TESIS DOCTORAL

Estudio metagenómico de la comunidad de virus y de su interacción con la microbiota en la cavidad bucal humana

Marcos Parras Moltó

Madrid, 2019

Estudio metagenómico de la comunidad de virus y de su interacción con la microbiota en la cavidad bucal humana

Memoria presentada por Marcos Parras Moltó para optar al título de Doctor por la Universidad Autónoma de Madrid

Esta Tesis se ha realizado en el Centro de Biología Molecular Severo Ochoa bajo la supervisión del Tutor y Director Alberto López Bueno, en el Programa de Doctorado en Biociencias Moleculares (RD 99/2011)

Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular
Centro de Biología Molecular Severo Ochoa (CBMSO)
Madrid, 2019

El Dr. Alberto López Bueno, Profesor Contratado Doctor en el Departamento de Biología Molecular de la Universidad Autónoma de Madrid (UAM) e investigador en el Centro de Biología Molecular Severo Ochoa (CBMSO):

CERTIFICA:

Haber dirigido y supervisado la Tesis Doctoral titulada "**Estudio metagenómico de la comunidad de virus y de su interacción con la microbiota en la cavidad bucal humana**" realizada por D. Marcos Parras Moltó, en el Programa de Doctorado en Biociencias Moleculares de la Universidad Autónoma de Madrid, por lo que autoriza la presentación de la misma.

Madrid, a 23 de Abril de 2019,

Alberto López Bueno

La presente tesis doctoral ha sido posible gracias a la concesión de una “Ayuda para Contratos Predoctorales para la Formación de Doctores” convocatoria de 2013 (BES-2013-064773) asociada al proyecto SAF2012-38421 del Ministerio de Economía y Competitividad.

Durante esta tesis se realizó una estancia de dos meses en el laboratorio del Catedrático Francisco Rodríguez Valera, director de grupo de investigación: *Evolutionary Genomics Group* de la Universidad Miguel Hernández de Elche (San Juan de Alicante), gracias a una “Ayuda a la Movilidad Predoctoral para la Realización de Estancias Breves en Centros de I+D” convocatoria de 2015 (EEBB-I-16-11876) concedida por el Ministerio de Economía y Competitividad.

AGRADECIMIENTOS

Quisiera agradecer en primer lugar a Alberto López Bueno el haberme dado la oportunidad de desarrollar esta tesis bajo su tutela y dirección. Recuerdo como si fuera ayer aquella entrevista por Skype a pocos días de finalizar el plazo de entrega de documentación para la FPI, como si fuera ayer el día en que eché a volar. Gracias por tu paciencia (esto quisiera resaltarlo especialmente), compromiso y entrega. Gracias por ser tan buen profesional tanto como docente como científico, por demostrarme que quien quiere puede, que con esfuerzo se consiguen las cosas y que los límites no provienen de uno mismo. Más que un jefe, un amigo y compañero. Aquel Viernes llamando por primera vez a Paco, rellenando la documentación necesaria para la estancia antes de las 15.00, hora en que se cumplía el deadline...grandes momentos que meteré por siempre en la nube de mi memoria. Mención especial a Patricia y Ana, por haber sido mi motor al principio, y es que este grupo no siempre fue cosa solo de dos.

Agradecer al 203 y al 224 su acogida. Gracias a Rocío por ser un par de manos para un manco de las pipetas, gracias a ti hoy presumo de "saber" hacer mini y maxipreps, poca broma para un bioinformático. A Mari Carmen por su alegría y su atención, a Carolina y Leyre por ayudarme siempre que han podido, a Bruno por aportar cada día ese punto friki que alegra las mañanas, a Dani por enseñarme tan bien lo que es ser un *machanguito* y Antonio por darme cabida bajo su techo y haberme permitido así conocer a tan maravillosa familia. El *killo* siempre estará para esas cañas de los Viernes a la una. A mis compañeros y amigos predocs así como al postdoc que siempre nos ha cuidado. A Rastrojo, por su sabiduría, paciencia, por haber sido mi segundo maestro en esta aventura. Una auténtica máquina. A Domingo, porque tu sistema de puntos me lo llevaré donde vaya. Murcia está orgullosa de haber sido cuna de semejante esperpento de la verborrea hispanohablante. Que nunca falte *paparajote a pajera abierta*. A Carlos, por ser mi parigual videojugabilistico. Y es que aunque haya sido una pena que llegases a mitad de la partida, aquí estamos ya, frente al jefe final. A Graciela, por haberme llevado a los rincones más maravillosos de la geografía española. Ya no me imagino la vida sin una visita anual a Asturias. Gracias por evitarme ese último culín aquel día, en aquella cena, en aquel lugar. Gracias a Pepa por su desparpajo y su alegría desde primera hora de la mañana. Vaya fiesta montabamos cada vez que traías comida para compartir. A cada uno de los predocs del 224: Nooshin, Carlos, Rebeca, Tania, y todas aquellas personas que eventualmente han residido bajo el techo de nuestro actual hogar. Gracias Pepe por prestarnos ese hueco en el que hemos podido desarrollar nuestras locuras científicas. Y no me olvido de la "locura" de cada tarde. Gracias Rosa por darnos una última chispa de energía cada día cuando ya no podíamos más, tu alegría es contagiosa. En resumen, gracias a todas las personas que he

tenido la suerte de conocer en el CBM, como Silvia, pero en especial a ese *vato* que me llenó de ilusión e historias desde el primer día. Gracias Callejas por apoyarme, ayudarme, enseñarme y preocuparte por mi en todo momento. Echaré de menos esas charlas durante la comida y las subsiguientes rondas de preguntas aleatorias. No, los españoles no solemos hacer armaduras en nuestros ratos libres. No es casualidad que lo tuyo sea la numismática, porque vales oro.

Gracias a Francisco Rodríguez Varela por abrirme las puertas de su laboratorio y permitirme aprender de su mano. A cada una de las personas que compartieron conmigo esos dos meses: Rafa, Ricardo, Rohit, José y especialmente a Mario. Gracias por tu dedicación y por tu alegría. No hubo mejor manera de conocer Alicante. Gracias a Alex Mira y Áurea, a Asier y José Manuel, por su colaboración en este proyecto. De la misma forma, agradecer al servicio de secuenciación del parque científico de Madrid y a los integrantes del grupo de bioinformática del CBM, especialmente a Alfonso, su labor estos años.

Gracias a Antoñito por ser mi familia en Madrid, un amigo para lo bueno y para lo malo. Que no nos quede un festival de electrónica por pisar. A Borja y Sandra, por ser ese trozo de Madrid que llevaré por siempre en la mochila allá donde vaya.

Agradecer a todos lo que han formado parte de mi mundo durante esos años de viajes a *La Terreta*. He tenido la suerte de compartir tiempo y vivencias con personas maravillosas que han crecido bajo el techo del IATA, demostrando que la mejor ciencia es la que se hace desde un grupo de amigos tan grande, dentro y fuera del laboratorio. A Fani por enseñarme, ayudarme, escucharme y abrirme un nuevo mundo de oportunidades. Nunca dejes de luchar por tus sueños. Gracias por confiar en mí. A Walter por su alegría y por ser un profesional de los pies a la cabeza. Aún tengo ganas de uno de esos cursos de bautismo (de buceo, no nos echemos las manos a la cabeza). A Alba por ser siempre una persona tan maravillosa. Necesito que me expliques el secreto para mantener siempre una sonrisa en la cara, llueve o truene. Y en definitiva al resto de *la secta*, la que yo conocí: Adri, Anto, Aurora, Ceci, Lucía, Ric y Sara. Cada uno me habéis aportado para ayudarme a ser lo que soy hoy. Bien saben cada una de las horchatas, paellas y buñuelos, cada grano de arena de la playa y mililitro de crema solar, que luché por entrar en *la secta*, pero nada oye, no hubo manera. Esa manera de ser felices y ese buen rollo, patentadlo.

La familia que se elige, qué suerte tuve con la mía. Casas rurales, cumpleaños, quedadas en cualquier ciudad o país. La suerte de generar una amistad como esta, imperturbable por el paso del tiempo o las condiciones personales de cada uno, es oro. A Mer, Barbi, Marguy, CdC, CrisR, María, Angela, Anabel, Álvaro, Marta, La negra y a cada uno de vuestros *+1*. A mis niños, *osú* que cuatro. Curro, Carlos, Javi y Paco, *¡Ay, mi Parco!*. No sería de justicia poner solo una frase relacionada con

cada uno, debería escribir un libro entero en cada caso. Sin vosotros yo sería la mitad, quizá ni un cuarto. *Coatí* hasta la tumba.

Y a esa familia, la que fue mi primera familia cuando yo vivía aún asustado dentro de mi cascarón. Gracias Vane, Natalia y Anita por ser mis amigas entonces y ahora, por enseñarme que la locura es un estado más de la cordura y que sonriendo se vive más feliz.

Y este viaje termina donde empezó, en Sevilla, mi remanso de paz. A los que siempre estuvieron, están y estarán, defendiendo Sevilla Este. A Ignacio y Sandra por tan buenas charlas enfrentando nuestros puntos de vista sobre la vida, acompañadas siempre por un juego de mesa o un mando. A Tocho, por ayudarme a ahogar mis neuronas siempre que lo he necesitado, dentro de cualquiera de los significados que esa expresión pueda adquirir. A Juan Carlos por haber estado siempre ahí, por motivarme con su ilusión. A Andrés por acompañarme en las noches de vigilia frente a la pantalla. A Ale por acompañarme en esos paseítos de despeje diarios. A Vilches por haber recorrido junto a mí este camino durante tanto tiempo. Amigos. Que nunca nos falte un juego de mesa, una película, una partida a algún videojuego aleatorio, una noche hasta las tantas. Que *Casa Igna sea* nuestro punto de reunión y la *Batcueva* nuestro templo. *Dr. Google is in da jaus*, gracias también por vuestro apoyo Jimmy y Jolas.

A mi familia. A mis primas y primos, a mis tías. Gracias a mi madre por su fortaleza, atención y dedicación. Por no dejar que pasaran dos días sin hablar por teléfono con este hijo tan *dejao* que le ha tocado. A mi padre por preocuparse por mí siempre y ayudarme en lo imposible. A mi hermana por su apoyo. A mis abuelos y abuelas, quienes me apoyaron y confiaron en mí siempre, hasta el último día. Sin vosotros esta historia no habría podido tener un principio ni un final. Os quiero.

```
for i in {1..1000000}; do echo 'Gracias';done
```

If I got rid of my demons,
I'd lose my angels too

Tennesse Williams

ÍNDICE

ABREVIATURAS Y ANGLICISMOS

ADN Ácido desoxirribonucleico

AMG *Auxiliary metabolic genes*

ARN Ácido ribonucleico

Barcode Etiqueta o secuencia de unos pocos nucleótidos que sirve para identificar el origen de la muestra de la cual procede la secuencia

Biofilm Biopelícula formada por microorganismos

BIOM *Biological observation matrix*

BLAST *Basic local alignment search tool*

CBMSO Centro de Biología Molecular Severo Ochoa

Cluster En esta Tesis se utiliza este anglicismo bajo dos acepciones diferentes: conjunto de unidades de cómputo para el análisis bioinformático de los datos y agrupación de *contigs* que presentan un grado de similitud de secuencia determinado

COG *Clusters of orthologous groups*

Contig Secuencia de ADN ensamblada *de novo* a partir de lecturas secuenciadas

Core Microorganismos propios de un ambiente y compartidos entre una mayoría de los individuos en los que ese ambiente ha sido estudiado

CPU *Central processing unit*

CRISPR *Clustered regularly interspaced short palindromic repeats*

CT *Cycle threshold*

EAR Estomatitis Aftosa Recurrente

EDTA *Ethylenediaminetetraacetic acid*

EGTA *Ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid*

FDA *Food and drug administration*

FDR *False discovery rate*

FISABIO Fundación para el Fomento de la Investigación Sanitaria y Biomédica

FISH *Fluorescence in situ hybridization*

GAAS *Genome relative Abundance and Average Size*

GB *Gigabytes*

HMP *Human microbiome project*

HOMD *Human oral microbiome database*

HPV *Human papilloma virus*

HTLV *Human T-lymphotropic virus*

KEGG *Kyoto encyclopedia of genes and genomes*

LASL *Linker amplified shotgun library*

LB *Lysogeny broth*

MCL *Markov cluster algorithm*

MCS *MiSeq control software*

MDA *Multiple displacement amplification*

Megacluster En esta Tesis se utiliza este anglicismo para definir la agrupación de *contigs* completos o casi completos en función de la similitud de sus secuencias medida según la distancia Sørensen-Dice modificada

MEGAN *Metagenome Analyzer*

MRS *MiSeq reporter software*

MRSA *Methicillin-resistant Staphylococcus aureus*

MVM *Minute virus of mice*

NGS *Next generation sequencing*

NMDS *Non-metric multidimensional scaling*

ORF *Open reading frame*

OTU *Operational taxonomic unit*

PCM Parque científico de Madrid

PCR *Polymerase chain reaction*

PFAPA *Periodic fever, aphthous stomatitis, pharyngitis and adenitis*

PHACCS *Phage communities from contig spectrum*

PHAST *Phage search tool*

PVDF *Polyvinylidene fluoride*

RAM *Random-access memory*

RPKM *Reads per kilobase of contig, per million mapped reads*

Score Puntuación asignada por *BLAST* para un alineamiento

SDS *Sodium dodecyl sulfate*

Shotgun Método de fragmentación al azar del ADN empleado para la preparación de librerías de secuenciación masiva

SISPA *Sequence-independent, single-primer amplification*

TB *Terabytes*

TTMV Torque Teno mini virus

UPV Universidad del País Vasco

VIH Virus de la inmunodeficiencia humana

WGA *Whole genome amplification*

WGS *Whole genome sequencing*

WR *Western Reserve*

SUMMARY

Viruses are key players regulating microbial ecosystems. Exploration of viral assemblages is now possible thanks to the development of metagenomics, the most powerful tool for studying viral ecology. Unfortunately, several sources of bias lead to the misrepresentation of certain viruses within metagenomics workflows. The oral cavity is a major portal of entry for human viruses, but it is dominated by highly personalized and time-persistent bacteriophage assemblages. Most of them follow lysogenic life cycles, deploying complex strategies to manage bacterial homeostasis. Although bacterial dysbiosis underlies common oral pathologies, the cause of these bacteria replacements remains obscure, and it is theorized that bacteriophages might play an important role.

In this thesis, we assessed the bias induced by viral enrichment and random amplification methods on mock assemblages of DNA viruses and human saliva viromes, using qPCR and deep sequencing. We observed that low-force centrifugation, 0.45µm filtration and iodixanol cushions preserved the original composition of nuclease-protected viral genomes. Comparison of unamplified and randomly amplified saliva viromes revealed that multiple displacement amplification induced stochastic bias from picograms of DNA template, but systematic bias from nanograms. This systematic bias was mainly due to under-amplification of sequences with extreme %GC, a negative bias shared with PCR-based methods that showed, in addition, high-coverage peaks in sequences with low linguistic complexity. Ordination plots of contig profiles showed overlapping of related amplified and unamplified saliva viromes and strong separation from unrelated saliva viromes. This result suggests that random amplification bias has a minor impact on beta diversity studies.

In order to gain insight into the diversity of viruses in the oral cavity, and their contribution to maintain healthy bacterial communities, we addressed the largest metagenomic study of viruses reported to date for this ecosystem: 43 Gbp. Oral viromes were dominated by bacteriophages of the *Caudovirales* order, with a small percentage of eukaryotic viruses, including four new human viruses of the *Anelloviridae* and *Papillomaviridae* families. 444 nearly full-length viral genomes, grouped in 31 megaclusters, were poorly related to those in the databases. However, some of the most abundant and ubiquitous megaclusters were also found in other published oral viromes. Bacteria communities differs substantially between dental plaque and oral mucosa, whereas this difference was at the limit of the statistical significance for virus assemblages. No differences in the microbiomes and viromes were found between healthy individuals and patients affected by caries or recurrent aphthous stomatitis.

Bacteriophage host-prediction by five consistent bioinformatic approaches revealed that *Actinobacteria* was the most frequent host in the oral cavity, in spite of its fourth position in 16S metagenomes of the same samples. Clustering of bacteriophages that infect the same phylum supports the importance of virus-host coevolution, and suggests that the host phylum is better taxonomic criteria than virus morphology for classification within the *Caudovirales* order. The identification of hundreds of lysins encoded by oral bacteriophages, together with their host-range prediction, might boost enzibiotics research as an alternative to antibiotics for restoring healthy oral microbiota.

RESUMEN

Los virus juegan un papel clave en la regulación de los ecosistemas microbianos. La metagenómica nos permite explorar en detalle las comunidades de virus y es la herramienta más potente para estudiar su ecología. Desafortunadamente, varios de sus pasos introducen sesgos que alteran la composición original de estas comunidades. La cavidad bucal es el portal de entrada más importante para los virus humanos y está dominada por poblaciones de bacterias que difieren entre individuos pero persisten en el tiempo. La mayoría son lisogénicos y despliegan múltiples estrategias para mantener el equilibrio de las comunidades bacterianas. En algunas patologías frecuentes de la boca la microbiota está alterada y se desconoce si los bacteriófagos juegan un papel importante.

