peptidase_M23
vB_SauS_690	59			
vB_SauS_308	64	4649		
vB_SauS_321c	57	3467	SSF48371	ARM repeat
vB_SauS_713	54	3467		
vB_SauS_832	64	3467		
vB_SauS_277g	56	3464		
vB_SauS_775	59	3464		
vB_SauS_Mh4	62	3464		

vB_SauS_I73	60	3128		
vB_SauS_287	62	2885		
vB_SauS_760	62	2870		
vB_SauS_Mh1	64	2870	PF05017	TMP
vB_SauS_Mh15	67	731		

Estas diferencias en longitud y composición de dominios sugieren la existencia de estrategias estructurales alternativas para la determinación de la longitud de la cola, manteniendo no obstante una posición genómica conservada.

Identificación de las RBPs candidatas

Dado que las RBPs suelen estar pobremente anotadas a nivel funcional, se adoptó una estrategia posicional para su identificación. Para cada genoma, se definió una región candidata a RBPs comprendida entre la TMP y el fin del casete. Dentro de esta región se extrajeron todas las proteínas codificadas, obteniéndose un conjunto de 133 proteínas candidatas distribuidas entre los distintos fagos.

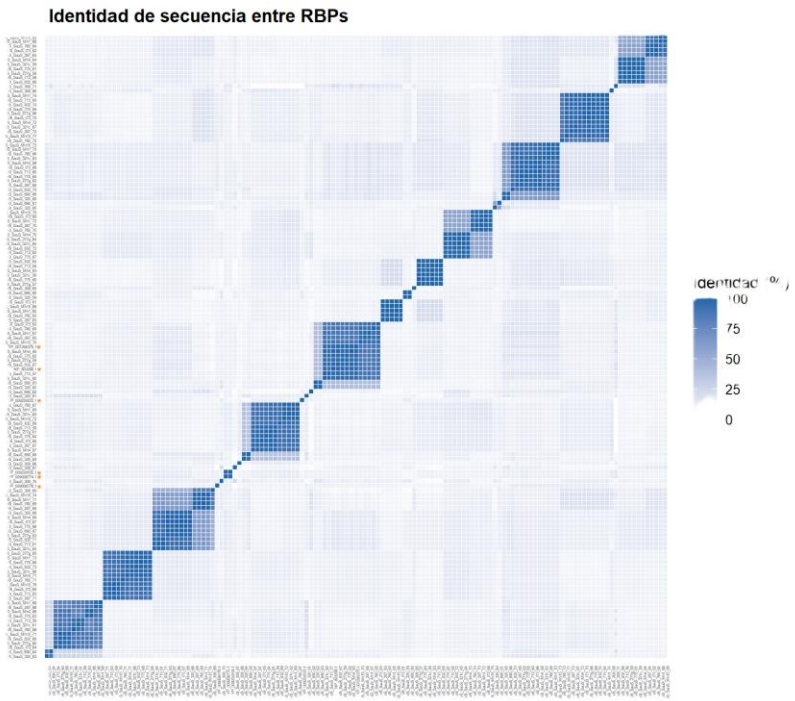

Figura 23: Heatmap de identidad porcentual (PID) entre RBPs candidatas y RBPs de referencia.

Con un"" se muestran las RBPs de referencia*

Para evaluar la relación evolutiva entre las RBPs candidatas y RBPs previamente caracterizadas, se realizó un alineamiento múltiple conjunto que incluyó tanto las proteínas candidatas como un conjunto de RBPs de referencia descritas en la literatura. A partir de este alineamiento se calculó la identidad porcentual par-a-par (PID) (Anexo Tabla 17) y se construyó un heatmap de similitud (Fig.23).

El análisis del heatmap reveló la formación de nueve grandes grupos principales, en los cuales varias RBPs candidatas se agruparon estrechamente con RBPs de referencia específicas (las dos RBPs conocidas de Φ11 y80α). En particular, esas dos proteínas de referencia mostraron altos valores de identidad (>85–95%) con las proteínas menores de cola (mTP) de once fagos nuestros, indicando una fuerte conservación de secuencia. Resulta llamativo que, los tres fagos restantes son vB_SauS_320, vB_SauS_690 y vB_SauS_308, los mismos que muestran una organización genómica muy disímil a los fagos restantes y TMPs también distintas. En esos casos, el porcentaje de identidad fue de 40.54% para vB_SauS_320 y vB_SauS_690 y 20% para vB_SauS_308 (Fig. 24). A su vez, vB_SauS_308 es el único fago cuya RBP no se encuentra en una mTP sino que, además, es una proteína hipotética cuya distancia a la TMP es el doble (en nt.) que las otras RBPs.

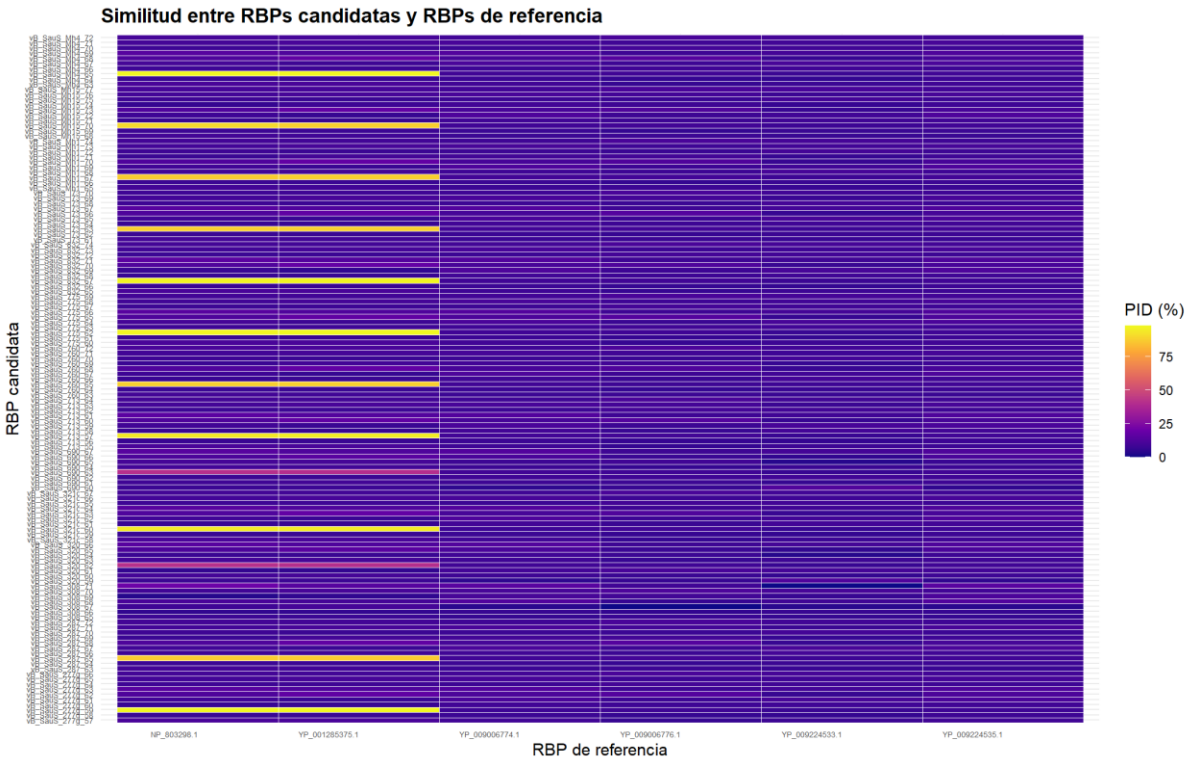

Figura 24: Porcentaje de identidad de secuencia entre RBPs de referencia y candidatas

Módulo de Lisis

Los bacteriófagos dependen de la expresión coordinada de los genes del módulo de lisis —holina y endolisina— para liberar exitosamente las partículas virales producidas durante el ciclo replicativo. La

divergencia de secuencia de dichos genes o productos génicos ha sido utilizada como criterio para clasificar fagos estafilocócicos Las holinas son proteínas formadoras de poros que se insertan en la membrana citoplasmática y permiten la translocación de las endolisinas, enzimas con actividad hidrolasa que degradan el peptidoglucano bacteriano. Las endolisinas se caracterizan por poseer un dominio de unión a la pared celular (CBD) y uno o más dominios enzimáticamente activos (EAD) con actividad lítica que actúan sobre enlaces peptídicos o glicosídicos del peptidoglucano.

Análisis estructural preliminar del módulo de lisis

En una primera aproximación (Método 1), se analizaron las holinas codificadas por los fagos estudiados, observándose longitudes génicas que oscilaron entre 255 y 438 nucleótidos, con valores discretos de 438, 303, 276 y 255 nt. Estos tamaños coinciden con los polimorfismos de longitud previamente reportados para holinas de fagos estafilocócicos (486, 438, 435, 423, 303, 276, 273, 255 y 216 nt), lo que respalda la correcta identificación de este componente del módulo de lisis.

En paralelo, el análisis manual de las endolisinas basado en arquitectura de dominios permitió una clasificación inicial en cinco tipos funcionales (Tipos I a V). Sin embargo, durante esta etapa se detectaron ambigüedades asociadas a la anotación automática de dominios Pfam, particularmente en el reconocimiento de dominios tipo SH3 en CBDs no homólogos, lo que dificultó la discriminación directa entre ciertos tipos de endolisinas.