En esta tesis evaluamos el sesgo introducido durante los pasos de enriquecimiento de virus y amplificación al azar de sus genomas en comunidades sintéticas de virus de ADN y en viromas de saliva humana. Los pasos de centrifugación a baja velocidad, filtración en 0,45µm y colchones de iodixanol preservan la composición original de virus. Sin embargo, la comparación de viromas de saliva antes y después de la amplificación reveló que la amplificación por desplazamiento múltiple de banda inducía un sesgo estocástico desde picogramos de ADN molde, y sistemático desde nanogramos. El sesgo sistemático en viromas de saliva se debió principalmente a la peor amplificación de secuencias con %CG extremos, un sesgo negativo que es compartido con todos los métodos basados en *PCR*, que a su vez muestran picos de alta cobertura en secuencias de baja complejidad lingüística. Un sistema de ordenación de viromas en base a las abundancias de *contig* compartidos mostró solapamiento de los viromas de saliva amplificados y no amplificados procedentes de la misma muestra y una fuerte separación de los procedentes de individuos distintos. Estos resultados sugieren que los sesgos de la amplificación apenas afectan a los estudios de diversidad beta.

Con el objetivo de conocer mejor la diversidad de virus en la cavidad oral y cuál es su contribución al mantenimiento de comunidades bacterianas saludables, hemos abordado el estudio metagenómico de virus más importante que se ha hecho hasta la fecha para este ecosistema: 43 Gpb. Los viromas de la boca estaban dominados por bacteriófagos del orden *Caudovirales*, con un pequeño porcentaje de virus eucarióticos, incluyendo cuatro virus humanos nuevos de las familias *Anelloviridae* y *Papillomaviridae*. Los 444 genomas virales casi completos que hemos obtenido estaban lejanamente relacionados con los virus de las bases de datos y se agrupaban en 31 *megaclusters*. Algunos de los más abundantes y ubicuos en nuestros viromas, lo son también en otros viromas bucales publicados. Las comunidades bacterianas difieren sustancialmente entre placa dental y mucosa bucal, pero esta diferencia estaba en el límite de la significación estadística para los viromas. No encontramos diferencias en los microbiomas, ni en los viromas, de individuos sanos y afectados por caries o estomatitis aftosa recurrente. Empleando cinco aproximaciones bioinformáticas distintas para la predicción del hospedador de los bacteriófagos obtuvimos resultados consistentes, siendo *Actinobacteria* el hospedador más frecuente en la cavidad bucal, a pesar de ocupar sólo la cuarta posición en los microbiomas 16S de las mismas muestras. La agrupación de los bacteriófagos que infectan bacterias del mismo filo respalda la importancia de la co-evolución entre virus y hospedadores, y sugiere que el filo del hospedador es un criterio taxonómico

mejor que la propia morfología de los virus para la clasificación de los *Caudovirales*. La identificación de cientos de lisinas codificadas en los genomas de los bacteriófagos bucales, junto con la predicción de su hospedador, podría impulsar la investigación de los enzibióticos, como alternativa a los antibióticos, para restaurar una microbiota bucal saludable.

INTRODUCCIÓN

1. Microbiota humana

Las comunidades de microorganismos representan un porcentaje significativo de la masa total de nuestro cuerpo, y en términos numéricos, las células procariotas de estos ecosistemas microbianos superan en diez veces el número de células humanas (Sender et al., 2016). Estas comunidades están formadas por microorganismos de las tres ramas del árbol de la vida: bacterias, eucariotas y arqueas, e incluyen también un número aún mayor de virus. En 2001 se utilizó por primera vez el término *microbioma* para nombrar a "la comunidad ecológica de microorganismos comensales, simbiontes y patogénicos que literalmente comparten nuestro espacio corporal" (Lederberg y McCray, 2001). Hoy en día, está más aceptado el término *microbiota* para definir a dicha comunidad de microorganismos, mientras que se utiliza *microbioma* para referirse al conjunto de genes de dicha comunidad (Clemente et al., 2012). Gracias al desarrollo de la metagenómica, impulsada por la aparición de las técnicas de secuenciación masiva, disponemos de un mejor conocimiento de la diversidad de microorganismos en ambientes como la piel (Oh et al., 2016), la placenta (Aagaard et al., 2014), el tracto respiratorio (Cui et al., 2014), la saliva (Nasidze et al., 2009), la mucosa oral (Dewhirst et al., 2010), la vagina (Goltsman et al., 2018) o el tracto gastrointestinal (Quigley, 2013). Esta diversidad varía en individuos sanos dependiendo del compartimento estudiado, encontrando entre 152 y 17.546 especies diferentes de bacterias en ojo e intestino, respectivamente (Lloyd-Price et al., 2016). Además de bacterias, la microbiota humana está compuesta por varias especies diferentes de arqueas (Dridi et al., 2011) y una importante diversidad de hongos, como los 40-80 géneros encontrados en algunas zonas de la piel (Findley et al., 2013). Otros constituyentes patogénicos y comensales habituales pero no tan frecuentes son los protozoos (blastocystis, parabasálidos, entamoeba y chilomastix principalmente) y los nemátodos (Parfrey et al., 2014; Wade, 2013).

La microbiota intestinal humana juega un papel fundamental en el metabolismo humano: permiten una absorción eficiente de los nutrientes (Krajmalnik-Brown et al., 2012), nos proporciona vitaminas como las del grupo B (folatos, riboflavina o vitamina B12) que no podemos sintetizar, producidas principalmente por bacterias ácido lácticas (Gu y Li, 2016), y es crucial para un correcto desarrollo del sistema inmunitario (Ximenez y Torres, 2017). Un claro indicador de la importancia de la microbiota en nuestra salud es la conexión entre una microbiota intestinal alterada y el desarrollo de enfermedades mentales como el Alzheimer (Kowalski y Mulak, 2019), esquizofrenia (Lv et al., 2017) o la depresión (Winter et al., 2018).

Pese a superar en diez veces el número de microorganismos celulares, los virus siguen siendo el componente menos estudiado de la microbiota humana. Las comunidades virales asociadas a humanos están ampliamente dominadas por virus que infectan bacterias (bacteriófagos) (Breitbart et al., 2003).

La mayor parte de estos virus no han sido aún aislados y su papel en la regulación de las comunidades bacterianas no ha sido estudiado en profundidad. Junto a los bacteriófagos, existe también una menor proporción de virus patógenos humanos mejor caracterizados (Moustafa et al., 2017; Wylie et al., 2014).

1.1. Metagenómica

La metagenómica, impulsada por el desarrollo de tecnologías de secuenciación masiva de ADN, ha desempeñado un papel clave para conocer los microbiomas humanos. Tradicionalmente la obtención de información genética de las bacterias presentes en ecosistemas microbianos ha presentado dificultades derivadas de problemas para cultivar en laboratorio la inmensa mayoría de los microorganismos que existen en la naturaleza (hecho conocido como la "anomalía del contaje en placa") (Staley y Konopka, 1985), junto con el elevado coste de la secuenciación de ADN. Con el propósito de superar esta limitación surge la metagenómica como: "la aplicación de técnicas genómicas modernas para el estudio directo de comunidades de microorganismos en su entorno natural, evitando la necesidad de aislar y cultivar cada una de las especies que componen la comunidad" (Chen y Pachter, 2005). El primer estudio metagenómico de procariotas se llevó a cabo en aguas hidrotermales secuenciando el gen marcador que codifica por ARN ribosómico 5S (Stahl et al., 1984). Pocos años después se publicó el primer estudio metagenómico basado en la secuenciación del gen que codifica para el ARN ribosómico 16S (Schmidt et al., 1991; Woese, 1987). La secuenciación de amplicones de este gen marcador sigue siendo, a día de hoy, el método más empleado para explorar comunidades bacterianas. Posteriormente, la metagenómica funcional permitió identificar genes con funciones específicas presentes en bacterias no aisladas mediante la preparación de librerías de ADN provenientes de muestras naturales en plásmidos o en cromosomas artificiales de bacterias (BACs) (Handelsman et al., 1998). La secuenciación de estas librerías se denomina metagenómica *shotgun* o secuenciación de genoma completo (Breitbart et al., 2002; Tyson et al., 2004; Venter et al., 2004). Existe en la actualidad cierta controversia acerca de si el término metagenómica debe ser empleado en exclusiva para la secuenciación directa de genomas fragmentados (metagenómica s*hotgun*), al ser el único método que proporciona información de todos los genes de los genomas de la comunidad. Sin embargo, otros autores piensan que también es adecuado su uso para la secuenciación de genes marcadores como ARNr 16S, ya que este método proporciona información genética de la comunidad completa de microorganismos de un ambiente aunque esté basada en el estudio de un único gen conservado (Laudadio et al., 2018). El estudio metagenómico de las comunidades de virus (viromas) presenta una serie de dificultades adicionales a las de los microorganismos celulares. La principal es la ausencia de genes marcadores universales entre todas las familias virales, lo que obliga a la utilización de técnicas metagenómicas *shotgun* de secuenciación de genomas completos, una aproximación técnica y económicamente más costosa que requiere un análisis más complejo (Thurber et al., 2009). Además, debido al menor tamaño que tienen los genomas virales en comparación con los celulares, los protocolos de metagenómica de virus exigen una serie de pasos

iniciales de enriquecimiento de partículas virales y amplificaciones inespecíficas que pueden introducir sesgos en la composición de la comunidad (**sección 2.3. de Introducción**) (Parras-Moltó y López-Bueno, 2018). Pese a ello, la metagenómica de virus está provocando una verdadera revolución en el campo de la virología, permitiendo vislumbrar la enorme diversidad genética de las comunidades naturales de virus, lo que representa un primer paso necesario para entender cómo los virus regulan las comunidades microbianas. Una ventaja sobre los metagenomas de bacterias es que, como consecuencia del menor tamaño de los genomas virales, es factible ensamblar varios genomas completos de virus directamente desde metagenomas, facilitando la identificación de nuevas especies.

1.2. Secuenciación masiva

El impulso definitivo de la metagenómica ocurrió a mediados de la década pasada, gracias al desarrollo de la secuenciación masiva o de siguiente generación (*next generation sequencing, NGS*). Estas técnicas se basan en la miniaturización y paralelización de las reacciones de secuenciación incrementando exponencialmente el número de secuencias obtenidas y reduciendo enormemente su coste económico (Shendure y Ji, 2008). Actualmente, la tecnología de secuenciación masiva más utilizada es la comercializada por Illumina®. Esta compañía ofrece una amplia variedad de plataformas de secuenciación por síntesis que permiten obtener, desde 4 millones de secuencias en los equipos *iSeq 100 System*, hasta 20.000 millones de secuencias de los equipos *NovaSeq 6000 System q* (https://www.illumina.com/). Frente a las secuencias pareadas cortas (hasta 2x300pb) de alta calidad (>99,5% de fiabilidad) generadas por los secuenciadores de Illumina®, han surgido en los últimos años otras plataformas que hacen secuenciación de molécula única en tiempo real. Estas tecnologías proporcionan secuencias de menor calidad pero de mayor tamaño. Así, Pacific Biosciences (PacBio®, https://www.pacb.com/) genera 0,05-0,5 millones de secuencias por carrera con un tamaño medio de 10-35 kpb. En cuanto a la calidad de las secuencias, éstas presentan un error medio de un 13%, que se corrige a través de la secuenciación repetida de la misma molécula, generando "lecturas circulares consenso cortas" con una calidad superior al 99% (Pootakham et al., 2017). Otra compañía que compite en el ámbito de secuencias de gran tamaño es Oxford Nanopore Technologies© (https://nanoporetech.com/). Esta compañía proporciona un número muy variable de secuencias, que al igual que PacBio®, presentan aún una baja fiabilidad de secuencia, pero que pueden llegar hasta los dos millones de nucleótidos (Payne et al., 2018) en equipos de muy pequeño tamaño conectados a un ordenador portátil o a un teléfono móvil (Quick et al., 2014).

1.3. Impacto de la metagenómica en el estudio de la microbiota humana

El uso de las técnicas de secuenciación masiva en estudios metagenómicos ha provocado una auténtica revolución en el conocimiento de la microbiota humana. Uno de los proyectos de metagenómica más importantes llevados a cabo hasta la fecha ha sido el Proyecto Microbioma Humano (*HMP*, (Human

Microbiome Project Consortium, 2012). Sólo en este proyecto se generaron 5.177 metagenomas humanos de hasta 14 ecosistemas diferentes de 242 individuos sanos distintos. Alguno de sus objetivos principales fueron: entender cuáles son los componentes microbianos que contribuyen a la fisiología normal de nuestro cuerpo en estado de salud, analizar la estabilidad de nuestra microbiota en el transcurso del tiempo, estudiar las diferencias entre la microbiota de individuos sanos o la existencia de un núcleo (*core*) microbiano común a todos los humanos (Nash et al., 2017; Turnbaugh et al., 2009). Desde entonces se han llevado a cabo un gran número de estudios comparativos de los microbiomas en individuos sanos y afectados por algún tipo de enfermedad que han desvelado en algunos casos importantes alteraciones en las comunidades de bacterias en situaciones de enfermedad (disbiosis). Por ejemplo, en obesidad se ha observado un aumento de *Firmicutes* en detrimento de *Bacteroidetes* (Ley et al., 2006), en cirrosis se ha observado un aumento de *Streptococcus* y *Veillonella* (Qin et al., 2014). También se ha detectado disbiosis en enfermedades como Crohn (Wang et al., 2016), diabetes (Qin et al., 2012) o incluso enfermedades mentales (Kowalski y Mulak, 2019; Lv et al., 2017; Winter et al., 2018). Un objetivo común a muchos de estos estudios es la identificación de biomarcadores de enfermedad con valor clínico. En este sentido, un campo actualmente activo en investigación es la identificación con potentes herramientas estadísticas de determinados genes directamente en metagenomas *shotgun* como posibles biomarcadores de enfermedad (Wu et al., 2018).

El enorme volumen de información generado en los estudios genómicos y metagenómicos ha propiciado el desarrollo de una gran variedad de herramientas bioinformáticas y métodos estadísticos para su análisis (Afgan et al., 2018; Caporaso et al., 2010; Package et al., 2000), así como la ampliación de las infraestructuras que mantienen las diferentes bases de datos *online* de secuencias y cuyo crecimiento ha sido exponencial desde la aparición de la secuenciación masiva (Benson et al., 2018; Chen et al., 2017; Glöckner et al., 2017) (**Fig. 1**).

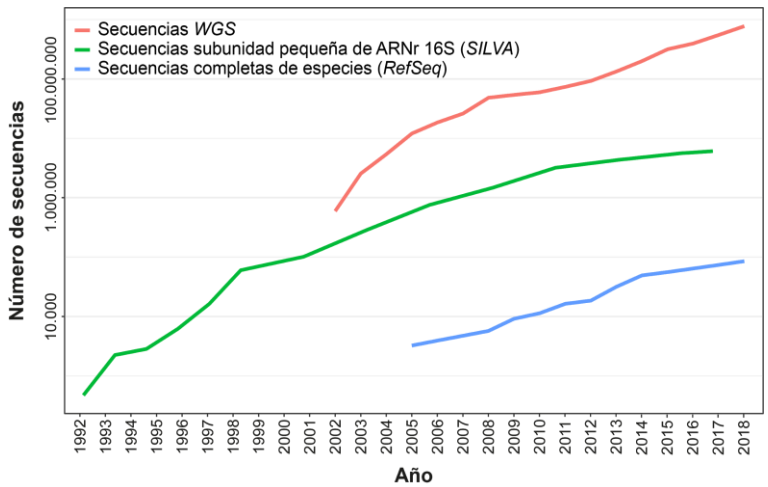

Figura 1. Evolución del número de secuencias contenidas en las bases de datos *WGS* (*GenBank*) (https://www.ncbi.nlm.nih.gov/genbank/statistics/), *RefSeq* (*GenBank*) y *SILVA* (https://www.arb-silva.de/). Desde 1992, el número de nucleótidos depositados en el *GenBank* se ha duplicado aproximadamente cada 18 meses.

2. Estudio de las comunidades virales

Los virus constituyen el componente más abundante y genéticamente diverso de los ecosistemas microbianos, pero también el menos conocido. Se estima que la biosfera alberga $1x10^{31}$ partículas virales (Fuhrman, 1999; Suttle, 2007) que encierran una enorme diversidad genética, en su mayor parte desconocida (Angly et al., 2006). Pese a la visión negativa que les hemos atribuido por su capacidad de producir enfermedades en humanos, animales de ganadería y en cultivos de plantas, la inmensa mayoría de los virus que existen en la naturaleza son virus que infectan otros microorganismos. Los bacteriófagos son los principales constituyentes de estas poblaciones. Los bacteriófagos fueron descubiertos a principios del siglo XX (d'Herelle, 1917; Twort, 1915) y su estudio ha sido clave para alcanzar muchos de los hitos más importantes de la biología molecular (Salmond y Fineran, 2015). El papel que estos bacteriófagos desempeñan en el control de las comunidades de bacterias de nuestro cuerpo ha sido poco estudiado.