Curación del conjunto de datos y entrenamiento del clasificador SVM (Método 2)

Para resolver estas limitaciones, se desarrolló un clasificador supervisado basado en Support Vector Machines (SVM), entrenado sobre un conjunto curado de 295 endolisinas. El modelo utilizó un kernel Gappy Pair (k = 1, m = 2), adecuado para capturar patrones locales de secuencia aminoacídica relevantes para la función. Dado el fuerte desbalance entre clases (Fig. 25), se implementó un esquema de pesos de clase, permitiendo penalizar diferencialmente los errores en los grupos minoritarios y evitar un sesgo hacia las clases más representadas.

El conjunto de datos fue dividido aleatoriamente en un 70% para entrenamiento (N = 206) y un 30% para prueba (N = 89). El desempeño del modelo se evaluó mediante múltiples métricas complementarias, mostrando un rendimiento global elevado, con una exactitud (accuracy) del 98,9% y una exactitud balanceada (balanced accuracy) del 95%, indicando un desempeño robusto incluso en presencia de clases minoritarias. El coeficiente de correlación de Matthews (MCC = 0,981) confirmó la alta calidad de la clasificación multiclase.

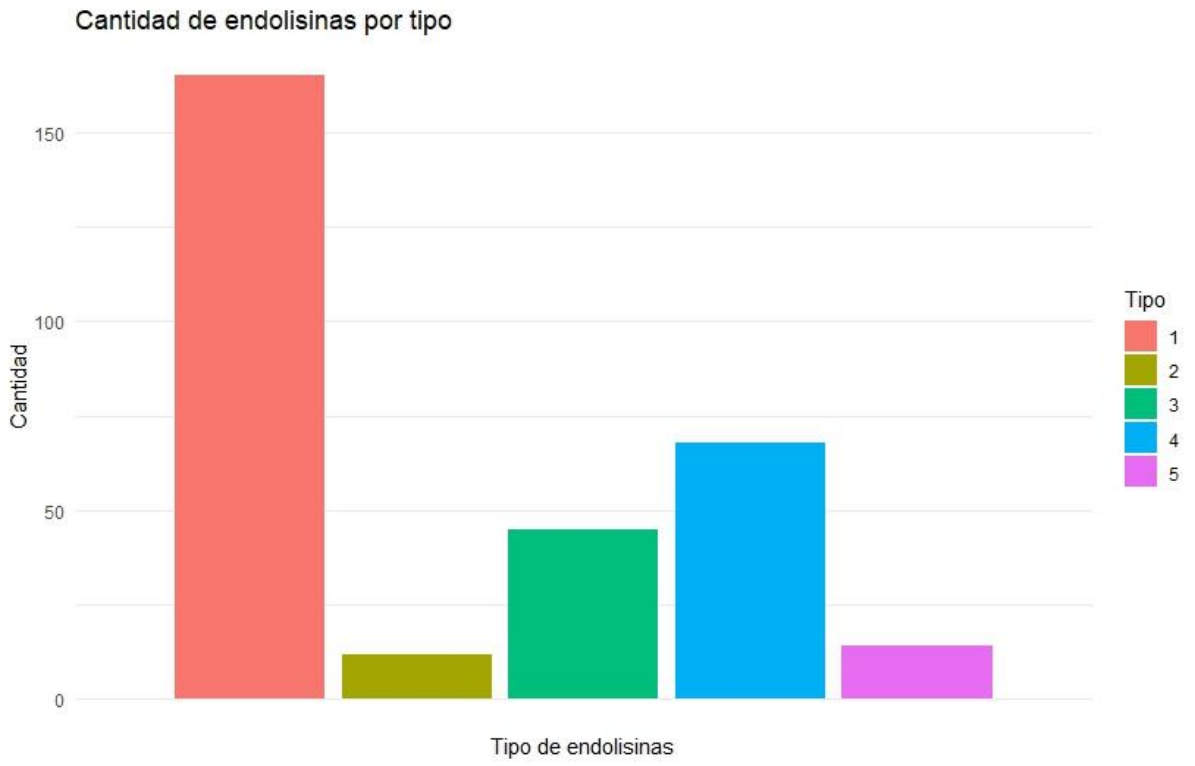

Figura 25: Cantidad de endolisinas por tipo (Tipo_I a Tipo_V, se excluye Tipo_VI del análisis)

El análisis por clase reveló sensibilidades del 100% para los grupos 1, 3, 4 y 5, mientras que el grupo 2 —el menos representado en el dataset— presentó una sensibilidad del 75%, aunque con una especificidad del 100%. Estas métricas reflejan que el uso de pesos de clase permitió mitigar eficazmente el desbalance sin sacrificar la precisión global del modelo.

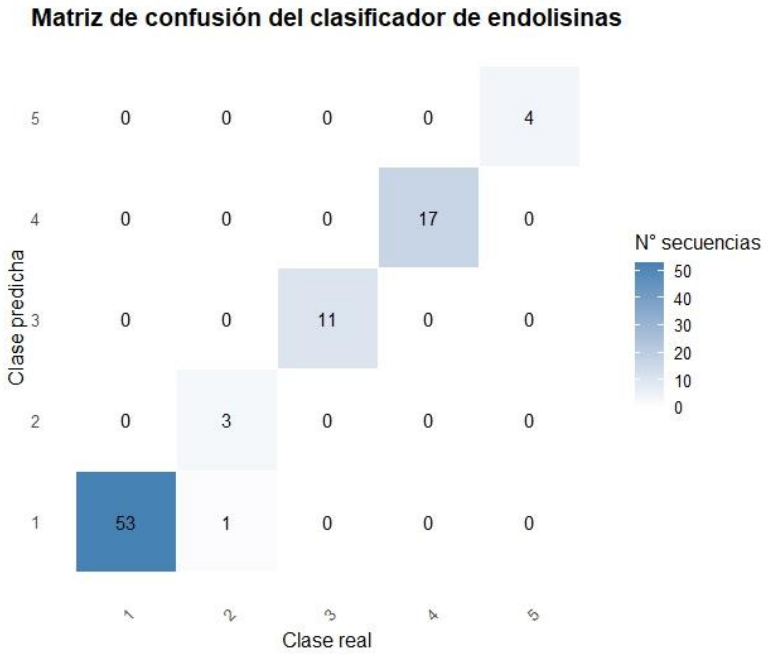

Figura 26: Matriz de confusión del clasificador de endolisinas

La figura 26 muestra la matriz de confusión del clasificador SVM, donde se observa una correspondencia casi perfecta entre las clases reales y las predichas, con un número mínimo de errores concentrados en los grupos minoritarios.

Por su parte, la figura 27 presenta la distribución de longitudes proteicas de las endolisinas utilizadas durante la curación del dataset. Este análisis permitió identificar y excluir secuencias con longitudes atípicas(<200aa.), asegurando la homogeneidad estructural de los grupos y contribuyendo a la estabilidad del modelo.

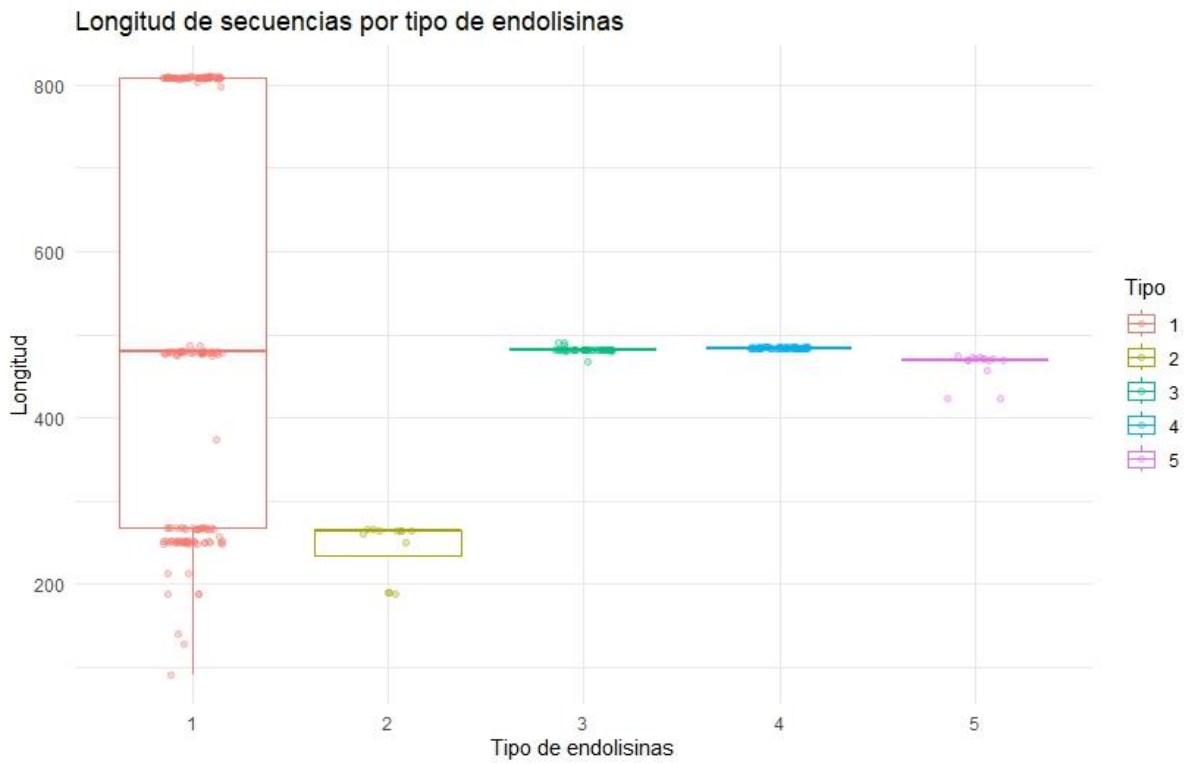

Figura 27: Longitud de secuencias de aminoácidos de endolisinas según el tipo

Predicción de endolisinas en nuevos fagos y comparación con el Método 1

Una vez validado, el modelo SVM fue aplicado a la predicción del tipo funcional de 14 endolisinas correspondientes a fagos activos contra *S. aureus*. Todas las secuencias fueron asignadas de manera inequívoca a uno de los cinco grupos definidos.