2.1. Control de las comunidades microbianas por bacteriófagos

El modelo más sencillo de regulación de la composición microbiana por virus es el conocido como *Killing-the-Winner* o "muerte del ganador". Este modelo de interacción está basado en una relación depredador-presa en la que la actividad lítica de algunos virus controla la abundancia de aquellos hospedadores que superan un determinado umbral de densidad. De esta forma, consiguen amortiguar cambios drásticos en la población bacteriana manteniendo una alta diversidad (Rodriguez-Brito et al., 2010; Winter et al., 2010). Este tipo de interacción tiene un papel muy importante en el control de comunidades microbianas en ambientes naturales, sin embargo, parece no ser tan importante en los ecosistemas microbianos asociados a humanos, ya que muchos viromas humanos están dominados por bacteriófagos de la familia *Siphoviridae* con un ciclo vital predominantemente lisogénico (Minot et al., 2011; Reyes et al., 2010). En general, la relevancia del ciclo lisogénico de los bacteriófagos queda de manifiesto al observar que en torno a un 60-70% de todos los genomas bacterianos secuenciados contienen profagos (Casjens, 2003; Paul, 2008). La dinámica de interacción entre los bacteriófagos lisogénicos y sus hospedadores es mucho más compleja y variada que la de los virus exclusivamente líticos. Existen muchos ejemplos de relaciones de mutualismo donde los profagos pueden proporcionar a sus hospedadores una serie de genes que mejoran su viabilidad, como genes de resistencia a antibióticos (Quirós et al., 2014), genes de virulencia como la toxina colérica codificada en el profago CTX integrado en el genoma de *Vibrio cholerae* (Waldor y Mekalanos, 1996), genes de la maquinaria fotosintética (Sharon et al., 2009), o del metabolismo de aminoácidos o azúcares, denominándose estos últimos como "genes metabólicos auxiliares" (*Auxiliary Metabolic Genes, AMG*) (Breitbart et al., 2007). Un ejemplo de la complejidad de este tipo de asociaciones es la transferencia horizontal de *CRISPRs* mediada por bacteriófagos, donde el hospedador puede adquirir inmunidad frente a otros bacteriófagos y el bacteriófago propaga un mecanismo de defensa contra bacteriófagos competidores (Minot et al.,

2013, 2011). En términos globales, la relación entre lisogenia y densidad de hospedadores no está clara. Algunos estudios proponen que el estado de lisogenia es más frecuente a baja densidad de hospedador (Jiang y Paul, 1998; Paul, 2008), apoyando la idea del modelo *Kill-the-Winner*, donde la transición entre lisogenia y lisis se da por el aumento de la densidad de éstos. Sin embargo, estudios recientes proponen un modelo que explicaría la convergencia de una elevada proporción de bacteriófagos lisogénicos con densidades de hospedadores altas (*Piggyback-the-Winner*) (Knowles et al., 2016).

Con respecto a la incidencia directa que los bacteriófagos pueden tener en la salud humana, se ha propuesto un mecanismo por el cual los bacteriófagos se disponen en las mucosas como barrera no inmunológica ante la posible colonización por bacterias patógenas (Barr et al., 2013).

2.2. Metagenómica de virus

Pese a que los virus dominantes en la mayor parte de los ambientes naturales son bacteriófagos, casi todos los esfuerzos de la comunidad científica se han centrado en estudiar virus eucarióticos que infectan humanos, animales o cultivos de plantas. Además, la caracterización de nuevos bacteriófagos ha estado tradicionalmente limitada por las dificultades para cultivar sus hospedadores. Como consecuencia, las comunidades de virus se han estudiado en menor profundidad que el resto de microorganismos celulares, y las bases de datos de genomas virales son las que tienen menos información. Por todo ello, la metagenómica ha tenido un gran impacto dentro del campo de la virología, permitiéndonos acceder a estas comunidades de bacteriófagos pobremente estudiadas (Angly et al., 2006; Reyes et al., 2010), así como identificar nuevos virus humanos, como nuevas variantes divergentes del virus ébola (Towner et al., 2008), nuevos rhabdovirus (Grard et al., 2012) o arenavirus (Briese et al., 2009). A diferencia de los microorganismos celulares, los virus no poseen un gen marcador universal conservado entre todas sus familias virales, por lo que el estudio de los viromas está basado en la secuenciación de genomas completos fragmentados (metagenómica *shotgun*). Teniendo en cuenta la gran cantidad de secuencias obtenidas en estos estudios de metagenómica *shotgun* con los sistemas de secuenciación *NGS* actuales, y el pequeño tamaño de los genomas virales en comparación con los genomas celulares, resulta relativamente fácil ensamblar *de novo* varios genomas completos de virus directamente desde metagenomas.

En los últimos años hemos asistido a la publicación de numerosos estudios metagenómicos de virus procedentes de diferentes hábitats del cuerpo humano. Uno de los ecosistemas microbianos mejor estudiados mediante técnicas metagenómicas es el tracto intestinal. Los primeros estudios estimaron que la diversidad de bacteriófagos en este compartimento varía entre 10-984 genotipos diferentes (Minot et al., 2011; Reyes et al., 2010), donde las familias más representadas son *Siphoviridae*, *Myoviridae*, *Podoviridae*, en mucha menor proporción *Microviridae*, y un porcentaje muy menor de virus eucarióticos. Además, en estos trabajos se demostró que los viromas del tracto intestinal son muy

diferentes en su composición de bacteriófagos entre individuos adultos distintos, incluso entre gemelos monocigóticos (Reyes et al., 2010), aunque son similares durante la infancia (Lim et al., 2015; Reyes et al., 2015). Pese a la gran variabilidad del viroma intestinal entre adultos, también se ha podido identificar un núcleo de bacteriofagos compartidos (Broecker et al., 2017). Otra característica de estos viromas es que son estables en el tiempo (Reyes et al., 2010), aunque el tipo de dieta puede ser uno de los factores que explique su variabilidad entre individuos distintos (Minot et al., 2011). En este sentido, también se ha observado que estos viroma tienden a converger en individuos que conviven en la misma casa (Ly et al., 2016). Estudios recientes han observado que las comunidades virales pueden variar en enfermedades como Crohn o colitis ulcerosa, donde se ha observado un aumento en la diversidad de *Caudovirales* en paralelo a una reducción de la diversidad de sus hospedadores (Norman et al., 2015), o en infecciones crónicas de *Clostridium difficile* donde se ha observado una disminución de la riqueza de *Caudovirales* (Zuo et al., 2018).

El estudio de la comunidad de virus de ARN ha sido abordado en numerosas ocasiones para buscar posibles agentes etiológicos de enfermedades derivadas del tracto intestinal. El primer estudio de viromas de ARN en heces humanas encontró que el virus mayoritario es un virus de plantas proveniente de la ingesta de alimentos (*Pepper mild mottle virus*) (Zhang et al., 2006). También se han detectado otros virus que pueden tener un impacto mayor en la salud del tracto intestinal como picornavirus y picobirnavirus en pacientes de diarrea (Holtz et al., 2014; Smits et al., 2014), y varios picornavirus en parálisis flácida (Victoria et al., 2009). Por el contrario, otros estudios han determinado que algunos virus como los norovirus puede ejercer una función homeostática en el intestino estimulando el correcto desarrollo de la microvellosidades intestinales y supliendo la función de bacterias comensales (Kernbauer et al., 2014).

Aunque en menor profundidad, también se han estudiado los viromas de ADN de otros ecosistemas asociados a humanos. Los viromas de orina (Rani et al., 2016; Santiago-Rodriguez et al., 2015) están también dominados por bacteriófagos con una pequeña proporción de virus eucarióticos como papilomavirus, polyomavirus y anellovirus. El viroma de la piel está dominado por bacteriófagos de las familias *Siphoviridae* y *Myoviridae* con un importante repertorio de genes de virulencia y resistencia a antibióticos, así como genes indicativos de un ciclo vital predominantemente lisogénico. La piel también contiene una población minoritaria de virus eucarióticos como papilomavirus y Poxvirus (Hannigan et al., 2015). En sangre, se han empleado técnicas metagenómicas diversas que han permitido encontrar una gran abundancia de secuencias de virus de las familias *Anelloviridae* y *Herpesviriridae*, seguidas de otros virus menos frecuentes como adenovirus, VIH, *HTLV*, los virus de la hepatitis B y C, parvovirus B19 o el virus de la gripe (Breitbart y Rohwer, 2005; Furuta et al., 2015; Moustafa et al., 2017). En muestras de ADN libre en sangre de personas tratadas con inmunosupresores y antivirales antes de recibir un trasplante de órganos se ha observado un aumento drástico en los niveles de

secuencias de anellovirus, pudiendo representar un marcador útil de inmunodepresión (De Vlaminck et al., 2013).

2.3. Sesgos metodológicos en el estudio metagenómico de las comunidades de virus

El estudio metagenómico de las comunidades virales requiere en muchos casos la aplicación de metodologías de enriquecimiento de partículas virales debido a la mayor cantidad de material genético procedente de organismos celulares. Desafortunadamente, estos procedimientos pueden alterar sustancialmente la composición original de las comunidades virales. Los protocolos de enriquecimiento más usados (Parras-Moltó y López-Bueno, 2018; Pride et al., 2011a; Thurber et al., 2009) se basan en centrifugaciones a baja velocidad, filtraciones con filtros de 0,22-0,45μm, ruptura de membranas celulares con cloroformo, gradientes de CsCl y tratamientos con nucleasas para la digestión de material genético no protegido. Alguno de estos pasos, como los gradientes de CsCl, sesgan las poblaciones virales hacia los bacteriófagos con colas (Castro-Mejía et al., 2015; Kleiner et al., 2015; Thurber et al., 2009). El cloroformo rompe la membrana de los virus con envuelta (Willner et al., 2011) y la elección entre los filtros de 0,22 o 0,45μm durante la filtración tiene por un lado un efecto en el grado de contaminación bacteriana, pero por otro lado también influye en la eliminación de virus de gran tamaño como los miembros de las familias *Poxviridae*, *Asfarviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnavirus*, *Mimiviridae* y *Marseilleviridae* (Colson et al., 2013; Hoyles et al., 2014; Popgeorgiev et al., 2013), así como los llamados bacteriófagos Jumbo, con genomas de hasta 200 kpb (Yuan y Gao, 2017).

Otro de los puntos críticos en la obtención de viromas es la cantidad de genomas virales de ADN y ARN obtenida, que en muchos casos está por debajo de las cantidades mínimas requeridas para la preparación de librerías de secuenciación masiva. Esta limitación es especialmente crítica cuando se trabaja con muestras procedentes de ambientes extremos (López-Bueno et al., 2009) o con determinadas muestras humanas (Breitbart y Rohwer, 2005; Parras-Moltó y López-Bueno, 2018). Para solucionar este problema se emplean métodos de amplificación inespecífica o aleatoria del ADN viral. Estos métodos también pueden introducir sesgos que alteren las proporciones relativas de virus presentes en la comunidad original. Entre los métodos más usados se encuentran la "amplificación de cebador único independiente de secuencia" (*SISPA*) (Breitbart y Rohwer, 2005; Froussard, 1992), que se basa en el uso de oligonucleótidos pseudo-degenerados de aproximadamente 20 nucleótidos con una región de 6-12 nucleótidos aleatorios en el extremo 3' y una región de secuencia conocida en el extremo 5' que permite su posterior amplificación por *PCR*. Este tipo de amplificación puede generar un sesgo debido a la preferencia de su región constante por secuencias similares de la muestra. Otra de las técnicas más utilizadas es la llamada *LASL* (Breitbart et al., 2002), basado en la amplificación por *PCR* de librerías obtenidas por la ligación de marcadores específicos en los extremos de las secuencias (Angly et al., 2006). Por último, otro de los métodos más utilizados es la amplificación por "desplazamiento múltiple de banda" (*MDA*) (Blanco et al., 1989; Dean et al., 2002), que a diferencia de los anteriores métodos

consiste en una amplificación isotérmica con cebadores aleatorios de 6 nucleótidos modificados y la polimerasa de alta fidelidad y procesividad del bacteriófago *Φ29*. Uno de los sesgos más conocidos de este método es la sobreamplificación de genomas circulares pequeños (Kim et al., 2008).

3. Cavidad bucal humana

La cavidad bucal constituye el punto de conexión entre el mundo exterior y los tractos respiratorio y gastrointestinal, siendo el principal portal de entrada de muchos patógenos. Las principales morfologías bacterianas se describieron en 1683 por Antonie van Leeuwenhoek utilizando lentes de aumento sencillas y muestras de placa dental (Anderson, 2016). Actualmente se estima que la boca alberga más de 1.000 especies bacterianas diferentes (Dewhirst et al., 2010), varios tipos de arqueas (Matarazzo et al., 2011; Vianna et al., 2006), protozoos (Belda-Ferre et al., 2012), y hongos (Ghannoum et al., 2010), además de entre 300 y 2.000 especies distintas de virus (Pride et al., 2011a). Estos microorganismos crecen en ambientes tan dispares como la mucosa oral o la placa dental, y están en continua comunicación por la acción de la saliva. Por otro lado, estas poblaciones están sujetas a profundos cambios fisicoquímicos, como la alteración del pH por la fermentación de los alimentos y la acción tamponadora de la saliva, la ingesta de alimentos muy variados o la exposición transitoria a compuestos como los del tabaco (Camelo-Castillo et al., 2015). Además, las poblaciones de microorganismos de la boca sufren disminuciones drásticas y periódicas derivadas de prácticas comunes de higiene oral como el cepillado y el enjuague con antisépticos (Dewhirst et al., 2010; Proctor y Relman, 2017; Wade, 2013). La alimentación juega un papel importante en el desarrollo de ciertas comunidades microbianas. Por ejemplo, el consumo de carbohidratos aumenta la concentración de *Lactobacillaceae* (Kato et al., 2017) o la de bacterias acidogénicas y acidúricas, como *Streptococcus mutans* (Morhart y Fitzgerald, 1976; Wade, 2013), considerada entonces como una de las causas primarias del desarrollo de caries.

La microbiota asociada a la placa dental evoluciona con los años. La colonización del diente comienza en el nacimiento (Nelson-Filho et al., 2013), y durante la niñez el diente es colonizado en primer lugar por una gran variedad de bacterias del género *Streptococcus* (alrededor de un 80% del total de bacterias) (Nyvad y Kilian, 1987) como *Streptococcus mitis* y *Streptococcus sanguinis*, seguidos en menor proporción por *Actinomyces* (Li et al., 2004). En la edad adulta se pueden detectar hasta 15 filos distintos, donde predominan *Fusobacteria*, *Firmicutes*, *Actinobacteria*, *Proteobacteria*, *Bacteroidetes* y *Espiroquetas* (Dewhirst et al., 2010). Se han detectado diferencias en la microbiota dental según el ambiente estudiado. Así, en la placa supragingival hay una mayor abundancia de *Actinomyces* y en placa subgingival de *Prevotella* (Ziouani et al., 2015). También se han observado diferencias importantes en la composición microbiana de la placa dental en función de la localización precisa de la pieza dental (Simón-Soro et al., 2013). En la mucosa oral de individuos adultos predominan bacterias del género *Streptococcus* seguidas de bacterias de los géneros *Haemophilus*, *Neisseria*, *Prevotella* y *Veillonella* (Bik et al., 2010). La saliva está formada por una mezcla de bacterias procedentes de todos los ambientes

de la boca, sin embargo su composición es más parecida a la de la mucosa oral que a la de la placa supragingival (Gomar-Vercher et al., 2018).

3.1. Comunidades virales de la cavidad bucal humana

La inmensa mayoría de los virus de la boca son bacteriófagos con cola del orden *Caudovirales*, principalmente de la familia *Siphoviridae*, que tienen un ciclo vital predominantemente lisogénico (Abeles et al., 2014), seguidos por bacteriófagos de la familia *Myoviridae*, de ciclo preferentemente lítico. Las comunidades de bacteriófagos de la boca son estables en el tiempo (Abeles et al., 2014; Ly et al., 2016) y muy diferentes entre individuos, ya que la mayoría de los virus son detectados en unos pocos individuos. Aunque tenemos un conocimiento menor comparado con los viromas del tracto intestinal, la boca también alberga algún bacteriófago y varios virus eucarióticos comunes entre individuos distintos (Pérez-Brocal y Moya, 2018; Wylie et al., 2014). En la cavidad bucal las comunidades de bacteriófagos de individuos que cohabitan tienden a converger (Ly et al., 2016; Robles-Sikisaka et al., 2013), sugiriendo cierto grado de plasticidad. A pesar de que estudios anteriores demostraron una relación entre el viroma y el género del donante de la muestra (Abeles et al., 2014), estudios recientes con un mayor número de muestras estudiadas han visto que no existe una tendencia clara (Pérez-Brocal y Moya, 2018). El impacto del tratamiento con antibióticos en la comunidad de bacterias de la boca también provoca alteraciones importantes en sus comunidades de virus (Abeles et al., 2015).

Pese a que las comunidades de virus de la cavidad bucal están dominadas por bacteriófagos, podemos encontrar algunos virus eucarióticos en proporciones bajas. En la mayoría de adultos es frecuente encontrar *Roseolovirus* (*HHV-6* y *HHV-7*), papilomavirus, anellovirus y adenovirus (Pérez-Brocal y Moya, 2018; Wylie et al., 2014).