Los resultados obtenidos coincidieron plenamente con la clasificación manual previamente reportada mediante el Método 1 (Tabla 14), lo cual era esperable considerando el alto desempeño del modelo. Esta concordancia valida tanto la aproximación manual inicial como la capacidad del clasificador para generalizar correctamente sobre nuevas secuencias.

En todos los fagos analizados, los dominios catalíticos correspondieron a un dominio CHAP (cisteína-histidina-dependiente amidohidrolasa/peptidasa) y un dominio amidasa (AMI) de tipo 2 o 3, con la única excepción del fago vB_SauS_308, cuya endolisina presenta exclusivamente un dominio CHAP. Este fago fue clasificado de manera consistente como portador de una endolisina Tipo I.

Integración funcional holina–endolisina

La integración de las predicciones de endolisinas con la información de holinas permitió caracterizar de manera completa el módulo de lisis de cada fago. Se observó que:

- Seis fagos presentaron endolisinas Tipo III asociadas a holinas de 438 nt.
- Cinco fagos mostraron endolisinas Tipo IV junto con holinas de 303 nt.
- Dos fagos poseían endolisinas Tipo V y holinas de 276 nt.
- Un único fago (vB_SauS_308) presentó una endolisina Tipo I asociada a una holina de 255 nt.

Esta correspondencia refuerza la existencia de combinaciones funcionales conservadas entre holinas y endolisinas, sugiriendo una co-evolución de ambos componentes del módulo de lisis.

Un caso particular lo constituye el fago vB_SauS_Mh15, cuya endolisina está codificada por dos genes adyacentes separados por un intrón de grupo I, una característica previamente reportada en fagos estafilocócicos como K, X2, G1 y 85. Determinar si ambos genes forman parte de una única endolisina funcional requerirá evidencia experimental adicional.

Tabla 14: Resultados del análisis del casete de lisis para los catorce fagos

Fago	Predicción Tipo	Clasificación manual	Holina (nt.)	Holina (Pfam)
vB_SauS_308	1	I	255	PF04531
vB_SauS_321c	3	III	438	PF04531
vB_SauS_760				
vB_SauS_I73				
vB_SauS_Mh1				
vB_SauS_Mh4				
vB_SauS_Mh15				
vB_SauS_277g	4	IV	303	PF04688
vB_SauS_320				
vB_SauS_690				
vB_SauS_775				
vB_SauS_832				
vB_SauS_287	5	V	276	PF04688
vB_SauS_713				

La concordancia entre la clasificación manual y la predicción automática refuerza la utilidad del enfoque supervisado no solo como herramienta de validación, sino como un método escalable para la anotación funcional de endolisinas en grandes conjuntos de fagos genómicamente caracterizados.

Análisis de endolisinas con ausencia de CBDs

Los dominios de unión a la pared celular (CBDs, *cell wall binding domains*) de las endolisinas son componentes clave para la interacción rápida y de alta afinidad con la bacteria hospedadora. Dependiendo del epítopo reconocido, estos dominios pueden modular tanto la especificidad del reconocimiento como la eficiencia de la actividad catalítica de la enzima lítica (125). En este contexto, resulta particularmente llamativa la ausencia de CBDs anotados en las endolisinas de tipo I y tipo V y, específicamente, en las endolisinas de vB_SauS_308 (Tipo I) y vB_SauS_287 y vB_SauS_713 (Tipo V).

Con el objetivo de evaluar si esta ausencia correspondía a una verdadera falta de dominios de unión o, alternativamente, a la presencia de dominios estructuralmente conservados, pero no detectables mediante métodos basados en similitud de secuencia, se realizó un análisis estructural comparativo mediante predicción de estructura tridimensional.

Las estructuras de las endolisinas fueron predichas utilizando AlphaFold. En todos los casos, los modelos obtenidos presentaron valores elevados de confianza, con puntuaciones pLDDT superiores a 70 en la mayor parte de la estructura, y valores menores en las regiones de enlace (linkers) entre dominios, lo cual es consistente con la mayor flexibilidad esperada en dichas regiones. Estos resultados indican una alta confiabilidad global de las estructuras predichas.

Las estructuras modeladas fueron posteriormente alineadas y superpuestas utilizando ChimeraX. La comparación estructural entre las endolisinas de vB_SauS_308 y vB_SauS_277g, así como entre vB_SauS_713 y vB_SauS_277g, mostró un buen alineamiento global, con diferencias locales principalmente en la disposición de láminas β y hélices α (Figura 28).

A pesar de estas variaciones estructurales, el análisis reveló la presencia de un plegamiento compatible con dominios funcionales del tipo SH3, aunque no idénticos a los CBDs clásicos previamente caracterizados. Este resultado sugiere que, aun en ausencia de dominios anotados por herramientas de detección basadas en secuencia, estas endolisinas podrían conservar regiones estructurales con potencial capacidad de interacción con la pared celular bacteriana.

Desde un punto de vista aplicado, estos hallazgos resultan relevantes para el diseño racional de endolisinas sintéticas y quiméricas, ya que evidencian la posibilidad de identificar y reutilizar dominios

estructurales funcionales no evidentes a nivel de secuencia, ampliando el repertorio de módulos disponibles para la construcción de enzimas líticas con propiedades optimizadas.

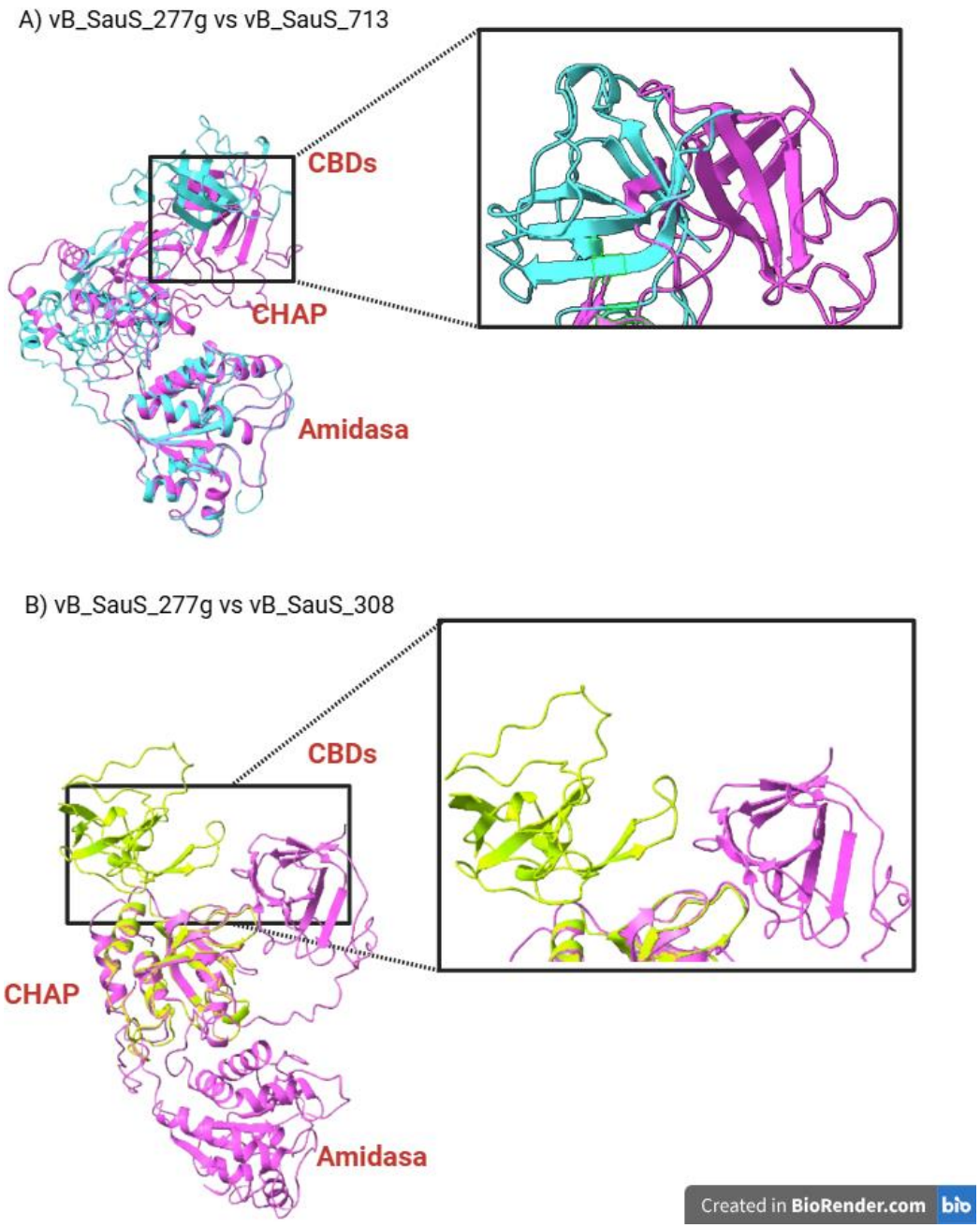

Figura 28: Alineamiento estructural de las endolisinas modeladas mediante AlphaFold. A) vB_SauS_277g (grupo IV, en magenta) y vB_SauS_713 (grupo V, en celeste). B) vB_SauS_277g (magenta) y vB_SauS_308 (grupo I, en verde).