3.2. Enfermedades de la cavidad bucal

Desde hace mucho tiempo se sabe que algunas bacterias juegan un papel fundamental en el desarrollo de enfermedades frecuentes de la cavidad bucal como caries o periodontitis (Marsh, 2010). En alguno de estos casos su participación en el proceso de la enfermedad parece claro, como el caso de *Streptococcus mutans* y el desarrollo de caries (Fitzgerald y Keyes, 1960), siendo una de las primeras especies, junto con *Streptococcus mitis*, *Streptococcus oralis* o *Streptococcus sanguinis*, en colonizar el *biofilm* de placa dental (Jenkinson y Lamont, 2005; Marsh, 2004). De igual modo, el grupo de bacterias que conforman la triada roja (*Porphyromonas gingivalis*, *Tannerella forsythensis*, *Treponema denticola*) juegan un papel fundamental en el desarrollo de periodontitis (Socransky et al., 1998). Estudios más recientes proponen que estas enfermedades pueden ser resultado de desequilibrios en las comunidades bacterianas, lo que implicaría a un número mayor de bacterias (Belda-Ferre et al., 2012; Hajishengallis, 2015; Jenkinson y Lamont, 2005; Mira et al., 2017). Además, una mala salud bucal o alteraciones en la microbiota oral se ha relacionado también con algunas enfermedades sistémicas como úlceras y cáncer

de estómago (Ndegwa et al., 2018), o con enfermedades cardiovasculares (Kholy et al., 2015). En este sentido, se ha demostrado la participación directa en enfermedad coronaria de algunos bacteriófagos frecuentes en la cavidad bucal como *Streptococcus mitis phage SM1*, que expresa proteínas de adhesión a plaquetas y facilita la formación de *biofilms* de *Streptococcus mitis*, lo que acaba produciendo endocarditis (Willner et al., 2011). La detección de altos niveles de *Streptococcus mitis* en boca puede ser indicador también de cáncer pancreático (Farrell et al., 2012), o altos niveles de *Capnocytophaga*, *Selenomonas*, *Veillonella* y *Neisseria* en muestras de saliva pueden tener una relación con el desarrollo de cáncer de pulmón (Yan et al., 2015). Más recientemente se ha observado que *Fusobacterium nucleatum*, procedente previsiblemente de la cavidad bucal, está implicado en el desarrollo de cáncer colo-rectal (Kostic et al., 2013; Rubinstein et al., 2013; Thomas et al., 2019). A pesar de las implicaciones médicas y económicas que tienen las enfermedades de la cavidad oral, a día de hoy no existen tratamientos eficientes y no invasivos frente a estas afecciones.

Aunque los hábitos alimenticios y la higiene oral determinan en gran parte la estabilidad de una microbiota oral sana (Morhart y Fitzgerald, 1976; Wade, 2013), se ha sugerido en algún caso y demostrado en otros, la presencia de algunas bacterias con potencial antagonista frente a las cariogénicas en personas que nunca han padecido caries (Belda-Ferre et al., 2012; Corby et al., 2005; López-López et al., 2017).

3.2.1. Caries

La caries es una enfermedad que afecta a cerca del 90% de la población mundial (Dye et al., 2015; Petersen, 2003), englobando un total de 3.900 millones de individuos con caries no tratadas (Marcenes et al., 2013), y constituye la enfermedad crónica más extendida entre los niños de entre 5 y 17 años a nivel global (Bagramian et al., 2009). La caries surge como resultado de la fermentación de carbohidratos por parte de bacterias acidogénicas que crecen en la placa supragingival de los dientes y la consiguiente selección y colonización de algunas bacterias acidófilas cariogénicas (Loesche et al., 1975; Simón-Soro y Mira, 2015).

Tradicionalmente se ha definido al género *Streptococcus* como el agente etiológico de la caries dental, especialmente a *Streptococcus mutans* y *sobrinus*, junto a *Lactobacilli* (van Houte, 1994). Sin embargo, gracias a la metagenómica, ahora sabemos que *Streptococcus mutans* solo supone un 0,1% de la comunidad bacteriana de la placa dental y es el causante de solo un 0,7-1,6% de las lesiones por caries (Gross et al., 2012; Simón-Soro et al., 2013). Las lesiones por caries se dan por la acción de un gran rango de especies bacterianas como las pertenecientes a *Streptococcales*, *Latobacillus*, *Actinobacteria*, donde destacan las *Bifidobacterium*, *Veillonella*, *Prevotella*, *Atopobium* o *Corynebacterium* (Aas et al., 2008; Fejerskov, 2004; Simón-Soro y Mira, 2015) y ocasionalmente levaduras como *Candida* (Jean et

al., 2018). Esto convierte a la caries en una enfermedad polimicrobiana, donde una comunidad de más de un centenar de bacterias y algunas levaduras juegan un papel fundamental en su desarrollo.

Individuos que nunca han sufrido caries carecen de *Streptococcus mutans*, pero presentan otras especies como *Streptococcus parasanguinis*, *Abiotrophia defectiva*, *Streptococcus mitis*, *Streptococcus oralis* y *Streptococcus sanguinis* (Belda-Ferre et al., 2012), cuya colonización se ha sugerido que otorga protección frente al desarrollo de la caries (Aas et al., 2008). El mecanismo anticaries de estas bacterias no se conoce en todos los casos, pero una bacteria del género *Streptococcus* recientemente aislada de placa dental de individuos que nunca han desarrollado caries (Belda-Ferre et al., 2012), *Streptococcus dentisani*, es capaz de frenar el crecimiento de bacterias eminentemente cariogénicas como *Streptococcus mutans*, *Streptococcus sobrinus* o *Prevotella intermedia* al secretar al medio bacteriocinas específicas y frenar la caída drástica del pH mediante la producción de amonio desde arginina (Camelo-Castillo et al., 2014; López-López et al., 2017). La viabilidad y eficacia de la administración como probiótico de esta bacteria está siendo actualmente evaluada en ensayos clínicos (https://clinicaltrials.gov/ct2/show/NCT03522363).

3.2.2. Periodontitis

La periodontitis es una enfermedad polimicrobiana que causa la destrucción de los ligamentos y el hueso alveolar que soporta el diente, generando inflamación, dolor y en última instancia, la pérdida de las piezas dentales (Kinane et al., 2017). La periodontitis es una enfermedad que afecta a un 20-50% de la población adulta (Eke et al., 2015; Nazir, 2017). Tradicionalmente se ha determinado que son tres las especies bacterianas causantes de la periodontitis, conocidas como el "complejo rojo": *Porphyromonas gingivalis*, *Tannerella forsythensis* y *Treponema denticola* (Socransky y Haffajee, 2005). De nuevo, la metagenómica ha ampliado el número de bacterias asociadas con esta enfermedad, entre las que se incluye *Aggregatibacter actinomycetemcomitans* (Hajishengallis y Lamont, 2012) y bacterias de los géneros *Anaeroglobus*, *Bulleidia*, *Desulfobulbus*, *Filifactor*, *Mogibacterium*, *Phocaeicola*, *Schwartzia*, *TM7* o *Streptococcus* (Camelo-Castillo et al., 2015). Aunque la causa de la disbiosis en estas comunidades bacterianas se desconoce por el momento, el aumento de la actividad lítica de bacteriófagos de la familia *Myoviridae* en el espacio subgingival podría desempeñar un papel importante en la enfermedad (Ly et al., 2014). También se ha sugerido la implicación de virus humanos como *HHV-4* y citomegalovirus (Jakovljevic et al., 2015; Zhu et al., 2015).

3.2.3. Estomatitis aftosa recurrente

La estomatitis aftosa recurrente (afta) es la enfermedad más común de la mucosa oral y la lengua. Su forma menor es la causante del 70-85% de los casos de aftas, presentándose como una o varias úlceras circulares con un margen eritematoso y un centro amarillo grisáceo que persisten durante 7-14 días

(Tarakji et al., 2015). Se desconoce su etiología, pero se han propuesto varios factores como posibles causantes, incluyendo factores locales (como un traumatismo), factores nutricionales (como la deficiencia de folato y el complejo de la vitamina B), factores inmunológicos, hormonales, estrés, alergias alimentarias y factores microbianos (Porter et al., 1998). La comunidad de bacterias de las aftas presenta una disminución de *Firmicutes* y un aumento de *Proteobacteria* (Hijazi et al., 2015). Aunque no existe consenso en la comunidad científica, también se han asociado con esta enfermedad algunos virus eucarióticos como *HSV1*, varicella-zoster, citomegalovirus o adenovirus (Natah et al., 2004; Pedersen y Hornsleth, 1993).

3.2.4. Cáncer bucal

Se conocen varios microorganismos de la cavidad bucal que pueden participar en el desarrollo de cáncer en la boca. Los papilomavirus de tipo 16 y 18 son los principales responsables de cáncer en cabeza y cuello (Kreimer et al., 2005). También se ha visto que en cáncer oral hay un incremento de ciertas bacterias como *Capnocytophaga gingivalis, Prevotella melaninogenica* y *Streptococcus mitis* (Mager et al., 2005), y más recientemente se han asociado estados de disbiosis en los que se produce un aumento de bacterias de los géneros *Fusobacterium, Dialister, Peptostreptococcus, Filifactor, Peptococcus, Catonella* y *Parvimonas* (Börnigen et al., 2017; Zhao et al., 2017).

4. Uso terapeútico de los bacteriófagos o sus productos génicos

4.1. Fagoterapia

El uso clínico de mezclas de bacteriófagos para combatir enfermedades bacterianas se conoce como fagoterapia o terapia fágica. Debido a la continua adquisición de resistencias a antibióticos por parte de una gran variedad de bacterias patogénicas, la fagoterapia se presenta como una prometedora alternativa a los antibióticos. Sus orígenes se remontan al descubrimiento de los bacteriófagos, cuando uno de sus descubridores, Felix d'Herelle, propuso el uso de un bacteriófago para el tratamiento eficiente de la disentería, en 1919 (d'Herelle, 1917; Summers, 1999). Unos años más tarde se utilizaron bacteriófagos para tratar una infección de la piel producida por *Staphylococcus* (Bruynoghe et al., 1921). El descubrimiento de la alta eficacia de los antibióticos hizo que, en la década de los 40, las investigaciones basadas en fagoterapia pasaran a un segundo plano. Sin embargo, en Europa del Este, la terapia fágica se siguió utilizando con éxito para tratamientos frente a bacterias de los géneros *Staphylococcus, Pseudomonas, Klebsiella* y *Escherichia*, entre otras bacterias (Babalova et al., 1968; Bogovazova et al., 1992; Carlton, 1999; Perepanova et al., 1995; Slopek et al., 1984, 1983; Weber-Dąbrowska et al., 2001).

Debido a la aparición de bacterias con multirresistencias a casi todos los antibióticos disponibles, se han financiado numerosos proyectos de investigación en todo el mundo para explorar la eficacia de la fagoterapia en la clínica (Furfaro et al., 2018). La validez de la fagoterapia se ha demostrado en modelos

animales de infección con *Staphylococcus aureus*, incluyendo cepas resistentes a meticilinas (*MRSA*) (Grunenwald et al., 2018), *Pseudomonas aeruginosa* (Forti et al., 2018), *Klebsiella pneumoniae* (Chadha et al., 2017) y *Acinetobacter baumannii* (Cha et al., 2018). También se ha publicado algunos estudios con resultados prometedores en personas con infecciones de estas bacterias (Chan et al., 2018; Międzybrodzki et al., 2012; Schooley et al., 2017; Wright et al., 2009).

En la actualidad, la fagoterapia se emplea en la industria alimentaria para el biocontrol de patógenos bacterianos. Por ejemplo, la *Food and Drugs Administration* (*FDA*) en Estados Unidos tiene aprobado el uso de bacteriófagos contra bacterias como *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli O157:H7* o *Mycobacterium tuberculosis* bajo la clasificación de "considerados como seguros" en las cadenas de producción y envasado de los alimentos (Coffey et al., 2010; Endersen et al., 2014; Monk et al., 2010; Nannapaneni y Soni, 2015). En clínica, a día de hoy, no existen productos o tratamientos basados en terapia fágica licenciados para su uso en enfermedades humanas en la Unión Europea o Estados Unidos. Sin embargo, hay registrados al menos 12 ensayos clínicos basados en el uso de bacteriófagos para el tratamiento de diferentes infecciones, en su mayoría desarrollados en Estados Unidos y Francia (https://clinicaltrials.gov/) (Furfaro et al., 2018), y estamos asistiendo a la aparición de decenas de empresas biotecnológicas relacionadas con la fagoterapia (http://www.companies.phage.org/). Algunos de los aspectos en los que se está trabajando actualmente en este campo son la modificación por ingeniería genética de los bacteriófagos para ampliar su rango de hospedador, redirigiendo, por ejemplo, bacteriófagos que infectan *Escherichia coli* hacia bacterias patogénicas como *Yersinia pseudotuberculosis* o *Klebsiella 390* (Ando et al., 2015), el uso de mezclas complejas de bacteriófagos para superar la limitación del estrecho rango de hospedador de muchos de estos bacteriófagos, o el desarrollo de un marco de regulación adecuado para este tipo de terapias (Pirnay et al., 2018) como el propuesto en el proyecto europeo *Phagoburn* (Rose et al., 2014).

La fagoterapia presenta una serie de ventajas frente al uso de antibióticos como son: una mayor especificidad, ya que muchos bacteriófagos tienen un rango de hospedador estrecho; la auto-amplificación de las partículas virales; o la baja toxicidad en humanos. Entre sus desventajas podría estar el desarrollo de sistemas de defensa bacterianos específicos frente a estos bacteriófagos (Carlton, 1999), a través de los *CRISPRs*.

4.2. Tratamientos con proteínas bacteriolíticas codificadas en el genoma de bacteriófagos

Las lisinas y las holinas son dos tipos de proteínas codificadas en el genoma de los bacteriófagos que trabajan en conjunción para lograr la lisis bacteriana y permitir la liberación de la progenie viral. Las lisinas, al igual que las lisozimas eucarióticas, rompen los enlaces que forma el peptidoglicano en las paredes bacterianas provocando el lisado de estas células por osmólisis. Estas enzimas pueden hidrolizar los enlaces entre los monosacáridos que forman el peptidoglicano (N-acetilglucosaminidasas y N-

acetilmuramidasas), el enlace con la L-alanina de los péptidos que unen cadenas de peptidoglicanos (N-acetilmuramoil-L-alanina amidasa) o los enlaces entre los aminoácidos de estos péptidos (endopeptidasas). Las holinas, por otro lado, son proteínas transmembrana que forman poros en la membrana de las bacterias permitiendo el acceso de las lisinas a la pared celular (Young, 1992). Además de estas proteínas implicadas en la lisis de las bacterias en las etapas finales del ciclo vital de los bacteriófagos, éstos presentan otras actividades enzimáticas asociadas a componentes estructurales de la cola de los viriones que producen una hidrólisis local de peptidoglicano durante la entrada de la partícula viral (Kanamaru et al., 2005).

El uso potencial en clínica de las lisinas de los bacteriófagos como alternativa a los antibióticos, ha llevado a acuñar el término de enzibióticos (Nelson et al., 2001), y a varios grupos de investigación a explorar su actividad y especificidad contra determinadas bacterias (Schmelcher et al., 2012). Algunos de los primeros trabajos publicados describieron la capacidad de las lisinas de los bacteriófagos pneumococales Cp-1 y Dp-1, A511 y C1 para hidrolizar la pared celular de *Streptococcus pneumoniae* (García et al., 1987; Loeffler, 2001; Nelson et al., 2001; Sheehan et al., 1997) y *Streptococcus pyogenes* (Nelson et al., 2001), respectivamente. Estudios más recientes han demostrado la eficacia de alguna de estas lisinas de bacteriófagos en modelos animales (Grandgirard et al., 2008; Jado et al., 2003), mejorando el efecto de los antibióticos contra *Streptococcus pneumoniae* (Corsini et al., 2018; Rodriguez-Cerrato et al., 2007) o contra *Clostridium difficile* (Wang et al., 2015), incluso reduciendo la formación de *biofilms* bacterianos (Domenech et al., 2011; Szafrański et al., 2017). También se han obtenido resultados muy prometedores contra *Bacillus cereus* empleando lisinas del bacteriófago γ (Schuch et al., 2002), o contra *Staphylococcus aureus* (incluyendo *MRSA*) con la endolisina del bacteriófago *ΦMR11* (Rashel et al., 2007) o la lisina del bacteriófago *ΦH5* que lo elimina hasta niveles no detectables en leche pasteurizada (Gutiérrez et al., 2018; Obeso et al., 2008). Uno de los principales problemas del uso de lisinas es la falta de eficacia frente a bacterias gram-negativas, ya que su membrana externa de lipopolisacáridos impide su acceso a la pared de peptidoglicanos. El único caso documentado de acción sobre bacterias gram-negativas es el de la lisinas que actúan sobre *Acinetobacter baumannii*, posiblemente debido a su dominio C-terminal cargado positivamente que puede ayudar a atravesar la membrana externa (Lood et al., 2015). En este sentido, se han desarrollado lisinas recombinantes contra *Pseudomonas aeruginosa* que tienen fusionadas proteínas de desestabilización de lipopolisacáridos, denominadas artilisinas (Briers et al., 2014). Actualmente existen dos ensayos clínicos en humanos utilizando diferentes lisinas contra *Staphylococcus aureus* (Czaplewski et al., 2016). Uno de ellos emplea lisina CF-301 para el tratamiento de *Staphylococcus aureus* en endocarditis, el cual ha presentado resultados positivos en fase 2 y va a comenzar la fase 3 (Grunenwald et al., 2018; Indiani et al., 2019). Staphefekt XDR.300 es una solución antiséptica ya disponible en el mercado y efectiva contra *MRSA* en infecciones de piel humana y que deriva de la endolisina del bacteriófago *SAP-1* de *Staphylococcus aureus* (https://www.staphefekt.com/en/).

OBJETIVOS

La boca constituye el principal portal de entrada de patógenos y presenta una microbiota diversa, compleja y sometida a numerosos factores abióticos cambiantes. Pese a que cambios en la microbiota bucal se han asociado con algunas patologías frecuentes de este ecosistema como la caries, periodontitis y estomatitis aftosa recurrente, falta mucho para comprender qué desencadena dichas alteraciones. Pese a que los virus presentan múltiples mecanismos de regulación de las comunidades microbianas, los viromas de la boca humana han sido pobremente estudiados. La presente tesis doctoral trata de conocer mejor estas comunidades de virus en individuos con diferentes estados de salud y enfermedad, empleando diversas aproximaciones metagenómicas y herramientas bioinformáticas. Los objetivos específicos que nos planteamos fueron:

1.- Evaluar los sesgos en la composición de las comunidades virales que se introducen durante los pasos de enriquecimiento de partículas virales y amplificación aleatoria de genomas en estudios metagenómicos de virus de ADN.