En conjunto, el análisis estructural sugiere que la ausencia de CBDs detectables a nivel de secuencia en determinadas endolisinas no implica necesariamente la falta de regiones funcionales implicadas en la interacción con la pared celular bacteriana. La conservación de plegamientos compatibles con

dominios tipo SH3 indica que estos módulos podrían mantenerse a nivel estructural pese a una baja similitud de secuencia, lo que limita su detección mediante enfoques bioinformáticos clásicos. Este hallazgo refuerza la importancia de complementar los análisis basados en dominios con herramientas de predicción estructural, particularmente en sistemas altamente mosaicos como los genomas fágicos. Asimismo, pone de manifiesto la plasticidad evolutiva de los módulos de lisis y abre la posibilidad de que endolisinas sin CBDs anotados conserven mecanismos alternativos de reconocimiento del hospedador, con implicancias tanto evolutivas como biotecnológicas.

CONCLUSIONES

En el presente trabajo se anotaron y publicaron 14 nuevos genomas de bacteriófagos activos contra *Staphylococcus aureus*, contribuyendo al repositorio público de secuencias y posicionando a la producción local dentro de los diez países con mayor número de genomas de fagos depositados. Paralelamente, se construyó una base de datos local que integra 524 genomas de fagos estafilocócicos y un total de 62.440 proteínas, constituyendo un recurso robusto para estudios comparativos y análisis a gran escala. Este conjunto de datos puso en evidencia, además, el aumento reciente en la secuenciación de fagos jumbo, los cuales representan un reservorio particularmente relevante de diversidad genética para futuros estudios funcionales y evolutivos.

Se desarrolló y validó un flujo de trabajo bioinformático orientado a la semi-automatización de la anotación genómica y el análisis modular de fagos. Este enfoque permitió reducir drásticamente los tiempos de análisis —de meses a minutos— facilitando la integración de estudios in silico con experimentos de wet-lab, y reservando la curación manual para etapas críticas del proceso. La modularidad y reproducibilidad del pipeline lo convierten en una herramienta escalable y adaptable a nuevos conjuntos de datos.

El análisis funcional y estructural de los módulos genómicos permitió profundizar en los determinantes moleculares asociados a la biología de los fagos estafilocócicos. En particular, se evidenció la localización y posible organización estructural de dominios de unión a la pared celular (CBDs) de endolisinas previamente no caracterizados. De manera independiente y concomitante a este trabajo, uno de estos dominios fue incorporado a la base de datos InterPro en junio de 2024 (126), lo que refuerza la relevancia y validez del análisis realizado.

Asimismo, en comparación con estudios previos, se identificó en el fago vB_SauS_308 la presencia de un factor de virulencia (PVL) que no había sido reconocido con anterioridad, subrayando la importancia de revisitar genomas previamente anotados mediante enfoques integradores y actualizados.

En relación con el módulo de lisogenia, se propuso una nueva clasificación de integrasas basada en similitud global de secuencia, independiente de reconstrucciones filogenéticas tradicionales y del punto de corte seleccionado por el operador. Este enfoque permitió definir grupos evolutivamente coherentes de manera objetiva y reproducible, resolviendo limitaciones asociadas a la fragmentación en singletons y proporcionando un marco sólido para la clasificación automática de nuevas integrasas.

Desde una perspectiva metodológica, el pipeline desarrollado para integrasas separa explícitamente la definición de los grupos evolutivos (enfoque no supervisado) de la asignación automática de nuevas

secuencias (enfoque supervisado), alineándose con buenas prácticas bioinformáticas y criterios biológicos robustos. Esta estrategia no solo reproduce clasificaciones previamente reportadas, sino que mejora la generalización y estabilidad del análisis frente a conjuntos de datos heterogéneos o incompletos.

De manera complementaria, el clasificador supervisado basado en Support Vector Machines (SVM) desarrollado para la clasificación funcional de endolisinas demostró un desempeño elevado y consistente. La alta concordancia entre las predicciones automáticas y la clasificación manual, junto con la coherencia observada entre holinas y endolisinas dentro del módulo de lisis, respalda la aplicabilidad del modelo como una herramienta confiable para el análisis sistemático de fagos estafilocócicos. Este enfoque permite automatizar procesos tradicionalmente manuales y constituye una base sólida para futuros estudios orientados al diseño racional de aplicaciones terapéuticas basadas en fagos.

Finalmente, el conocimiento generado sobre los determinantes genómicos y proteicos involucrados en la interacción fago–hospedador sienta las bases para el diseño racional de cócteles de fagoterapia dirigidos a maximizar el rango de hospedadores de *S. aureus*. En este contexto, los resultados obtenidos contribuyen no solo al desarrollo de herramientas bioinformáticas avanzadas, sino también a la comprensión de procesos biológicos fundamentales, como la co-evolución entre fagos y bacterias y el impacto de los fagos en la evolución de *S. aureus*.

ANEXO

Consideraciones especiales

Todas las bases de datos utilizadas en este trabajo fueron actualizadas en febrero de 2025. Específicamente para las secuencias de bacteriófagos, no se esperaría a posterior un gran número de secuencias nuevas en base de datos del NCBI y, en consecuencia, que represente un cambio en la tendencia de los análisis.

Este trabajo fue realizando utilizando R Software (versión 4.4.2) en la aplicación R Studio (versión 2025.09.2) en Windows 11Pro (versión 24H2). Recursos claves como librerías de R o softwares de uso externo se resumen en la tabla 18.

Tablas

Tabla 15: Secuencias de bacteriófagos eliminadas de la base de datos

Nombre Bacteriófago	Gi
phiSa2wa-st5.3	2309819071
phiSa2wa-st5.4	2309819147
phiSa2wa-st5.8	2320102815
phiSa2wa-st5.9	2320102891
vB_SauR_MO29M	2753834889
SP6	402761649
SA537ruMSSAST97	1432226297
phiSa2wa-st80.2	2416400196
phiSa2wa-st80.4	2416400265
phiSa2wa_st93mssa	1321071470
Sa2wa_st8	1700519185
Sb1M_9832	1817916759
SAOMS1	1971775925
VB_SavM_JYL01	1471157087
LSA2308	1955069323
PVL-Sa2GN1	1013822632
v_SauR_MO29B	2753834909
SpT252	1103735309
S-CoN_Ph22	2582090220
J2	2587925052
BESEP8	1867215600

Tabla 16: Importancia de cada variable en el modelo RandomForest de clasificación de fagos

Variable	MeanDecreaseAccuracy	MeanDecreaseGini
Endolysin_3	21,46	5,88
S..aureus	30,44	5,53
Endolysin_6	9,75	4,73
PF04531	21,53	4,69
PF13102_PF00589	3,14	4,64
otro	15,44	4,50
PF02899_PF00589	15,13	4,37
PF04630_like	28,58	4,28
PF20155_PF01551	0,00	4,25
pol_III	14,16	4,22
PF00476	21,10	4,17
S..epidermidis	28,57	4,1
TIGR01603	29,20	3,90
PF04688	23,55	3,86
PF05521	30,13	3,83
PF10145_PF01551_like	24,77	3,77
PF05105	21,93	3,50
PF14659_PF00589	16,69	3,42
PF05135	27,08	3,27
PF05065	28,41	2,96
PF03837	13,99	2,88
PF05272	15,16	2,84
PF05119	25,15	2,81
PF04586	20,61	2,80
S..pseudintermedius	25,05	2,78
PF14657_PF13102_PF00589	0,95	2,75
PF02452	7,23	2,73
PF16838	18,80	2,69
PF05352	18,73	2,61
PF22768	0,56	2,49
PF03354_PF20441_like	22,69	2,30
PF00239_PF07508_PF13408_like	14,25	2,30
PF00154	15,80	2,26
PF04860	23,81	2,24