2.- Caracterizar mediante técnicas metagenómicas las comunidades de virus de ADN de la mucosa bucal y la placa dental humana.

3.- Estudiar posibles cambios en las comunidades virales de la placa dental y de la mucosa oral en enfermedades frecuentes de la cavidad bucal como caries y estomatitis aftosa recurrente.

4.- Predecir mediante diversas herramientas bioinformáticas el hospedador bacteriano de los principales grupos de bacteriófagos de la cavidad bucal.

5.- Identificar componentes clave del viroma bucal humano, como bacteriófagos o enzibióticos, que pudieran ser utilizados para restablecer ecosistemas bucales saludables.

MATERIALES Y MÉTODOS

1. Trabajo en poyata (*wet lab*)

1.1. Materiales

1.1.1. Colección de virus para la construcción de una comunidad sintética

La comunidad sintética se formó con siete virus de ADN de diferentes características genéticas y estructurales (**Fig. 2**): el virus Vaccinia cepa Western Reserve (WR) se caracteriza por ser un virus grande (250x360nm) con envuelta y fue cedido por el doctor Antonio Alcamí (Centro de Biología Molecular Severo Ochoa: CBMSO) tras una purificación en colchón de sacarosa; los bacteriófagos λ, cedido por Dionisio Ureña (CBMSO) y Φ29, proporcionado por la doctora Margarita Salas (CBMSO), presentan algunas de las características más comunes de los bacteriófagos mayoritarios, ya que contienen un ADN de banda doble y morfología de cabeza y cola. Estos virus se purificaron dos veces en gradientes de CsCl; el bacteriófago con estructura filamentosa M13 fue proporcionado también por la doctora Margarita Salas desde un gradiente de CsCl; el Circovirus porcino 2a (PCV2a), procedente de un sobrenadante de cultivo clarificado por centrifugación a baja velocidad, fue cedido por el doctor Joaquim Segalés (Centre de Recerca en Sanitat Animal. Universidad Autónoma de Barcelona). Estos dos últimos virus se caracterizan por presentar un genoma de ADN pequeño y circular; el Parvovirus diminuto de ratón cepa p (MVMp), proporcionado por el doctor José María Almendral (CBMSO), se caracteriza por tener un tamaño pequeño y un ADN lineal de cadena sencilla. Este virus se purificó mediante un gradiente de sacarosa y un gradiente isopícnico de CsCl; el adenovirus humano 5 (AdenoV) es un virus sin envuelta y con un genoma de ADN lineal bicatenario. Este virus fue cedido por la doctora Carmen San Martín (Centro Nacional de Biotecnología) y procedía de una doble purificación en gradientes de CsCl.

1.1.2. Muestras de la cavidad bucal humana

El protocolo de obtención de muestras de mucosa bucal, placa dental y saliva, y su posterior uso, fue explicado por personal cualificado (odontólogos en el caso de placa dental y estomatólogos en el caso de saliva y mucosa oral) a todas las personas que voluntariamente se presentaron al estudio. Las muestras fueron recogidas tras la firma de un consentimiento informado que contó con la aprobación de los comités de ética de la Universidad del País Vasco (UPV) y del Centro Superior de Investigación en Salud Pública (CSISP)-Fundación FISABIO de la Comunidad Valenciana. También se obtuvo un informe favorable del Comité de Ética de la Investigación de la Universidad Autónoma de Madrid (CEI-31-792).

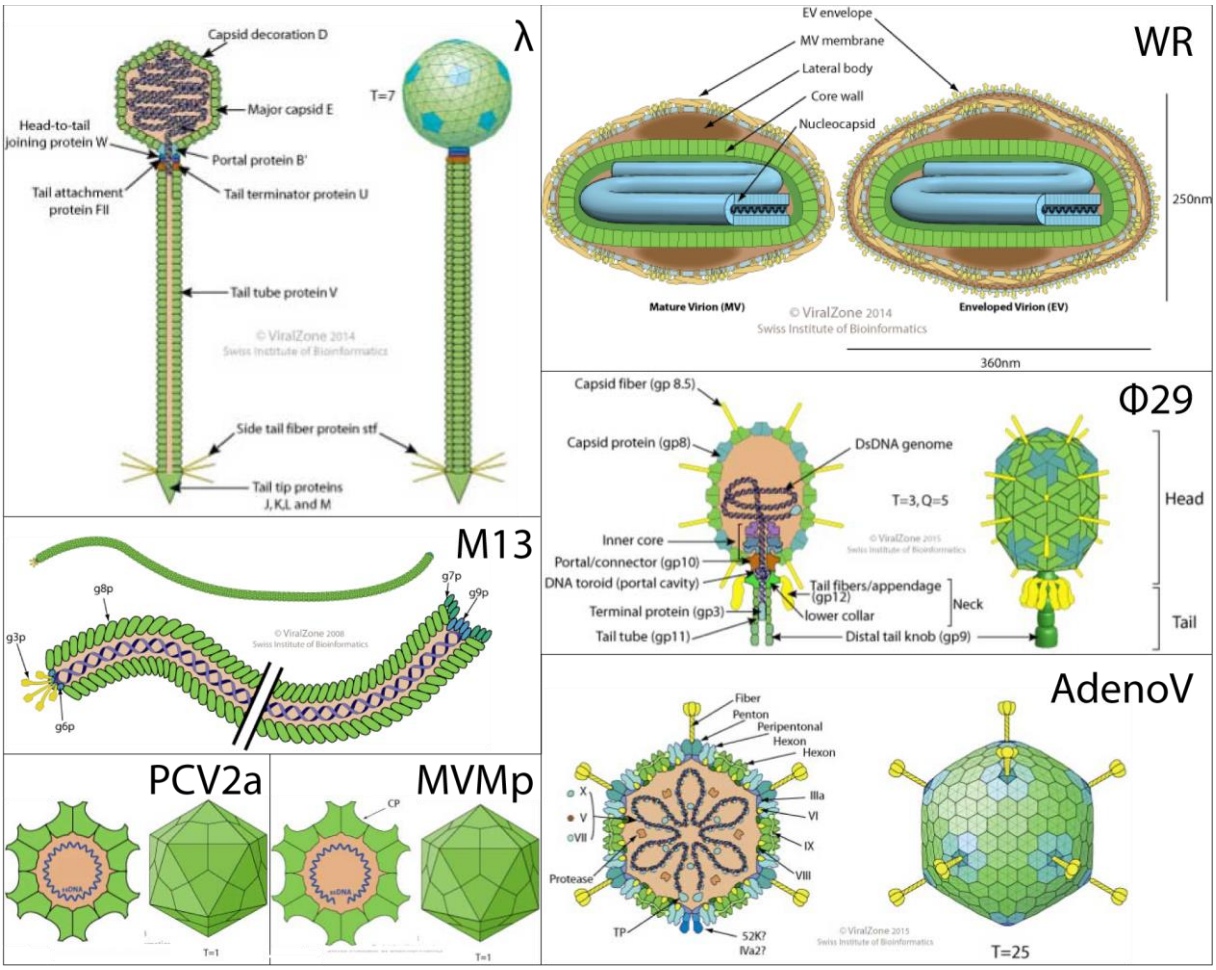

Figura 2. Representación gráfica de los siete virus que conforman las comunidades sintéticas. Las imágenes proceden de la web *ViralZone* (www.expasy.org/viralzone, SIB Swiss Institute of Bioinformatics (Hulo et al., 2011)**).**

1.1.2.1. Muestras de mucosa oral

Las muestras de mucosa fueron extraídas de individuos sin relación de parentesco por los doctores Asier Eguía y José Manuel Aguirre Urizar (Departamento de Estomatología II de la UPV). Estas muestras procedían de pacientes que padecían Estomatitis Aftosa Recurrente (EAR) y de individuos sanos. Debido a la etiología desconocida de la enfermedad y ante la posibilidad de que existan causas distintas, se recogieron muestras exclusivamente de pacientes con aftas simples (dos a cuatro episodios anuales) y menores (5-10 mm de tamaño) en epitelio no queratinizado. Se descartaron aquellos pacientes o individuos sanos con indicación o sospecha de:

- Enfermedades autoinmunes relacionadas con aftas complejas: enfermedad de Behcet, lupus eritematoso sistémico, síndrome de Reiter, enfermedad de Crohn, colitis ulcerosa y enfermedad celíaca.

- Alteraciones hematológicas: síndromes mielodisplásicos, neutropenia cíclica, síndrome *PFAPA* (fiebre periódica que se asocia a faringitis, adenopatías y aftas orales recurrentes) y síndrome de Sweet.
- Inmunodeficiencias (como la infección por VIH).
- Lesiones de origen infeccioso: virales (como herpes o virus boca-mano-pie) o bacterianas.
- Enfermedades mucocutáneas: liquen plano, pénfigo y penfigoide.
- Neoplasias.
- Traumatismo bucal (como el uso de prótesis, quemaduras o mordeduras).
- Periodontitis.
- Gingivitis severa.
- Candidiasis.
- Procesos febriles en el último mes.
- Procesos infecciosos en vías respiratorias o sistema digestivo en el último mes.
- Toma de antibióticos en el último mes.

Para la recolección de las muestras de mucosa oral de individuos sanos se frotó un hisopo (*Catch-All*™ *Sample Collection Swabs*, Epicentre) por la zona vestibular izquierda, derecha e inferior de la boca durante unos 10 sg, con cuidado de no tocar piezas dentales. En el caso de pacientes con EAR, se tomaron muestras por contacto directo y repetido del hisopo con la úlcera o afta, evitando en la medida de lo posible el sangrado. Cada hisopo se sumergió en 800 µl de tampón SM 1x (50 mM Tris pH 7,5, 100 mM NaCl y 10 mM $MgSO_4$) suplementado con sacarosa al 20% y se agitó presionando contra las paredes del tubo para asegurar que la muestra se transfiriera al fluido durante 30 sg. El tampón y la sacarosa fueron previamente filtrados con filtros de jeringa con un tamaño nominal de poro de 0,22µm (*PVDF syringe filters*, Millipore). Las muestras se preservaron congeladas a -20ºC hasta su llegada al CBMSO y a -80ºC hasta su utilización. La sacarosa minimiza el impacto de la descongelación sobre las cápsidas virales como pudimos comprobar en estudios previos de cuantificación de genomas virales protegidos de la acción de nucleasas en preparados congelados con y sin sacarosa (datos no mostrados).

1.1.2.2. Muestras de placa supragingival

Las muestras de placa dental de pacientes sin relación de parentesco fueron extraídas por los doctores Alex Mira y Áurea Simón (CSISP-FISABIO). Estas muestras procedían de pacientes que sufrían caries activa (individuos enfermos) y de individuos que no habían sufrido nunca una caries registrada clínicamente (individuos sanos). Las muestras se recogieron utilizando una cureta de metal para extraer la placa supragingival de varias piezas hasta conseguir masa suficiente. En las muestras de individuos con caries se evitó tocar directamente la zona afectada por caries. La cureta se sumergió en 400 µl de tampón SM 1x con sacarosa al 20% previamente filtrados y se agitó con fuerza durante 5 sg para facilitar

la dispersión del material biológico en el tampón. Las muestras se preservaron congeladas a -20ºC hasta su llegada al CBMSO y a -80ºC hasta su utilización.

Los criterios de exclusión de estas muestras fueron:

- Haber tomado antibióticos durante el último mes.
- Padecer enfermedades orales como periodontitis.
- No tener al menos 28 piezas dentales naturales.
- Haberse cepillado los dientes en las 24h previas al muestreo.

1.1.2.3. Muestras de saliva

Se recogieron muestras de saliva de los mismos individuos que proporcionaron muestras de mucosa oral. Se depositaron aproximadamente 3 ml de saliva no estimulada de cada individuo en tubos cónicos y estériles de polipropileno de 15 ml (Falcon™) y se diluyeron en tres volúmenes de tampón SM 1x filtrados con el fin de evitar la colmatación de los filtros y la inversión de fases durante la fenolización posterior. Las muestras de saliva se recogieron sin estimulación porque ésta induce sesgos en la composición de la comunidad microbiana (Gomar-Vercher et al., 2018) A continuación, se realizaron tres ciclos de 15 sg de vortex y hielo para evitar el sobrecalentamiento que podría afectar a la estabilidad de las cápsidas virales, respectivamente. Las muestras se preservaron congeladas a -20ºC hasta su llegada al CBMSO y a -80ºC hasta su utilización.

Adicionalmente, se prepararon dos mezclas a partir de salivas de siete y nueve individuos aparentemente sanos con el objetivo de tener suficiente material como para abordar directamente la preparación de librerías de secuenciación masiva sin amplificación al azar. Seis muestras de saliva de ambas mezclas procedían de las mismas personas y fueron tomadas con un intervalo de una semana.

1.2. Métodos

1.2.1. Cuantificación de genomas virales mediante *PCR* cuantitativa y preparación de comunidades sintéticas

1.2.1.1. Preparación de estándares para *PCR* cuantitativa

La determinación del número de genomas virales contenidos en partículas intactas de siete virus de ADN (WR, λ, Φ29, M13, MVMp, AdenoV y PCV2a) se hizo mediante *PCR* cuantitativa utilizando estándares de concentración conocida. Los estándares usados consistieron en una serie de plásmidos que contenían alguna región clonada de cada genoma viral. Estos plásmidos fueron cedidos, en la mayoría de los casos, por los mismos grupos que proporcionaron los preparados virales. El estándar de WR consistió en un plásmido pcDNA3-V5His (Invitrogen) con el gen *B18R* de Vaccinia WR clonado entre las dianas KpnI y XbaI; el estándar de AdenoV era un plásmido pUC19 que tiene clonado el gen *pIII* de Adenovirus

humano 5; el estándar del parvovirus MVMp era un plásmido pSVtk que tiene clonado la región de los genes estructurales de MVMp; el estándar de PCV2a tiene clonado el genoma completo del virus en la diana SacII del plásmido pCR2.1TOPO (Invitrogen); los estándares de Φ29 y λ se construyeron en nuestro grupo por Ana Rodríguez Galet a partir de los fragmentos de restricción Hind III de 1.150 y 564 nucleótidos respectivamente, que se clonaron en ambos casos en el mismo sitio de restricción del plásmido pcDNA_3 (Invitrogen). El estándar del bacteriófago M13, a diferencia de todos los anteriores, consistió en un producto de *PCR* de 645 nucleótidos obtenido usando los oligonucleótidos M13 primer_F HindIII y M13 primer_R EcoRV (**Anexos - Oligonucleótidos**). Todos los estándares consistentes en plásmidos fueron linearizados con enzimas de restricción de sitio único (**Fig. 3**).

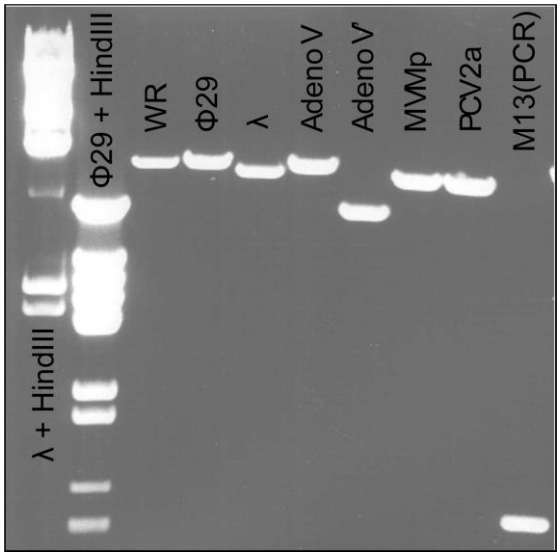

Figura 3. Gel de agarosa con los estándares linearizados usados para la determinación del número de genomas protegidos por cápsidas o envueltas intactas de siete virus de ADN. El plásmido con el fragmento del genoma de MVMp se linearizó con el enzima de restricción BamHI, el plásmido con el fragmento de PCV2a se linearizó con el enzima XhoI y los plásmidos con los fragmentos de genoma de WR, AdenoV, λ y Φ29 se linearizaron con XbaI. Para nuestro estudio utilizamos AdenoV y no AdenoV'. Los marcadores moleculares fueron los genomas de los bacteriófagos λ y Φ29 digeridos con HindIII.

Todos los estándares fueron cuantificados mediante *NanoDrop™ 1000* (Thermo Scientific) y *Quant-iT™ PicoGreen® dsDNA Assay*. Los oligonucleótidos empleados en las *PCR* cuantitativas se diseñaron con el programa *Primer3Plus* (Untergasser et al., 2012) con los siguientes parámetros: 18-22 nucleótidos, temperatura de alineamiento de 58-62ºC, homopolímeros de tres nucleótidos máximo y tamaños de amplificación de 80-150 pb (**Anexos - Oligonucleótidos**).

1.2.1.2. Cuantificación por *PCR* cuantitativa

Para la cuantificación de los genomas virales presentes en partículas virales intactas de cada preparado de virus se hizo un tratamiento con una mezcla de nucleasas previo a la extracción de los genomas virales con proteinasa K, *SDS* y fenol/cloroformo (ver descripción pormenorizada de estos tratamientos en la siguiente sección). La amplificación por *PCR* cuantitativa de los genomas purificados y de los estándares se realizó en placas de 384 pocillos acopladas a dos termocicladores diferentes: (i) *ABI PRISM 7900HT SDS* usando el fluoróforo *QuantiTect SYBR1 Green PCR Kit (Qiagen, Courtaboeuf, France)*, y siguiendo las instrucciones del fabricante. El volumen final de reacción fue de 10 µl y el

protocolo de temperaturas consistió en una desnaturalización inicial a 95ºC 15 min, seguido por 40 ciclos de desnaturalización a 94ºC 15 sg, alineamiento a 60ºC 30 sg y elongación a 72ºC 30 sg; y (ii) *CFX384 Touch thermocycler* (BioRad) usando el fluoróforo *SsoFast EvaGreen Supermix* (BioRad), y siguiendo las instrucciones del fabricante. El volumen final de reacción fue de 10 µl y el protocolo de temperaturas consistió en una desnaturalización inicial a 95ºC 30 sg, seguido por 40 ciclos de desnaturalización a 95ºC 5 sg y alineamiento-elongación a 60ºC 5 sg. Para cada punto se hicieron triplicados técnicos y para cada pareja de oligos se añadió un control negativo con agua en lugar de ADN. La evaluación de la especificidad del proceso se realizó mediante la monitorización de un ciclo de desnaturalización a 95ºC durante 5 sg sobre el producto final amplificado, en ambos casos.

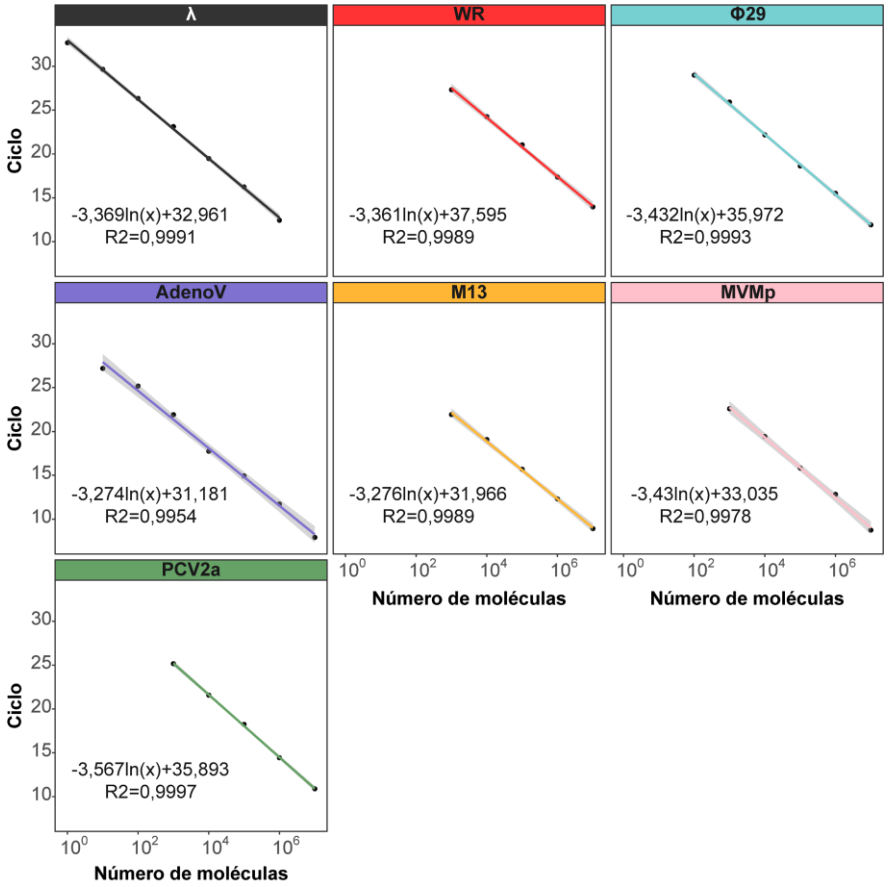

Figura 4. *PCR* cuantitativa desde cantidades conocidas de cada uno de los estándares de siete virus de ADN. Se representa la relación lineal entre el logaritmo del número de moléculas empleadas y el número de ciclos necesarios para detectar un determinado nivel de amplificación (Ct: *cycle threshold*).

Finalmente, los resultados fueron analizados en el software *SDS 2.4* (Applied Biosystem). La cuantificación absoluta de los genomas virales resultantes se realizó mediante interpolación en la recta obtenida con las diluciones seriadas de los estándar (**Fig. 4**). Las *PCRs* de los estándares tuvieron una eficiencia en torno al 91,21-105,82% en rangos dinámicos de 5-7 órdenes logarítmicas, y con valores de R^2 por encima de 0,996. La presencia de contaminaciones en las *PCR* cuantitativas se determinó

mediante curvas de desnaturalización (*melting*). Los análisis se representaron de forma gráfica utilizando el paquete *ggplot2* (Wickham, 2016) de *R v3.2.3* (R Core Team (2017), 2017).

1.2.2. Preparación de comunidades virales sintéticas

Para la preparación de comunidades virales sintéticas se hicieron mezclas balanceadas con igual cantidad de material genético procedente de los siete virus de ADN, preparadas en tampón SM 1x. Para ello, se determinó la concentración de genomas virales en partículas intactas por *PCR* cuantitativa (ver apartado anterior) y se mezclaron cantidades equivalentes de ADN hasta un total de 20 ng de ADN viral en la primera comunidad sintética y 120 ng en la segunda. En esta última solo se pudo incluir un número de moléculas de PCV2a equivalente a 0,88 ng por agotamiento del único preparado disponible.

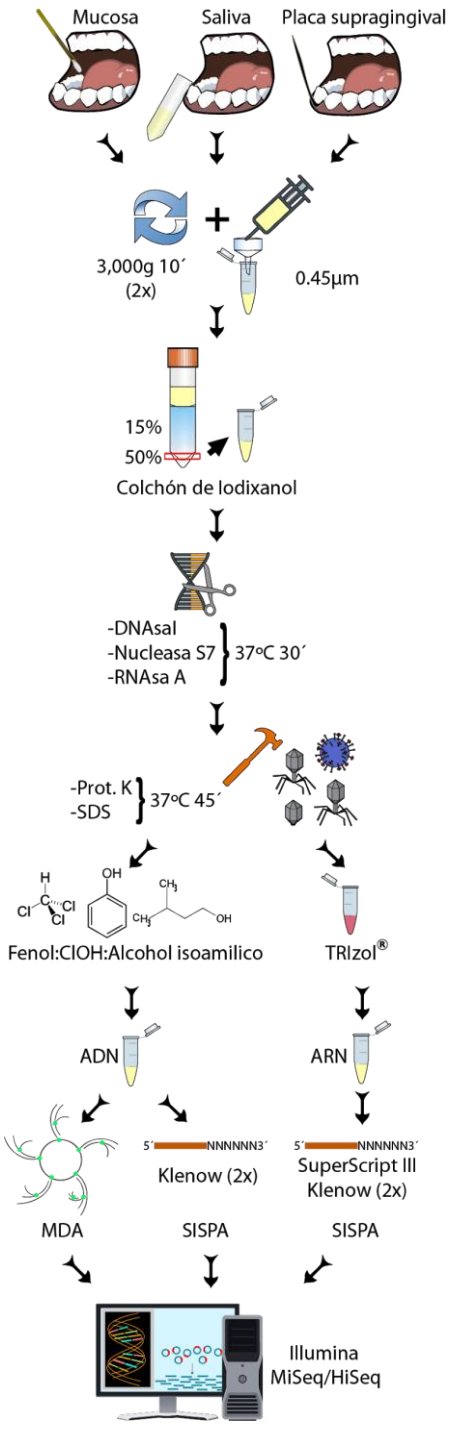

Figura 5. Esquema de los pasos seguidos para la obtención de viromas orales. Se representa de forma secuencial la extracción de muestras, la purificación de genomas virales, la amplificación al azar del material genético y la secuenciación masiva. La figura es una modificación de publicada en el capítulo del libro *The Human Virome* (Parras-Moltó y López-Bueno, 2018).

1.2.3. Enriquecimiento de partículas virales y purificación de genomas virales

Las muestras de la cavidad bucal y las comunidades sintéticas se procesaron siguiendo un protocolo sencillo de enriquecimiento de partículas virales y posterior extracción del material genético protegido por cápsidas o envueltas virales. El protocolo completo se encuentra publicado con más detalle en el capítulo del libro *The Human Virome* (Parras-Moltó y López-Bueno, 2018) (**Fig. 5**).

1.2.3.1. Centrifugación y filtración

A cada una de las muestras procedentes de placa dental y mucosa se le añadió un volumen de tampón SM 1x con el objetivo de diluir la concentración de sacarosa y evitar problemas durante los pasos de extracción con fenol y cloroformo, ya que habíamos previamente comprobado que la fenolización de un tampón con 20% de sacarosa provocaba la inversión de las fases. A continuación, las muestras se

agitaron mediante tres ciclos de vortex durante 20 sg e incubación a 4ºC para evitar el sobrecalentamiento. Posteriormente, las muestras se centrifugaron a 3.000 g durante 10 min y el sobrenadante se transfirió a tubos estériles de 1,5 ml de baja unión a proteínas, cierre seguro y libre de ADN humano y nucleasas (Eppendorf). Este paso se repitió con el objetivo de tener una muestra lo más libre posible de bacterias. La filtración se realizó usando filtros de 0,45µm (*PVDF syringe filters*, Millipore) para permitir el paso de algunos virus de gran tamaño, siendo conscientes de que estos filtros no eliminan la totalidad de las bacterias de pequeño tamaño.

Para evaluar la eficiencia de la centrifugación y filtración en la eliminación de bacterias, cultivos puros de *Escherichia coli*, *Staphylococcus aureus* y *Roseobacter litoralis* se incubaron a 37ºC con agitación hasta alcanzar una densidad óptica de 0,6 a 600 nm. En los dos primeros casos se usó medio *LB* (Bacto-triptona 10 g/L, extracto de levaduras 5g/L y NaCl 10 g/L) y en el caso de *Roseobacter litoralis* se utilizó medio *Marine broth 2216*. Alícuotas de 1 ml se centrifugaron según el protocolo indicado anteriormente y/o se filtraron a través de filtros de 0,45µm (*PVDF syringe filters*, Millipore). El número de bacterias viables formadoras de colonias resultantes se cuantificó mediante plaqueo en placas petri con medio *LB* semisólido (agar al 1,8% p/v) para *Escherichia coli* y *Staphylococcus aureus*, y en medio *Marine brooth 2216* suplementado con agar al 1,5% p/v para *Roseobacter litoralis*.

1.2.3.2. Concentración en colchones de iodixanol

Las muestras de diluyeron en tampón SM 1x filtrado hasta un volumen de 12 ml en tubos de ultracentrifugación de 17 ml (*Thinwall, Ultra-Clear*, transparente. Beckman). Con la ayuda de una pipeta Pasteur de vidrio estéril se depositaron 3 ml de iodixanol al 15% p/v (OptiPrep™) en el fondo de cada tubo. A continuación, y también con una pipeta Pasteur, se depositaron en el fondo del tubo 0,5 ml de iodixanol al 50% p/v, evitando la mezcla de ambas fases. Los tubos se centrifugaron a 18.000 g durante 16 h a 4ºC en un rotor Beckman *SW41Ti*. Las partículas virales se recuperaron desde la interfase 15-50% de iodixanol. Las muestras de mezclas de saliva se procesaron en tubos de 38,5 ml de polialómero (Beckman) escalando las cantidades anteriormente mencionadas y usando el rotor TST 28.38 (Kontron).

1.2.3.3. Extracción de genomas virales de ADN

Los sedimentos de la ultracentrifugación en colchones de iodixanol se resuspendieron en tampón de nucleasa 1x (10 mM tris pH 7,5, 10 mM $MgCl_2$, 2 mM $CaCl_2$) y se trataron con el siguiente cóctel de nucleasas: 250 U/ml de DNAse I (Roche), 250 U/ml de nucleasa S7 (Roche) y 100 µg/ml de RNAse A (Roche) a 37ºC durante 30 min. La reacción se detuvo incubando con agitación 5 min con 20 mM de *EDTA* y 2 mM de *EGTA*. A continuación, se rompieron las cápsidas y envueltas virales con un tratamiento a 37ºC 45 min con proteinasa K a 200 µg/µl (Roche) y *SDS* al 0,5%. El ADN se extrajo con un volumen de fenol (Merck Millipore) equilibrado en 10 mM de Tris-HCl a pH 7,5, agitación con

vortex durante 30 sg y centrifugación en minifuga a 8.000 g 4 min a temperatura ambiente. Con cuidado de no tocar las paredes del tubo, se recogió la fase acuosa en tubos estériles de 1,5 ml. De igual forma, se repitió este paso utilizando un volumen de fenol-cloroformo (1:1) y posteriormente utilizando un volumen de cloroformo-alcohol isoamílico (24:1). El material genético purificado se precipitó añadiendo 0,1 volúmenes de NaAc 3M pH 5,6, 20 µg de glicógeno de mejillón (Roche), y 2,5 volúmenes de etanol absoluto (Merck Millipore). Tras precipitar durante al menos 8 h a -80ºC, las muestras se centrifugaron a 12.000 g 30 min y se lavaron usando etanol al 70% preparado con agua ultrapura libre de nucleasas (Ambion) dos veces. Los sedimentos resultantes se secaron a temperatura ambiente cinco minutos y se resuspendieron en 20 µl de Tris-HCl 15 M pH 7,5 preparado con agua ultrapura libre de nucleasas (Ambion).

1.2.3.4. Cuantificación de material genético extraído

El ADN purificado se cuantificó utilizando la tecnología *PicoGreen*® (Invitrogen) en el Parque Científico de Madrid (PCM; Campus de Cantoblanco) siguiendo las instrucciones del fabricante. La utilización de fluoróforos nos permite detectar cantidades de ADN por debajo de 1-5 ng/µl, que es el límite de detección de los métodos espectrofotométricos como *NanoDrop™ 1000* (Thermo Scientific).

1.2.4. Métodos de amplificación al azar de genomas virales

La aplicación de métodos de amplificación al azar nos permite obtener cantidades de ADN suficientes para afrontar la preparación de librerías de secuenciación masiva a partir de muestras donde, por limitaciones técnicas, no podemos obtener suficiente material.

1.2.4.1. Amplificación independiente de secuencia con un oligonucleótido único (*SISPA*)

Este método de amplificación inespecífica está basado en una *PCR* que utiliza oligonucleótidos pseudodegenerados con una región de secuencia conocida en su extremo 5´ y una región degenerada de 6-12 nucleótidos en su extremo 3´ (**Anexos - Oligonucleótidos**).

En una primera etapa se incorporó el oligonucleótido pseudodegenerado de forma inespecífica en los genomas mediante dos rondas consecutivas de extensión durante 1 h a 37ºC del ADN con el fragmento Klenow de la ADN polimerasa I (NEBiolabs) utilizando 1 µl de una mezcla de dNTPs (10 mM cada uno), 60 pmol de uno de los oligonucleótidos pseudodegenerados, 2 µl de tampón Klenow 10x, 3,5 unidades de Klenow y agua ultra-limpia para RT-PCR (Ambion) hasta un volumen final de 20 µl. Entre las dos rondas de extensión se incubó a 75ºC 10 min para desnaturalizar y se suplementó de nuevo con la misma cantidad de Klenow.

En la segunda etapa de la amplificación se empleó 1-10 µl del producto de la primera etapa para hacer una *PCR* específica con un nuevo oligonucleótido que contenía exclusivamente la secuencia 5´

conservada del oligonucleótido pseudodegenerado. En el caso de la amplificación de ADN viral desde muestras de mezclas de salivas, pero no en el caso de las comunidades sintéticas, se utilizaron cócteles de cinco oligonucleótidos con 1-4 nucleótidos degenerados adicionales en su extremo 5´ para facilitar el proceso de identificación de los distintos agregados de secuencias idénticas que ocurren durante la secuenciación masiva en equipos de Illumina® (Wu et al., 2015). La *PCR* contenía 10 µl de tampón 5x del enzima Q5 de alta fidelidad (NEBiolabs); 1 µl de $MgCl_2$ 25 mM; 1,5 µl de la mezcla de dNTPs 10 mM; y 4 µl del oligonucleótido universal correspondiente 0,8 mM, 1,4 unidades de polimerasa Q5 y agua ultra-limpia RT-PCR hasta un volumen final de 50 µl. El protocolo de temperaturas de esta *PCR* consistió en una desnaturalización inicial a 98ºC 2 min, seguido de 35 ciclos de desnaturalización a 98ºC 10 sg, alineamiento a 65ºC 30 sg y elongación a 72ºC 75 sg, seguido de una elongación final de 72ºC 150 sg. Se utilizó un termociclador *MJ Mini™ Thermal Cycler* (BioRad). El resultado de la amplificación se analizó por electroforesis en un gel de agarosa al 1% (p/v), cargando 5 µl del producto de *PCR*. Una vez analizado, las muestras se cargaron en un nuevo gel y las bandas con tamaños entre 0,8 y 1,5 kpb se cortaron usando cuchillas desechables para minimizar contaminaciones cruzadas. El ADN contenido en las bandas cortadas se extrajo con el kit *QIAquick PCR Purification Kit* (Qiagen) siguiendo el protocolo descrito por el fabricante. La muestra finalmente se eluyó en 30 µl de Tris-HCl 15 M pH 7,5 preparado con agua ultrapura libre de nucleasas (Ambion).

1.2.4.2. Amplificación por desplazamiento múltiple de banda (*MDA*)

Debido a sus características de alta procesividad y fidelidad de copia, la polimerasa del bacteriófago *Φ29* juega un papel fundamental en varios métodos de amplificación al azar. En esta tesis se han empleado dos protocolos de amplificación al azar que emplean esta enzima pero difieren en el método de cebado. En el primero se utilizan hexanucleótidos degenerados modificados para evitar su degradación por las actividades exonucleasas del enzima (kit *Illustra Ready-To-Go GenomiPhi™ V2* y *V3 ADN Amplification Kits*; GE HealthCare). En el segundo, los oligonucleótidos cebadores son suministrados por la primasa de *Thermos thermofilus* que se incluye junto a la polimerasa de *Φ29* en la propia reacción (*kit TruePrime™ WGA Kit*; Sygnis Biotech).