PF07768	30,96	2,19
PF23980	16,10	2,13
PF05876	19,13	2,12
PF10145_PF18013_PF01551	0,00	1,95
PF06199	28,06	1,91
PF00082	12,37	1,85
PF13495_PF00589	0,00	1,79
PF00665_PF13333	0,00	1,75
Endolysin_1	25,82	1,75
PF01612_like	19,75	1,58
PF13523_PF01636	0,00	1,55
PF03237	21,18	1,47
PF05133	23,64	1,40
PF14657_PF14659_PF00589	19,72	1,36
S..sciuri	13,29	1,36
PF04466_PF17288_like	21,85	1,36
SSF48371	21,27	1,30
PF03592	17,81	1,10
Endolysin_4	16,84	0,85
PF10145	1,99	0,75
G3DSA.3.10.350.10	20,33	0,70
S..capitis	0,00	0,60
S..haemolyticus	-1,48	0,58
PF09587	0,00	0,56
PF21311	33,31	0,52
S..warneri	0,00	0,37
PF01551_PF01464	-0,81	0,36
G3DSA.2.40.10.270	27,25	0,34
PF02821	12,61	0,32
Endolysin_2	16,18	0,20
PF00589	13,79	0,17
PF13022	2,66	0,14
PF07968	6,47	0,10
TIGR02126	8,70	0,05
PF05017	16,17	0,05
G3DSA.3.40.470.10_G3DSA.3.30.420.10	15,42	0,04
SSF56672	12,93	0,02

PF01123_PF02876_like	6,92	0,02
G3DSA.1.10.246.150	10,94	0,02
SSF54328_SSF54328	0,00	0,00
G3DSA.1.10.150.130	0,00	0,00

Tabla 17: Identidad porcentual par-a-par (PID) del alineamiento de RBPs candidatas vs. referencia y distancia a TMPs en pb

Id_candidato	Id_referencia	PID	Fago	Distancia a TMP (nt.)
vB_SauS_277g_59	NP_803298.1	97,63	vB_SauS_277g	2885
vB_SauS_277g_59	YP_001285375.1	97,64	vB_SauS_277g	2885
vB_SauS_287_65	NP_803298.1	87,02	vB_SauS_287	2866
vB_SauS_287_65	YP_001285375.1	87,16	vB_SauS_287	2866
vB_SauS_308_71	NP_803298.1	20,00	vB_SauS_308	6500
vB_SauS_320_62	NP_803298.1	40,94	vB_SauS_320	2722
vB_SauS_320_62	YP_001285375.1	40,55	vB_SauS_320	2722
vB_SauS_321c_60	NP_803298.1	94,95	vB_SauS_321c	2885
vB_SauS_321c_60	YP_001285375.1	95,12	vB_SauS_321c	2885
vB_SauS_690_63	NP_803298.1	40,94	vB_SauS_690	2722
vB_SauS_690_63	YP_001285375.1	40,55	vB_SauS_690	2722
vB_SauS_713_57	NP_803298.1	94,48	vB_SauS_713	2885
vB_SauS_713_57	YP_001285375.1	95,590	vB_SauS_713	2885
vB_SauS_760_65	NP_803298.1	88,13	vB_SauS_760	2866
vB_SauS_760_65	YP_001285375.1	88,41	vB_SauS_760	2866
vB_SauS_775_62	NP_803298.1	97,16	vB_SauS_775	2885
vB_SauS_775_62	YP_001285375.1	98,27	vB_SauS_775	2885
vB_SauS_832_67	NP_803298.1	97,48	vB_SauS_832	2885
vB_SauS_832_67	YP_001285375.1	97,63	vB_SauS_832	2885
vB_SauS_I73_63	NP_803298.1	88,09	vB_SauS_I73	2866
vB_SauS_I73_63	YP_001285375.1	88,43	vB_SauS_I73	2866
vB_SauS_Mh15_70	NP_803298.1	90,51	vB_SauS_Mh15	2866
vB_SauS_Mh15_70	YP_001285375.1	88,25	vB_SauS_Mh15	2866
vB_SauS_Mh1_67	NP_803298.1	87,46	vB_SauS_Mh1	2866
vB_SauS_Mh1_67	YP_001285375.1	88,45	vB_SauS_Mh1	2866
vB_SauS_Mh4_65	NP_803298.1	97,64	vB_SauS_Mh4	2885
vB_SauS_Mh4_65	YP_001285375.1	97,63	vB_SauS_Mh4	2885

Tabla 18: Recursos claves

Softwares y Servicios	Fuente	Identificador
R	(127)	https://www.R-project.org/
R Studio	(128)	
Google Colab	(129)	https://colab.google/
AlphaFold 2	(123)	https://alphafoldserver.com/
InterproScan	(73,130)	https://www.ebi.ac.uk/interpro/
NCBI	(131)	https://www.ncbi.nlm.nih.gov/
Paquetes R		
AnnotationBustR	(132)	
BiocManager	(133)	
BiocParallel	(134)	
Biostrings	(135)	
dplyr	(136)	
GenomeInfoDb	(137)	
GenomicRanges	(138)	
ggplot2	(139)	
Gviz	(140)	
Kebabs	(141)	
kernlab	(142)	
msa	(143)	
read.gb	(144)	
rentrez	(145)	
Rsamtools	(146)	
ShortRead	(147)	

Material Suplementario

El siguiente código QR dirige al repositorio de GitHub que contiene el material suplementario asociado a la presente tesis. Dicho material no fue incorporado al manuscrito principal debido a limitaciones de extensión, pero forma parte integral del trabajo desarrollado.

BIBLIOGRAFÍA

Referencias

1. WHO Bacterial Priority Pathogens List 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance. 1st ed. Geneva: World Health Organization; 2024. 1 p.

2. Mancuso G, Midiri A, Gerace E, Biondo C. Bacterial Antibiotic Resistance: The Most Critical Pathogens. Pathogens. 12 de octubre de 2021;10(10):1310. doi:10.3390/pathogens10101310

3. Nagel T, Musila L, Muthoni M, Nikolich M, Nakavuma JL, Clokie MR. Phage banks as potential tools to rapidly and cost-effectively manage antimicrobial resistance in the developing world. Curr Opin Virol. abril de 2022;53:101208. doi:10.1016/j.coviro.2022.101208

4. Wan F, Torres MDT, Peng J, De La Fuente-Nunez C. Deep-learning-enabled antibiotic discovery through molecular de-extinction. Nat Biomed Eng. 11 de junio de 2024;8(7):854-71. doi:10.1038/s41551-024-01201-x

5. Pipitò L, Rubino R, D'Agati G, Bono E, Mazzola CV, Urso S, et al. Antimicrobial Resistance in ESKAPE Pathogens: A Retrospective Epidemiological Study at the University Hospital of Palermo, Italy. Antibiotics. 12 de febrero de 2025;14(2):186. doi:10.3390/antibiotics14020186

6. Piewngam P, Otto M. Staphylococcus aureus colonisation and strategies for decolonisation. Lancet Microbe. junio de 2024;5(6):e606-18. doi:10.1016/S2666-5247(24)00040-5

7. Cervantes-García E, García-González R, Salazar-Schettino PM. Características generales del Staphylococcus aureus.

8. Chambers HF, DeLeo FR. Waves of resistance: Staphylococcus aureus in the antibiotic era. Nat Rev Microbiol. septiembre de 2009;7(9):629-41. doi:10.1038/nrmicro2200

9. Chiani P, Michelacci V, Delibato E, Ventola E, Mannan S, Marra M, et al. Survival and genomic stability of Yersinia enterocolitica in environmental Acanthamoeba spp. Front Cell Infect Microbiol. 30 de octubre de 2025;15:1642352. doi:10.3389/fcimb.2025.1642352

10. Pires DP, Oliveira H, Melo LDR, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. Appl Microbiol Biotechnol. marzo de 2016;100(5):2141-51. doi:10.1007/s00253-015-7247-0

11. Deghorain M, Van Melderen L. The Staphylococci Phages Family: An Overview. Viruses. 23 de noviembre de 2012;4(12):3316-35. doi:10.3390/v4123316

12. Gummalla VS, Zhang Y, Liao YT, Wu VCH. The Role of Temperate Phages in Bacterial Pathogenicity. Microorganisms. 21 de febrero de 2023;11(3):541. doi:10.3390/microorganisms11030541

13. Palma M, Qi B. Advancing Phage Therapy: A Comprehensive Review of the Safety, Efficacy, and Future Prospects for the Targeted Treatment of Bacterial Infections. Infect Dis Rep. 28 de noviembre de 2024;16(6):1127-81. doi:10.3390/idr16060092

14. Kapoor A, Mudaliar SB, Bhat VG, Chakraborty I, Prasad ASB, Mazumder N. Phage therapy: A novel approach against multidrug-resistant pathogens. 3 Biotech. octubre de 2024;14(10):256. doi:10.1007/s13205-024-04101-8

15. Twort FW, Lond LRCP, M.R.C.S. AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES. Vol. 186. 1915;186(4814):1241-3.

16. Publications Service. On an invisible microbe antagonistic toward dysenteric bacilli: brief note by Mr. F. D'Herelle, presented by Mr. Roux. Res Microbiol. septiembre de 2007;158(7):553-4. doi:10.1016/j.resmic.2007.07.005

17. Subramanian A. Emerging roles of bacteriophage-based therapeutics in combating antibiotic resistance. Front Microbiol. 5 de julio de 2024;15:1384164. doi:10.3389/fmicb.2024.1384164

18. Sukumaran AT, Nannapaneni R, Kiess A, Sharma CS. Reduction of Salmonella on chicken breast fillets stored under aerobic or modified atmosphere packaging by the application of lytic bacteriophage preparation SalmoFreshTM,. Poult Sci. marzo de 2016;95(3):668-75. doi:10.3382/ps/pev332

19. Gutiérrez D, Rodríguez-Rubio L, Fernández L, Martinez B, Rodriguez A, García P. Applicability of commercial phage-based products against Listeria monocytogenes for improvement of food safety in Spanish dry-cured ham and food contact surfaces. Food Control. 1 de noviembre de 2016;73. doi:10.1016/j.foodcont.2016.11.007

20. Soffer N, Woolston J, Li M, Das C, Sulakvelidze A. Bacteriophage preparation lytic for Shigella significantly reduces Shigella sonnei contamination in various foods. Skurnik M, editor. PLOS ONE. 31 de marzo de 2017;12(3):e0175256. doi:10.1371/journal.pone.0175256

21. Kazi M, Annapure US. Bacteriophage biocontrol of foodborne pathogens. J Food Sci Technol. marzo de 2016;53(3):1355-62. doi:10.1007/s13197-015-1996-8

22. Dhulipalla H, Basavegowda N, Haldar D, Syed I, Ghosh P, Rana SS, et al. Integrating phage biocontrol in food production: industrial implications and regulatory overview. Discov Appl Sci. 7 de abril de 2025;7(4):314. doi:10.1007/s42452-025-06754-3

23. Ranveer SA, Dasriya V, Ahmad MF, Dhillon HS, Samtiya M, Shama E, et al. Positive and negative aspects of bacteriophages and their immense role in the food chain. Npj Sci Food. 3 de enero de 2024;8(1):1. doi:10.1038/s41538-023-00245-8

24. Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D. Exploring the prokaryotic virosphere. Res Microbiol. junio de 2008;159(5):306-13. doi:10.1016/j.resmic.2008.05.001

25. Ely B, Lenski J, Mohammadi T. Structural and Genomic Diversity of Bacteriophages. En: Tumban E, editor. Bacteriophages: Methods and Protocols [Internet]. New York, NY: Springer US; 2024. p. 3-16. Disponible en: https://doi.org/10.1007/978-1-0716-3549-0_1 doi:10.1007/978-1-0716-3549-0_1

26. Valencia-Toxqui G, Ramsey J. How to introduce a new bacteriophage on the block: a short guide to phage classification. Mukhopadhyay S, editor. J Virol. 22 de octubre de 2024;98(10):e01821-23. doi:10.1128/jvi.01821-23

27. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat Rev Microbiol. octubre de 2015;13(10):641-50. doi:10.1038/nrmicro3527

28. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. ISME J. 1 de julio de 2017;11(7):1511-20. doi:10.1038/ismej.2017.16

29. Ofir G, Sorek R. Contemporary Phage Biology: From Classic Models to New Insights. Cell. marzo de 2018;172(6):1260-70. doi:10.1016/j.cell.2017.10.045

30. Naureen Z, Dautaj A, Anpilogov K, Camilleri G, Dhuli K, Tanzi B, et al. Bacteriophages presence in nature and their role in the natural selection of bacterial populations. Acta Bio Medica Atenei Parm. 9 de noviembre de 2020;91(13-S):e2020024. doi:10.23750/abm.v91i13-S.10819

31. Chung KM, Liau XL, Tang SS. Bacteriophages and Their Host Range in Multidrug-Resistant Bacterial Disease Treatment. Pharmaceuticals. 16 de octubre de 2023;16(10):1467. doi:10.3390/ph16101467

32. Tormo MÁ, Ferrer MD, Maiques E, Úbeda C, Selva L, Lasa Í, et al. *Staphylococcus aureus* Pathogenicity Island DNA Is Packaged in Particles Composed of Phage Proteins. J Bacteriol. abril de 2008;190(7):2434-40. doi:10.1128/JB.01349-07

33. Ubeda C, Olivarez NP, Barry P, Wang H, Kong X, Matthews A, et al. Specificity of staphylococcal phage and SaPI DNA packaging as revealed by integrase and terminase mutations. Mol Microbiol. abril de 2009;72(1):98-108. doi:10.1111/j.1365-2958.2009.06634.x

34. Sobral R, Tomasz A. The Staphylococcal Cell Wall. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, editores. Microbiol Spectr. 19 de julio de 2019;7(4):7.4.12. doi:10.1128/microbiolspec.GPP3-0068-2019

35. Wang M, Buist G, van Dijl JM. *Staphylococcus aureus* cell wall maintenance – the multifaceted roles of peptidoglycan hydrolases in bacterial growth, fitness, and virulence. FEMS Microbiol Rev. 28 de octubre de 2022;46(5):fuac025. doi:10.1093/femsre/fuac025

36. Moller AG, Lindsay JA, Read TD. Determinants of Phage Host Range in *Staphylococcus* Species. Johnson KN, editor. Appl Environ Microbiol. junio de 2019;85(11):e00209-19. doi:10.1128/AEM.00209-19

37. Xia G, Wolz C. Phages of Staphylococcus aureus and their impact on host evolution. Infect Genet Evol. enero de 2014;21:593-601. doi:10.1016/j.meegid.2013.04.022

38. Thabet MA, Penadés JR, Haag AF. The ClpX protease is essential for inactivating the CI master repressor and completing prophage induction in Staphylococcus aureus. Nat Commun. 18 de octubre de 2023;14(1):6599. doi:10.1038/s41467-023-42413-0

39. Laanto E. Overcoming Bacteriophage Resistance in Phage Therapy. En: Tumban E, editor. Bacteriophages: Methods and Protocols [Internet]. New York, NY: Springer US; 2024. p. 401-10. Disponible en: https://doi.org/10.1007/978-1-0716-3549-0_23 doi:10.1007/978-1-0716-3549-0_23

40. Cervera-Alamar M, Guzmán-Markevitch K, Žiemytė M, Ortí L, Bernabé-Quispe P, Pineda-Lucena A, et al. Mobilisation Mechanism of Pathogenicity Islands by Endogenous Phages in Staphylococcus aureus clinical strains. Sci Rep. 13 de noviembre de 2018;8(1):16742. doi:10.1038/s41598-018-34918-2

41. Ackermann HW. Tailed Bacteriophages: The Order Caudovirales. Adv Virus Res. 1998;51:135-201. doi:10.1016/s0065-3527(08)60785-x

42. Kwan T, Liu J, DuBow M, Gros P, Pelletier J. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. Proc Natl Acad Sci. 5 de abril de 2005;102(14):5174-9. doi:10.1073/pnas.0501140102

43. Korn AM, Hillhouse AE, Sun L, Gill JJ. Comparative Genomics of Three Novel Jumbo Bacteriophages Infecting Staphylococcus aureus. Pfeiffer JK, editor. J Virol. 9 de septiembre de 2021;95(19):e02391-20. doi:10.1128/JVI.02391-20

44. Hatfull GF, Hendrix RW. Bacteriophages and their genomes. Curr Opin Virol. octubre de 2011;1(4):298-303. doi:10.1016/j.coviro.2011.06.009

45. Ingmer H, Gerlach D, Wolz C. Temperate Phages of *Staphylococcus aureus*. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, editores. Microbiol Spectr. 27 de septiembre de 2019;7(5):7.5.1. doi:10.1128/microbiolspec.GPP3-0058-2018

46. Fogg PCM, Colloms S, Rosser S, Stark M, Smith MCM. New Applications for Phage Integrases. J Mol Biol. julio de 2014;426(15):2703-16. doi:10.1016/j.jmb.2014.05.014

47. Goerke C, Pantucek R, Holtfreter S, Schulte B, Zink M, Grumann D, et al. Diversity of Prophages in Dominant *Staphylococcus aureus* Clonal Lineages. J Bacteriol. junio de 2009;191(11):3462-8. doi:10.1128/JB.01804-08

48. García P, Martínez B, Obeso JM, Lavigne R, Lurz R, Rodríguez A. Functional Genomic Analysis of Two *Staphylococcus aureus* Phages Isolated from the Dairy Environment. Appl Environ Microbiol. 15 de diciembre de 2009;75(24):7663-73. doi:10.1128/AEM.01864-09

49. Lo CY, Gao Y. DNA Helicase–Polymerase Coupling in Bacteriophage DNA Replication. Viruses. 31 de agosto de 2021;13(9):1739. doi:10.3390/v13091739

50. Pandey M, Syed S, Donmez I, Patel G, Ha T, Patel SS. Coordinating DNA replication by means of priming loop and differential synthesis rate. Nature. diciembre de 2009;462(7275):940-3. doi:10.1038/nature08611

51. Aksyuk AA, Rossmann MG. Bacteriophage Assembly. Viruses. 25 de febrero de 2011;3(3):172-203. doi:10.3390/v3030172

52. Fokine A, Rossmann MG. Molecular architecture of tailed double-stranded DNA phages. Bacteriophage. abril de 2014;4(2):e28281. doi:10.4161/bact.28281

53. Rao VB, Feiss M. The Bacteriophage DNA Packaging Motor. Annu Rev Genet. 1 de diciembre de 2008;42(1):647-81. doi:10.1146/annurev.genet.42.110807.091545

54. Casjens SR, Gilcrease EB. Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. En: Clokie MRJ, Kropinski AM, editores. Bacteriophages [Internet]. Totowa, NJ: Humana Press; 2009 [citado 21 de mayo de 2025]. p. 91-111. (Walker JM, ed. Methods in Molecular Biology). Disponible en: http://link.springer.com/10.1007/978-1-60327-565-1_7 doi:10.1007/978-1-60327-565-1_7