Para las amplificaciones que emplearon los kits de *Illustra Ready-To-Go GenomiPhi™*, se siguieron las recomendaciones del fabricante utilizando desnaturalización térmica y dejando progresar la amplificación isotérmica a 30ºC durante 2 h y 30 min. Las amplificaciones con los kits *TruePrime™* se llevaron a cabo siguiendo las recomendaciones del proveedor, que incluyen desnaturalización química y prolongando la elongación isotérmica a 30ºC durante 3 h. Los productos de las amplificaciones se precipitaron añadiendo 150 µl de etanol absoluto e incubando a temperatura ambiente 15 min. Luego se centrifugaron a 12.000 g 2 min y se lavaron dos veces con etanol al 70%. Esta precipitación reduce la cantidad de sales, nucleótidos y hexámeros aleatorios. Los sedimentos secos se resuspendieron en 50 µl de Tris-HCl 15 mM pH 7,5. Las amplificaciones resultantes se corrieron en un gel de agarosa al 0,7%

junto a un control negativo donde se sustituye la muestra por agua. Debido a que *MDA* acaba amplificando concatémeros de hexanucleótidos, consideramos como una amplificación positiva solo aquellas muestras que mostraban un producto de amplificación mayor que el observado en el control negativo. Por otro lado, para ayudar a la cuantificación del material genético amplificado y deshacer la estructura compleja a la que da lugar la amplificación por Φ29, los productos de amplificación se incubaron 5 min a 50ºC. La cuantificación se hizo usando *Picogreen*, que detecta específicamente ADN bicatenario, ya que los restos de hexanucleótidos alteran la cuantificación cuando se emplean técnicas espectrofotométricas.

1.2.5. *PCR* semicuantitativa para la estimación de la contaminación bacteriana

Para estimar el nivel de contaminación bacteriana que tenían productos amplificados al azar se analizó en cada caso la cantidad relativa del gen que codifica por el ARNr 16S mediante *PCR* con oligonucleótidos específicos (**Anexos - Oligonucleótidos**) (Frank et al., 2008). Para ello, se amplificaron diluciones seriadas de los productos de *MDA* siguiendo el protocolo de *PCR* descrito en el anterior trabajo: las muestras se desnaturalizaron a 95ºC 5 min y se sometieron a 25 ciclos de 1 min de desnaturalización a 95ºC, un paso de alineamiento a 51ºC 1 min y una extensión a 72ºC y 105 sg. Finalmente se realizó una extensión final a 72ºC durante 10 min. Los productos de amplificación de 1.496 pb se separaron en geles de agarosa al 1%. Como control positivo del nivel de contaminación bacteriana se usó una muestra disponible en el laboratorio y que presentaba un nivel de contaminación en torno al 10% de las secuencias según se había podido demostrar previamente por secuenciación masiva. La intensidad de las bandas de nuestros metagenomas en relación a la de este control fue utilizada como medida del grado aproximado de contaminación de nuestras muestras.

1.2.6. Purificación del ADN total asociado al sedimento bacteriano

Los sedimentos bacterianos provenientes del primer paso de centrifugación a baja velocidad para el enriquecimiento de virus se utilizaron para extraer ADN total y obtener microbiomas. Para ello, se resuspendieron en 500 µl de tampón de lisis 1x (*Tissue cell Lysis Solution*; MasterPure™ Kit EpicenterR), se les aplicó vortex 5 min a máxima velocidad y se incubaron a 37ºC durante 30 min con dos mg/ml de lisozima (Sigma-Aldrich). A continuación, se incubaron con 200 µg/ml de proteinasa K a 65ºC 15 min y, tras enfriar en hielo 3 min, se añadieron 300 µl de *MPC Protein Precipitation Reagent* (MasterPure™ kit; EpicenterR), se agitaron 10 min y centrifugaron a 14.000 rpms en minifuga 10 min a 4ºC. Finalmente, el ADN libre en el sobrenadante se precipitó con un volumen de isopropanol y se lavó dos veces con etanol al 70% antes de resuspenderlo en 30 µl de Tris-HCl 15 mM pH 7,5 preparado con agua ultrapura y libre de nucleasas (Ambion).

1.2.7. Secuenciación *shotgun* de genomas completos (*WGS*)

El ADN viral purificado y amplificado, y el ADN total proveniente de los sedimentos bacterianos, se fragmentó en tamaños de 700-900 pb mediante sonicación con *Biorruptor Plus* (Dioganode). A continuación, se prepararon librerías *NEBNext Ultra* (NEBiolabs) en el PCM, incluyendo cinco ciclos de amplificación por *PCR* para las muestras que se secuenciaron por *MiSeq* y ocho ciclos para las muestras de *HiSeq*. Todas las muestras secuenciadas se caracterizaron por poseer un tamaño de inserto grande, en torno 850-1000 pb, por lo que tras los ciclos de amplificación de las librerías se realizó una extensión final de 40 sg. Las moléculas de las librerías amplificadas con el tamaño deseado se extrajeron de geles de agarosa antes de la secuenciación. La secuenciación de los viromas se realizó en equipos *MiSeq* de Illumina® usando el kit *MiSeq Reagent Kit v3: 600 ciclos* que proporciona lecturas pareadas de 300 pb con una profundidad de 1,5-2 millones de lecturas por viroma, mientras que las muestras de ADN total (microbiomas) se secuenciaron en equipos *HiSeq2500* de Illumina® con el kit *HiSeq Rapid SBS Kit v2: 500 ciclos* logrando lecturas pareadas de 2x250pb con una profundidad de 15-20 millones de lecturas por carrera. Durante la preparación de las librerías, a cada muestra se le añadió una secuencia identificativa única (etiqueta o *barcode*) que permitió separar posteriormente las secuencias según su origen.

Para la secuenciación de microbiomas con la tecnología *Pacbio (RS II)* se utilizó el kit completo recomendado por la casa comercial (*Pacific Biociences*): *SMRTbell™ Template Prep Kit, DNA Polymerase Binding Kit, MagBead Kit y AMPure® PB beads*. Previo al paso de secuenciación, realizamos un paso para la selección de fragmentos grandes utilizando dos estrategias diferentes dependiendo de la masa de partida de nuestras muestras. Aquellas muestras con más de 750 ng de masa fueron tratadas con *BluePippin*, mientras que las muestras con menor cantidad de ADN de partida fueron tratadas con bolas magnéticas de *Ampure* (Beckman). En ambos casos se seleccionaron fragmentos >5 kpb. Esta tecnología de secuenciación proporcionó unas 100.000 secuencias por celda con un tamaño medio de inserto de 6 kpb y un tamaño de lectura medio de 16 kpb.

Finalmente, hemos participado en la fase beta de pruebas del secuenciador *Minion* de *Oxford Nanopore*. La preparación de las librerías se realizó a partir del ADN total amplificado con *MDA* siguiendo el protocolo previamente descrito (Ip et al., 2015), que a su vez estaba basado en el protocolo original del fabricante (Genomic DNA Sequencing Kit code SQK–MAP005). Tras 24 h de reacción de secuenciación se obtuvo un número muy bajo de secuencias con una longitud media de 4-10 kpb por lo que se decidió no continuar con esta tecnología.

1.2.8. Estudio metagenómico de la comunidad bacteriana mediante secuenciación del gen marcador ARNr 16S

Utilizando como molde el ADN total purificado de los sedimentos bacterianos, se amplificó el gen que codifica por el ARNr 16S a través de una *PCR* anidada de dos pasos. En el primer paso se amplificó desde la posición 341 a la 805 utilizando los oligonucleótidos Pro341-F y Pro805-R (Takahashi et al., 2014). A estos se les unió en su región 3' una pareja de secuencia adaptadoras: CS1FsL-Pro341f y CS2FsL-Pro805R (**Anexos - Oligonucleótidos**). Las amplificaciones se realizaron siguiendo el protocolo descrito en el trabajo anteriormente mencionado: desnaturalización inicial a 98ºC 30 sg, 25 ciclos de 10 sg de desnaturalización a 98ºC, un paso de alineamiento a 55ºC 30 sg y una extensión a 72ºC 30 sg finalizando con una única elongación final a 72ºC 2 min. En nuestro caso utilizamos el enzima Q5 Hot Start High-Fidely (*NEBiolabs*), 5 pmol de cada oligonucleótido y la *PCR* se hizo en un termociclador *MJ Mini™ Thermal Cycler* (*BioRad*). A continuación, realizamos una dilución 1/150 de los productos de *PCR* para reducir la concentración de los oligonucleótidos del paso anterior y utilizamos 6,5 µl de esta dilución para amplificar con otros 10 ciclos adicionales de *PCR* pero incorporando a la reacción 5 pmol de los oligonucleótidos P5Cs1 y un oligonucleótido de la serie P7bcCs2 para cada muestra. Este oligonucleótido incorpora una secuencia identificativa única (*barcode*) para cada producto de *PCR* (**Anexos - Oligonucleótidos**). Una mezcla equimolar de todos estos productos de *PCR* marcados se preparó tras su cuantificación en bioanalizador usando chips *ADN7500* (Agilent). Esta mezcla de amplicones se secuenció en equipos *MiSeq* de Illumina® tras preparar una librería usando el kit *MiSeq Reagent Kit v3: 600 ciclos* y se obtuvieron secuencias pareadas 2x300pb.

1.2.9. Clonaje de genomas completos de virus del papiloma humano

Las secuencias de dos genomas completos de nuevos papilomavirus humanos se amplificaron por *PCR* con oligonucleótidos específicos diseñados con *Primer3Plus* (**Anexos - Oligonucleótidos**) y la polimerasa *PrimeSTAR GXL ADN Polymerase* (Clontech) que permite la amplificación de grandes fragmentos de ADN (hasta 10 kpb). Se siguieron las instrucciones del fabricante: 30 ciclos de 98ºC durante 10 sg de desnaturalización, 60ºC de 15 sg de anillamiento y 68ºC de elongación 1 min/kpb. De esta manera se obtuvieron productos de *PCR* que se extrajeron de geles de agarosa al 1% (p/v) con el kit *QIAquick PCR Purification Kit* (*Qiagen*), eluyendo en 50 µl de Tris-HCl 10 M pH 8,5.

Para el caso del producto de *PCR* que incluye el genoma completo de Papilloma *HPV207* añadimos una cola de adeninas a su extremo 3´ incubando a 72ºC 20 min con 5 µl de tampón ThermoPol 10x (*NEB# B9004*), 1 mM de dATP, 0,2 µl de Taq polimerasa (*NEB*) y agua ultrapura libre de nucleasas (Ambion) hasta un volumen de reacción de 50 µl. Posteriormente, este producto de *PCR* se extrajo desde gel de agarosa y se ligó con el vector pGEM-T Easy Vector (Promega), siguiendo las instrucciones del

proveedor, durante 1 h a 16ºC y 30 min a temperatura ambiente. Por otro lado, el producto de amplificación del genoma completo de *HPV208* fue clonado directamente con el vector pSpark V (Canvax) siguiendo las instrucciones del fabricante, durante 60 min a 22ºC.

Los productos de la ligación se transformaron en bacterias ultracompetentes XL10-Gold (Agilent) mediante choque térmico y se sembraron en placas *LB*-Ampicilina (100 µg/ml) con Xgal (20mg/ml). Varias colonias blancas que habían incorporado el inserto se crecieron a 37ºC con agitación en 5 ml de *LB*-Ampicilina toda la noche. La miniprep de las células se realizó con el kit *ZR Plasmid Miniprep* (Zymo Research cat no. D4015) siguiendo las instrucciones del fabricante. La presencia de inserto se comprobó por tratamiento con las enzimas NcoI y XbaI para pGEM-t y BamH1 para pSpark V y electroforesis en geles de agarosa. La integridad de los extremos de los genomas clonados se comprobó por secuenciación Sanger utilizando los oligonucleótidos T7 y SP6 para pGEM-tEasy y M13 *forward* y *reverse* para pSpark V (**Anexos - Oligonucleótidos**). Los plásmidos se enviaron a HPV Information Centre, Karolinska Institutet (Suecia) para la confirmación de su secuencia por Sanger y asignación de un nombre en el caso de corresponderse a nuevos tipos de papilomavirus humanos.

2. Análisis bioinformático

2.1. Materiales

2.1.1. *Cluster* de cómputo

La mayor parte de los análisis bioinformáticos de esta tesis se realizaron utilizando el *cluster* de cómputo administrado por el Servicio de Bioinformática, y más recientemente por el servicio de Informática del CBMSO. Los recursos disponibles para los usuarios de este *cluster* son:

- 11 *CPUs* de 8 núcleos con 8 *GB* de *RAM*.
- 6 *CPUs* de 8 núcleos de 16 *GB* de *RAM*.
- 20 *CPUs* de 24 núcleos de 24 *GB* de *RAM*.
- 3 *CPUs* de 16 núcleos con 24 *GB* de *RAM*.
- 4 *CPUs* de 64 núcleos con 512 *GB* de *RAM*.
- Sistema de archivos con más de 91 *TB* de capacidad de almacenamiento.

2.1.2. Metagenomas secuenciados

Un total de 69 viromas y microbiomas fueron secuenciados utilizando equipos de secuenciación: 48 viromas en equipos *MiSeq* (Illumina®), 15 microbiomas en equipos *HiSeq2500* (Illumina®), y 6 microbiomas en equipos *PacBio (RSII)*. Adicionalmente, se secuenciaron 31 amplicones del gen marcador ARNr 16S mediante la tecnología *MiSeq* (Illumina®). En el caso de *MiSeq* y *HiSeq* se generaron secuencias pareadas contenidas en archivos digitales con extensión *.fastq* (R1 y R2), mientras que para PacBio® se generaron archivos con las lecturas crudas y otros archivos con las lecturas

ensambladas en *contigs*, todas en archivos con formato *.fastq*. Estas lecturas constituyen el material de entrada de nuestros análisis bioinformáticos. El software utilizado para la secuenciación y el análisis primario fue *MiSeq Control Software (MCS)* (Illumina®), mientras que el análisis secundario y la generación de los archivos *.fastq* se realizó con el software *MiSeq Reporter Software* (*MRS*) (Illumina®). La calidad de cada posición de las secuencias (*phred score*) está codificada según el código "Illumina 1.8+ Phred+33", donde la calidad de las lecturas va de 0 a 41. Las secuencias correspondientes a cada metagenoma fueron separadas atendiendo a sus identificadores (*barcodes*). En el caso de las secuencias procedentes de equipos *MiSeq* y *HiSeq2500* de Illumina® la separación se hizo sin permitir errores en las etiquetas de cada metagenoma, originando metagenomas denominados *M0*. La elección de trabajar con muestras *M0* se debe a que en estudios previos habíamos observado un grado significativo de contaminación cruzada entre muestras que se corrieron juntas en una misma carrera. Este problema se conoce como "sangrado de muestra", y consiste en la asignación errónea de las lecturas según el *barcode* (Mitra et al., 2015). La decisión de trabajar con lecturas *M0* nos hizo perder casi un 5% de las secuencias, pero redujo a menos de la mitad la contaminación cruzada entre metagenomas.

2.2. Métodos

Para el análisis de los viromas y microbiomas obtenidos en la anterior sección, desarrollamos un flujo de trabajo bioinformático que englobaba: preprocesado, ensamblaje de las lecturas, clasificación y anotación de los *contigs* resultantes, agrupación, estudios de diversidad alfa y beta y predicción de sus hospedadores (**Fig. 6**).

2.2.1. Preprocesado

2.2.1.1. Eliminación de secuencias de muy baja calidad

El PCM tiene un flujo de trabajo para la entrega de los resultados de secuenciación masiva en equipos de Illumina® que incluye la eliminación de secuencias de muy baja calidad que están marcadas con una *Y* identificativa en su cabecera. Este protocolo solo se aplica a secuencias M1 y no a secuencias M0 como las usadas en esta tesis, así que este paso lo tuvimos que hacer nosotros. Estas secuencias se eliminan automáticamente tras el proceso de demultiplexado.

2.2.1.2. Limpia de los adaptadores en los metagenomas obtenidos mediante *SISPA*

En los metagenomas que se generaron a partir de una amplificación *SISPA* se hizo un demultiplexado adicional en función del oligonucleótido empleado cuya secuencia fue posteriormente recortada incluyendo las 6-12 posiciones inespecíficas de su extremo 3´. Utilizamos el paquete de funciones *Biopieces* (Hansen, s. f.) con las funciones *find_adaptor*, para buscar en tres pasos consecutivos los 15 primeros nucleótidos del extremo 5', los 15 nucleótidos centrales y los 15 nucleótidos terminales del

extremo 3´ de cada oligonucleótido utilizado en *SISPA* permitiendo 1-2 errores, y la función *clip_adaptador* para recortarlas, generando un nuevo archivo con extensión *.fastq* como salida.