55. Esterman ES, Wolf YI, Kogay R, Koonin EV, Zhaxybayeva O. Evolution of DNA packaging in gene transfer agents. Virus Evol. 20 de enero de 2021;7(1):veab015. doi:10.1093/ve/veab015

56. Moller AG, Winston K, Ji S, Wang J, Davis MNH, Solís-Lemus CR, et al. Genes Influencing Phage Host Range in Staphylococcus aureus on a Species-Wide Scale. mSphere. enero de 2021;6:e01263-20. doi:10.1128/%20mSphere.01263-20

57. Wang IN, Smith DL, Young R. Holins: The Protein Clocks of Bacteriophage Infections. Annu Rev Microbiol. octubre de 2000;54(1):799-825. doi:10.1146/annurev.micro.54.1.799

58. Bałdysz S, Nawrot R, Barylski J. "Tear down that wall"—a critical evaluation of bioinformatic resources available for lysin researchers. Vives M, editor. Appl Environ Microbiol. 24 de julio de 2024;90(7):e02361-23. doi:10.1128/aem.02361-23

59. Fernandes S, Proença D, Cantante C, Silva FA, Leandro C, Lourenço S, et al. Novel Chimerical Endolysins with Broad Antimicrobial Activity Against Methicillin-Resistant *Staphylococcus aureus*. Microb Drug Resist. junio de 2012;18(3):333-43. doi:10.1089/mdr.2012.0025

60. Oliveira H, Melo LDR, Santos SB, Nóbrega FL, Ferreira EC, Cerca N, et al. Molecular Aspects and Comparative Genomics of Bacteriophage Endolysins. J Virol. 15 de abril de 2013;87(8):4558-70. doi:10.1128/JVI.03277-12

61. Schmelcher M, Donovan DM, Loessner MJ. Bacteriophage Endolysins as Novel Antimicrobials. Future Microbiol. octubre de 2012;7(10):1147-71. doi:10.2217/fmb.12.97

62. Nelson DC, Schmelcher M, Rodriguez-Rubio L, Klumpp J, Pritchard DG, Dong S, et al. Endolysins as Antimicrobials. En: Advances in Virus Research [Internet]. Elsevier; 2012 [citado 20 de noviembre de 2025]. p. 299-365. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780123944382000074 doi:10.1016/B978-0-12-394438-2.00007-4

63. Gerstmans H, Rodríguez-Rubio L, Lavigne R, Briers Y. From endolysins to Artilysin®s: novel enzyme-based approaches to kill drug-resistant bacteria. Biochem Soc Trans. 15 de febrero de 2016;44(1):123-8. doi:10.1042/BST20150192

64. Vázquez R, Gutiérrez D, Criel B, Dezutter Z, Briers Y. Diversity, structure-function relationships and evolution of cell wall-binding domains of staphylococcal phage endolysins [Internet]. Molecular Biology; 2025 [citado 19 de febrero de 2025]. Disponible en: http://biorxiv.org/lookup/doi/10.1101/2025.02.06.636829 doi:10.1101/2025.02.06.636829

65. Chang Y, Ryu S. Characterization of a novel cell wall binding domain-containing Staphylococcus aureus endolysin LysSA97. Appl Microbiol Biotechnol. enero de 2017;101(1):147-58. doi:10.1007/s00253-016-7747-6

66. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 26 de noviembre de 2012;41(D1):D36-42. doi:10.1093/nar/gks1195

67. Dedrick RM, Guerrero-Bustamante CA, Garlena RA, Russell DA, Ford K, Harris K, et al. Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. Nat Med. mayo de 2019;25(5):730-3. doi:10.1038/s41591-019-0437-z

68. Kilcher S, Loessner MJ. Engineering Bacteriophages as Versatile Biologics. Trends Microbiol. abril de 2019;27(4):355-67. doi:10.1016/j.tim.2018.09.006

69. Tang SS, Biswas SK, Tan WS, Saha AK, Leo BF. Efficacy and potential of phage therapy against multidrug resistant *Shigella* spp. PeerJ. 5 de abril de 2019;7:e6225. doi:10.7717/peerj.6225

70. Turner D, Adriaenssens EM, Tolstoy I, Kropinski AM. Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation. PHAGE. 1 de diciembre de 2021;2(4):170-82. doi:10.1089/phage.2021.0013

71. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 1 de octubre de 2000;16(10):944-5. doi:10.1093/bioinformatics/16.10.944

72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. octubre de 1990;215(3):403-10. doi:10.1016/S0022-2836(05)80360-2

73. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. InterPro: the protein sequence classification resource in 2025. Nucleic Acids Res. 6 de enero de 2025;53(D1):D444-56. doi:10.1093/nar/gkae1082

74. Besemer J. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 15 de junio de 2001;29(12):2607-18. doi:10.1093/nar/29.12.2607

75. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. julio de 2018;28(7):1079-89. doi:10.1101/gr.230615.117

76. Delcher A. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1 de diciembre de 1999;27(23):4636-41. doi:10.1093/nar/27.23.4636

77. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. diciembre de 2010;11(1):119. doi:10.1186/1471-2105-11-119

78. Pope WH, Jacobs-Sera D. Annotation of Bacteriophage Genome Sequences Using DNA Master: An Overview. En: Clokie MRJ, Kropinski AM, Lavigne R, editores. Bacteriophages [Internet]. New York, NY: Springer New York; 2018 [citado 20 de noviembre de 2025]. p. 217-29. (Methods in Molecular Biology). Disponible en: http://link.springer.com/10.1007/978-1-4939-7343-9_16 doi:10.1007/978-1-4939-7343-9_16

79. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: A Fast Phage Search Tool. Nucleic Acids Res. 1 de julio de 2011;39(suppl):W347-52. doi:10.1093/nar/gkr485

80. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 15 de julio de 2014;30(14):2068-9. doi:10.1093/bioinformatics/btu153

81. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. PHANOTATE: a novel approach to gene identification in phage genomes. Hancock J, editor. Bioinformatics. 1 de noviembre de 2019;35(22):4537-42. doi:10.1093/bioinformatics/btz265

82. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. eLife. 22 de julio de 2015;4:e08490. doi:10.7554/eLife.08490

83. Beggs GA, Bassler BL. Phage small proteins play large roles in phage–bacterial interactions. Curr Opin Microbiol. agosto de 2024;80:102519. doi:10.1016/j.mib.2024.102519

84. Jordan B, Weidenbach K, Schmitz RA. The power of the small: the underestimated role of small proteins in bacterial and archaeal physiology. Curr Opin Microbiol. diciembre de 2023;76:102384. doi:10.1016/j.mib.2023.102384

85. Fremin BJ, Bhatt AS, Kyrpides NC, Sengupta A, Sczyrba A, Maria Da Silva A, et al. Thousands of small, novel genes predicted in global phage genomes. Cell Rep. junio de 2022;39(12):110984. doi:10.1016/j.celrep.2022.110984

86. Guan J, Ji Y, Peng C, Zou W, Tang X, Shang J, et al. PhaGO: Protein function annotation for bacteriophages by integrating the genomic context [Internet]. arXiv; 2024 [citado 28 de enero de 2026]. Disponible en: http://arxiv.org/abs/2408.06402 doi:10.48550/arXiv.2408.06402

87. Kumar S, Stecher G, Suleski M, Sanderford M, Sharma S, Tamura K. MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. Battistuzzi FU, editor. Mol Biol Evol. 6 de diciembre de 2024;41(12):msae263. doi:10.1093/molbev/msae263

88. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 1 de mayo de 2014;30(9):1312-3. doi:10.1093/bioinformatics/btu033

89. Kahánková J, Pantůček R, Goerke C, Růžičková V, Holochová P, Doškař J. Multilocus PCR typing strategy for differentiation of *Staphylococcus aureus* siphoviruses reflecting their modular genome structure. Environ Microbiol. septiembre de 2010;12(9):2527-38. doi:10.1111/j.1462-2920.2010.02226.x

90. Abedon ST, Kuhl SJ, Blasdel BG, Kutter EM. Phage treatment of human infections. Bacteriophage. marzo de 2011;1(2):66-85. doi:10.4161/bact.1.2.15845

91. Pirnay JP, De Vos D, Verbeken G, Merabishvili M, Chanishvili N, Vaneechoutte M, et al. The Phage Therapy Paradigm: Prêt-à-Porter or Sur-mesure? Pharm Res. abril de 2011;28(4):934-7. doi:10.1007/s11095-010-0313-5

92. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res. 6 de enero de 2025;53(D1):D20-9. doi:10.1093/nar/gkae979

93. Rodriguez-Tome P. The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res. 1 de enero de 1996;24(1):6-12. doi:10.1093/nar/24.1.6

94. Cochrane G, Karsch-Mizrachi I, Takagi T, Sequence Database Collaboration IN. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 4 de enero de 2016;44(D1):D48-50. doi:10.1093/nar/gkv1323

95. Berman HM. The Protein Data Bank. Nucleic Acids Res. 1 de enero de 2000;28(1):235-42. doi:10.1093/nar/28.1.235

96. Burley SK, Bhatt R, Bhikadiya C, Bi C, Biester A, Biswas P, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. Nucleic Acids Res. 6 de enero de 2025;53(D1):D564-74. doi:10.1093/nar/gkae1091

97. Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Brief Bioinform. noviembre de 2016;17(6):1009-23. doi:10.1093/bib/bbv099

98. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. marzo de 1970;48(3):443-53. doi:10.1016/0022-2836(70)90057-4

99. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. marzo de 1981;147(1):195-7. doi:10.1016/0022-2836(81)90087-5

100. Shi H, Zhang X. Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms. Front Genet. 27 de febrero de 2020;11:105. doi:10.3389/fgene.2020.00105

101. Sievers F, Higgins DG. The Clustal Omega Multiple Alignment Package. En: Katoh K, editor. Multiple Sequence Alignment [Internet]. New York, NY: Springer US; 2021 [citado 25 de noviembre de 2025]. p. 3-16. (Methods in Molecular Biology). Disponible en: http://link.springer.com/10.1007/978-1-0716-1036-7_1 doi:10.1007/978-1-0716-1036-7_1

102. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 19 de agosto de 2004;5(1):113. doi:10.1186/1471-2105-5-113

103. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 6 de enero de 2025;53(D1):D523-34. doi:10.1093/nar/gkae997

104. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 20 de septiembre de 2021;49(16):9077-96. doi:10.1093/nar/gkab688

105. Suárez CA, Carrasco ST, Brandolisio FNA, Abatangelo V, Boncompain CA, Peresutti-Bacci N, et al. Bioinformatic Analysis of a Set of 14 Temperate Bacteriophages Isolated from Staphylococcus aureus Strains Highlights Their Massive Genetic Diversity. Tyne DV, editor. Microbiol Spectr. 31 de agosto de 2022;10(4):e00334-22. doi:10.1128/spectrum.00334-22

106. Tritt A, Eisen JA, Facciotti MT, Darling AE. An Integrated Pipeline for de Novo Assembly of Microbial Genomes. Zhu D, editor. PLoS ONE. 13 de septiembre de 2012;7(9):e42304. doi:10.1371/journal.pone.0042304

107. Kropinski AM, Prangishvili D, Lavigne R. Position paper: The creation of a rational scheme for the nomenclature of viruses of *Bacteria* and *Archaea*. Environ Microbiol. noviembre de 2009;11(11):2775-7. doi:10.1111/j.1462-2920.2009.01970.x

108. Goldfarb T, Kodali VK, Pujar S, Brover V, Robbertse B, Farrell CM, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. 6 de enero de 2025;53(D1):D243-57. doi:10.1093/nar/gkae1038

109. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 8 de enero de 2021;49(D1):D121-4. doi:10.1093/nar/gkaa967

110. Microsoft Corporation. Microsfot Power BI Desktop [Internet]. 2025. Disponible en: https://powerbi.microsoft.com

111. Waman VP, Bordin N, Alcraft R, Vickerstaff R, Rauer C, Chan Q, et al. CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. J Mol Biol. septiembre de 2024;436(17):168551. doi:10.1016/j.jmb.2024.168551

112. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. enero de 2009;37(suppl_1):D380-6. doi:10.1093/nar/gkn762

113. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 1 de julio de 2005;33(Web Server):W244-8. doi:10.1093/nar/gki408

114. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 8 de enero de 2021;49(D1):D412-9. doi:10.1093/nar/gkaa913

115. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. 3 de julio de 2017;45(W1):W30-5. doi:10.1093/nar/gkx343

116. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 8 de enero de 2019;47(D1):D687-92. doi:10.1093/nar/gky1080

117. Oliveira H, Sampaio M, Melo LDR, Dias O, Pope WH, Hatfull GF, et al. Staphylococci phages display vast genomic diversity and evolutionary relationships. BMC Genomics. diciembre de 2019;20(1):357. doi:10.1186/s12864-019-5647-8

118. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 1 de abril de 2011;27(7):1009-10. doi:10.1093/bioinformatics/btr039

119. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10(1):210. doi:10.1186/1471-2148-10-210

120. Kizziah JL, Manning KA, Dearborn AD, Dokland T. Structure of the host cell recognition and penetration machinery of a Staphylococcus aureus bacteriophage. Peschel A, editor. PLOS Pathog. 18 de febrero de 2020;16(2):e1008314. doi:10.1371/journal.ppat.1008314

121. Takeuchi I, Osada K, Azam AH, Asakawa H, Miyanaga K, Tanji Y. The Presence of Two Receptor-Binding Proteins Contributes to the Wide Host Range of Staphylococcal Twort-Like Phages. Pettinari MJ, editor. Appl Environ Microbiol. octubre de 2016;82(19):5763-74. doi:10.1128/AEM.01385-16

122. Botka T, Pantůček R, Mašlaňová I, Benešík M, Petráš P, Růžičková V, et al. Lytic and genomic properties of spontaneous host-range Kayvirus mutants prove their suitability for upgrading phage therapeutics against staphylococci. Sci Rep. 2 de abril de 2019;9(1):5475. doi:10.1038/s41598-019-41868-w

123. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 26 de agosto de 2021;596(7873):583-9. doi:10.1038/s41586-021-03819-2

124. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF ChimeraX : Tools for structure building and analysis. Protein Sci. noviembre de 2023;32(11):e4792. doi:10.1002/pro.4792

125. Schuch R, Cassino C, Vila-Farres X. Direct Lytic Agents: Novel, Rapidly Acting Potential Antimicrobial Treatment Modalities for Systemic Use in the Era of Rising Antibiotic Resistance. Front Microbiol. 3 de marzo de 2022;13:841905. doi:10.3389/fmicb.2022.841905

126. Vázquez R, Gutiérrez D, Grimon D, Fernández L, García P, Rodríguez A, et al. The New SH3b_T Domain Increases the Structural and Functional Variability Among SH3b-Like CBDs from Staphylococcal Phage Endolysins. Probiotics Antimicrob Proteins. 30 de julio de 2024. doi:10.1007/s12602-024-10309-0

127. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Disponible en: https://www.R-project.org/

128. Posit team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: Posit Software, PBC; 2024. Disponible en: http://www.posit.co/

129. Google Colaboratory [Internet]. Google; 2024. Disponible en: https://colab.research.google.com/

130. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 1 de mayo de 2014;30(9):1236-40. doi:10.1093/bioinformatics/btu031

131. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res. 6 de enero de 2025;53(D1):D20-9. doi:10.1093/nar/gkae979

132. Samuel Borstein and Brian O'Meara. AnnotationBustR: Extract Subsequences from GenBank Annotations [Internet]. Disponible en: https://github.com/sborstein/AnnotationBustR R package version 1.2.

133. Martin Morgan and Marcel Ramos. BiocManager: Access the Bioconductor Project Package Repository [Internet]. Disponible en: https://CRAN.R-project.org/package=BiocManager R package version 1.30.25.

134. Bioconductor Package Maintainer [Cre] MM [Aut]. BiocParallel [Internet]. Bioconductor; 2017 [citado 3 de febrero de 2026]. Disponible en: https://bioconductor.org/packages/BiocParallel doi:10.18129/B9.BIOC.BIOCPARALLEL

135. H. Pagès PA. Biostrings [Internet]. Bioconductor; 2017 [citado 3 de febrero de 2026]. Disponible en: https://bioconductor.org/packages/Biostrings doi:10.18129/B9.BIOC.BIOSTRINGS

136. Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan. dplyr: A Grammar of Data Manipulation [Internet]. 2023. Disponible en: https://CRAN.R-project.org/package=dplyr

137. Sonali Arora MM. GenomeInfoDb [Internet]. Bioconductor; 2017 [citado 3 de febrero de 2026]. Disponible en: https://bioconductor.org/packages/GenomeInfoDb doi:10.18129/B9.BIOC.GENOMEINFODB

138. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol. 8 de agosto de 2013;9(8):e1003118. doi:10.1371/journal.pcbi.1003118

139. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York. Disponible en: https://ggplot2.tidyverse.org doi:978-3-319-24277-4

140. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. En: Mathé E, Davis S, editores. Statistical Genomics [Internet]. New York, NY: Springer New York; 2016 [citado 3 de febrero de 2026]. p. 335-51. (Methods in Molecular Biology). Disponible en: http://link.springer.com/10.1007/978-1-4939-3578-9_16 doi:10.1007/978-1-4939-3578-9_16

141. Palme J, Hochreiter S, Bodenhofer U. KeBABS: an R package for kernel-based analysis of biological sequences. Bioinformatics. 1 de agosto de 2015;31(15):2574-6. doi:10.1093/bioinformatics/btv176

142. Karatzoglou A, Smola A, Hornik K, Zeileis A. **kernlab** - An *S4* Package for Kernel Methods in *R*. J Stat Softw. 2004;11(9). doi:10.18637/jss.v011.i09

143. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics. 15 de diciembre de 2015;31(24):3997-9. doi:10.1093/bioinformatics/btv494

144. Robin Mercier. read.gb: Open GenBank Files [Internet]. Disponible en: https://CRAN.R-project.org/package=read.gb

145. David J. Winter. rentrez: an R package for the NCBI eUtils API. Vol. 9. 2017;9(2):520-6.

146. Martin Morgan HP. Rsamtools [Internet]. Bioconductor; 2017 [citado 3 de febrero de 2026]. Disponible en: https://bioconductor.org/packages/Rsamtools doi:10.18129/B9.BIOC.RSAMTOOLS

147. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. 1 de octubre de 2009;25(19):2607-8. doi:10.1093/bioinformatics/btp450