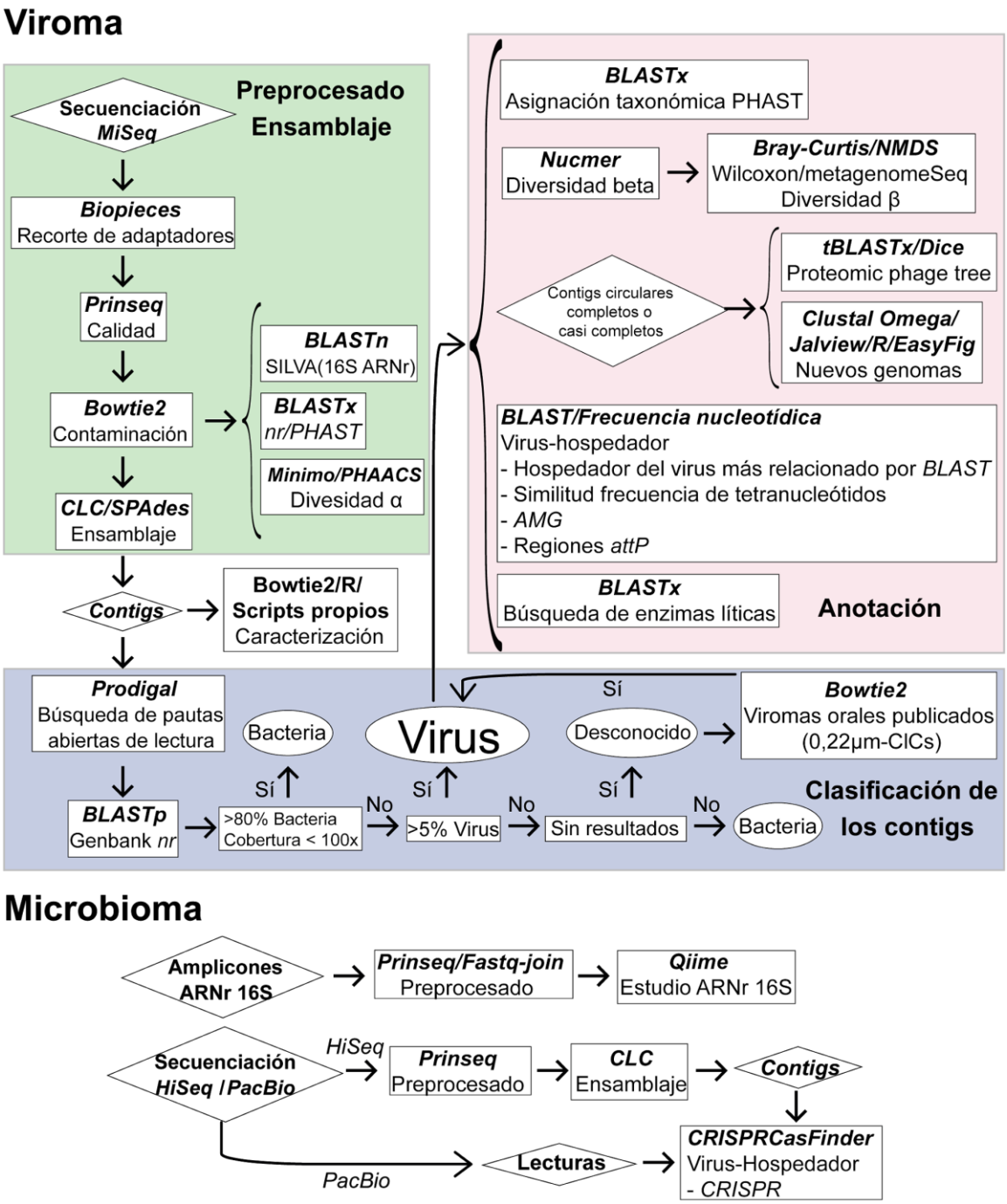

Figura 6. Flujo de trabajo global para el análisis bioinformático de los metagenomas de boca humana.

2.2.1.3. Filtración por calidad de secuencia

Las secuencias se filtraron según su calidad utilizando el software *Prinseq 0.19.3 lite* (Schmieder y Edwards, 2011). *Prinseq* se compone de un conjunto de herramientas que permite recortar o eliminar secuencias en función de una puntuación numérica de la calidad de cada posición denominada *phred*

quality score y que está codificada en el archivo *.fastq*. Con este programa eliminamos todas aquellas secuencias con más de 1% de Ns (–ns_max_p 1) o 3 Ns (–ns_max_n 3), se aplicó un filtro de complejidad basado en entropía que elimina secuencias por debajo de un umbral 50 de complejidad (–lc_method entropy –lc_threshold 50), se recortaron los extremos de cada secuencia si no alcanzaban un umbral de calidad media de 20 en ventanas de 2 nucleótidos con un paso de ventana de 1 nucleótido (–trim_qual_window 2 –trim_qual_step 1). Finalmente, solo aquellas secuencias con una longitud mínima de 100 nucleótidos (-min_len 100) y una calidad media global de 20 (-min_qual_mean 20) fueron consideradas para los siguientes análisis. Las secuencias de alta calidad generadas se almacenaron en archivos de salida multifasta sin calidad, con formato *.fasta*.

2.2.1.4. Eliminación de secuencias de genomas contaminantes conocidos

Utilizamos el alineador *Bowtie2* (Langmead y Salzberg, 2013), para buscar lecturas contaminantes que potencialmente pudieran encontrarse en nuestros metagenomas, como secuencias del genoma humano (*Genome Reference Consortium Human Build 37, GRCh37*), vectores y plásmidos (*UniVec; Junio 2014*) o el genoma del bacteriófago *ΦX174* (NC_001422.1). Este último genoma se utiliza como control interno por los sistemas de secuenciación de Illumina®. Se eliminaron las secuencias que alineaban bajo parámetros estrictos (--np 0 --n-ceil L,0,0.02 --rdg 0,6 --rfg 0,6 --mp 6,2 --score-min L,0,-0.2), que equivale a un 96% de identidad aproximadamente incluyendo inserciones y deleciones, a lo largo de la longitud total de la lectura. Se utilizó el comando *--un* para guardar solo aquellas lecturas que no alineaban con los genomas y bases de datos contaminantes. A estas secuencias se las designó secuencias de alta calidad sin contaminación conocida.

2.2.2. Estimación de los niveles de contaminación bacteriana

Se prepararon submuestras de 10.000 lecturas aleatorias desde los archivos de alta calidad sin contaminación conocida de cada metagenoma mediante la función *random_records* de *Biopieces.* Estas secuencias se compararon mediante *BLASTn* (Boratyn et al., 2012) contra la base de datos de 16S *SILVA* en su versión de Septiembre de 2014 (Pruesse et al., 2007). Los resultados obtenidos en formato tabular (-outfmt 8), se filtraron para eliminar aquellos resultados con un *e-value* < $1x10^{-10}$. El porcentaje de secuencias relacionadas con ARNr 16S para cada viroma se comparó con el obtenido para microbiomas obtenidos con ADN total para estimar el grado de contaminación.

2.2.3. Estimación de la composición taxonómica de las comunidades virales

Mediante la herramienta *random_records* de *Biopieces* se generaron archivos de 50.000 lecturas aleatorias a partir de los archivos de alta calidad sin contaminación conocida de cada metagenoma. Estos paquetes de lecturas se alinearon con *BLASTx* contra la base de proteínas no redundantes *nr* (Pruitt et al., 2005) del *GenBank*. Los resultados obtenidos con un *e-value* < $1x10^{-03}$ se categorizaron a nivel de

dominio. Utilizando las mismas submuestras, se evaluó la composición de familias virales de cada viroma mediante *BLASTx* contra la base de datos de proteínas virales *PHAST* en su versión de Agosto de 2017 (Zhou et al., 2011). Esta base de datos incluye las proteínas virales almacenadas en el *GenBank* y las proteínas virales encontradas en profagos caracterizados por la herramienta del mismo nombre. Solo aquellas secuencias con similitudes con proteínas de *PHAST* con *e-value* < $1x10^{-03}$ y *score* mínimo de 50 fueron categorizadas utilizando la herramienta *MEGAN v4.70.4* (Huson et al., 2016). Basándose en el mejor resultado de la secuencia más parecida, esta herramienta genera árboles taxonómicos a distintos niveles.

2.2.4. Estudio de diversidad alfa

El estudio de la diversidad alfa se realizó con la herramienta *Phage Communities from Contig Spectrum* (*PHACCS*, (Angly et al., 2005)). Esta herramienta utiliza la información contenida en el espectro de secuencias ensambladas en *contigs* para modelar la estructura de una comunidad viral y predecir su diversidad. Para ello, submuestras R1 de 100.000 lecturas de alta calidad y sin contaminación conocida de 300 pb fueron generadas con *random_records* del paquete *Biopieces* y se ensamblaron con el programa *Minimo* (*AMOS v3.1.0*) (Treangen et al., 2011). Este ensamblador *de novo* está basado en la teoría de grafos conocida como *Overlap-Layout-Consensus* y se usó con los parámetros por defecto: 98% de identidad y 35 nucleótidos de solapamiento mínimo. El tamaño y la abundancia media de los *contigs* se calculó con la herramienta *Genome relative Abundance and Average Size* (*GAAS)* (Angly et al., 2009), y el espectro de frecuencia de lecturas en *contigs* se calculó con la herramienta *Circonspect* (Angly et al., 2006). Con los datos procedentes de cada una de las herramientas calculamos la diversidad alfa con *PHACCS* utilizando los parámetros por defecto. Los resultados se expresaron en forma de índice de Shannon y riqueza estimada de especies.

2.2.5. Ensamblaje *de novo*

2.2.5.1. Eliminación de secuencias huérfanas y ordenación de lecturas pareadas

Los ensambladores basados en la teoría de grafos conocida como *De Bruijn Graph* permiten la reconstrucción de genomas completos o parciales (*contigs*) a partir de las lecturas secuenciadas. Uno de los ensambladores utilizados en esta tesis (*SPAdes*) requiere como dato de entrada, un archivo con las lecturas pareadas R1 y R2 alternadas en un único *multifasta*. Por ello, se buscaron y eliminaron aquellas secuencias no apareadas (huérfanas), cuya pareja había sido eliminada en anteriores pasos de preprocesado. Utilizamos el *script PairsOrphansGood_fasta_v2.py*, desarrollado por Ramón Peiró en el servicio de Genómica del CBMSO, para buscar en la cabecera de las secuencias información coincidente, y generar un nuevo archivo con las secuencias R1 y R2 alternadas y sin secuencias huérfanas. Las secuencias restantes se condensaron en un único archivo donde aparecían las secuencias

pareadas R1 y R2 alternadas, mediante el script *Interleave.py v0.1* (desarrollado por Mikael Karlsson, 2012).

2.2.5.2. Ensamblaje *de novo* basado en *De Bruijn Graph*

Para la reconstrucción de *contigs* desde los viromas y microbiomas, en esta tesis se han empleado dos ensambladores basados en *De Bruijn Graph*: *CLC Genomics Workbench v7.0.3* (*Qiagen*) *y SPAdes v3.11.1* (Bankevich et al., 2012), los cuales trabajan con subsecuencias de varias longitudes definidas (*k-mers*).

- *CLC* se empleó para el estudio de los viromas de mucosa bucal, placa dental y saliva. Las lecturas de alta calidad sin contaminación conocida de cada metagenoma se ensamblaron con este ensamblador de forma independiente con los parámetros por defecto, pero manteniendo solo los *contigs* >500 pb. Para el cálculo de la cobertura se estableció como parámetros de alineamiento un 97% de identidad en un 85% de solapamiento mínimo. Los viromas secuenciados por *MiSeq* y *HiSeq* se ensamblaron utilizando esta metodología.
- Por otro lado, *SPAdes* se utilizó para realizar un ensamblaje cruzado de submuestras de 500.000 secuencias pareadas R1+R2 de cada uno de los viromas utilizados en el estudio de los sesgos durante los procesos de purificación y amplificación. Se utilizaron los parámetros por defecto y *k-mers* con un tamaño de 27, 55, 77, 99 y 127 nucleótidos.

2.2.5.3. Estudio de los perfiles de cobertura

Cada *contig* generado en los viromas y microbiomas en esta tesis fue analizado para la obtención de una serie de parámetros y características como:

1. Los perfiles de cobertura de cada uno de los *contigs* se obtuvieron alineando las lecturas con *Bowtie2* utilizando parámetros estrictos (ver descripción anterior). Los mapas de alineamiento se extrajeron con *samtools* (*mpileup*) y se representaron en R utilizando la función *plot*.
2. Para el cálculo del perfil de complejidad lingüística nos basamos en la aproximación de *Trifonov* (Trifonov E.N., 1990) la cual relaciona la frecuencia de aparición de cada una de las bases que componen el ADN respecto a su frecuencia esperada. El cálculo se realizó en ventanas de 50 nucleótidos y pasos de 20 nucleótidos, usando el script propio *Trifonov_Complex.pl* (**Anexos - Script I**).
3. El estudio de sitios de unión preferente de los oligonucleótidos *SISPA* a secuencias específicas de nuestros *contigs* se hizo buscando subsecuencias de 8-15 nucleótidos dentro de los últimos 15 nucleótidos del extremo 3’ de cada oligonucleótido. Para ello se desarrolló un script propio: *Busqueda_Primers_Mapeo_SISPA.pl* (**Anexos - Script II**).
4. Para calcular el grado de homogeneidad de cobertura a lo largo de cada *contig* dentro de cada metagenoma se elaboraron curvas de Lorenz (Motley et al., 2014), que representan el porcentaje

acumulado de lecturas por posición del metagenoma. Un crecimiento lineal es indicativo de una mayor homogeneidad de cobertura a lo largo de la secuencia.

5. El porcentaje de CGs se calculó para cada *contig* en ventanas no solapantes de 100 nucleótidos y los perfiles se representaron con *plot*, del paquete *graphics* de R.
6. Se calculó la correlación de Pearson y el coeficiente de variación entre los perfiles de cobertura de aquellos *contigs* con una cobertura media > 100x en cada uno de los metagenomas que fueran objeto de comparación.

2.2.6. Clasificación taxonómica de los *contigs*

2.2.6.1. Asignación a dominios

Para clasificar los *contigs* de los viromas como virales o bacterianos en función de su contenido génico se desarrolló un *script* propio, llamando *ParseadorBLAST.pl v1.21* (**Anexos - Script III**). Para limitar el número de *contigs* fragmentados o poco abundantes consideramos sólo aquellos que cumplían al menos uno de los siguientes criterios:

1. Longitud mínima de 3.000 nucleótidos y cobertura 15x.
2. Longitud mínima de 10.000 nucleótidos y cobertura 4x.

Los *contigs* resultantes, denominados de aquí en adelante "*contigs* largos", se anotaron con el programa *Prodigal v2.6.3* (Hyatt et al., 2010) utilizando parámetros por defecto. Los genes predichos o pautas de lectura abiertas (*Open Reading Frames: ORFs*) se compararon mediante *BLASTx* contra la base de datos de proteínas *nr* del *GenBank*. Sólo se consideraron aquellos resultados con un *e-value* < $1x10^{-03}$. Para que un *contig* fuera clasificado como **bacteriano**, éste debía cumplir los siguientes dos requisitos:

1. Al menos el 80% de sus *ORFs* deberían encontrar una similitud en la base de datos con una proteína que no contuviera dentro de su nombre *phage*, *virus* o *capsid* o el nombre del organismo haga referencia directa a un virus.
2. La cobertura del *contig* debía ser menor a 100x, ya que un *contig* bacteriano procedente del genoma de una bacteria con un tamaño mínimo de 1 millón de nucleótidos y cobertura media 100x supondría al menos que el 15% de las secuencias del viroma son de esa bacteria, y esto había sido descartado previamente gracias al análisis del contenido en el gen 16S.

Aquellos *contigs* que no cumplían con estos criterios fueron analizados para ver si cumplían los criterios para ser *contigs* **virales**:

1. Al menos un 5% de los *ORFs* con resultados significativos por *BLASTx* contra la base de datos *nr* debían tener como mejor resultado una proteína que contuviera dentro de su nombre los términos *phage, virus, capsid* o que el nombre del organismo haga referencia directa a un virus.

2. Los *contigs* que no cumplían los requisitos para ser clasificados como bacterianos o virales se alinearon mediante *Bowtie2* con lecturas provenientes de cinco viromas humanos previamente publicados (Ly et al., 2014; Pride et al., 2011a; Reyes et al., 2010; Robles-Sikisaka et al., 2013; Willner et al., 2009) y que habían sido obtenidos tras filtración en 0,22µm y purificación en gradientes de cloruro de cesio. Este procedimiento garantiza la ausencia de contaminación bacteriana. Los *contigs* con más del 40% de su longitud cubierta por secuencias alineadas bajo parámetros estrictos de estos viromas se consideraron *contigs* virales.

Los *contigs* que no cumplían ninguno de estos requisitos se anotaron como *contigs* no clasificados.

2.2.6.2. Asignación taxonómica de los *contigs* virales

Los *ORFs* de los *contigs* virales obtenidos con *Prodigal* se compararon mediante *BLASTx* con la base de datos de proteínas virales *PHAST*. Para asignar estos *contigs* a la especie viral más cercana disponible en las bases de datos se consideraron solo los primeros cinco resultados significativos (*e-value* < $1x10^{-03}$) de cada *ORF* (siempre y cuando su *e-value* no superara en más de un 10% al *e-value* del mejor resultado obtenido, en cada caso). La asignación correspondió a la especie viral con mayor cantidad de resultados entre todos sus *ORF*.

2.2.7. Agrupación de *contigs* en *clusters*

Los *contigs* virales se agruparon en *clusters* siguiendo varias aproximaciones:

1. Alineamientos y agrupación por *Nucmer-MUMmer3* (Kurtz et al., 2004). *Nucmer* es un alineador que permite identificar regiones con diferente grado de similitud entre *contigs*. En este caso se agruparon en el mismo *cluster* los *contigs* que alineaban con >80% de identidad y una longitud mínima de solapamiento de 1.000 nucleótidos.
2. El método basado en *Markov Cluster Algorithm* (*MCL)* (Enright et al., 2002), es un algoritmo de agrupación de secuencias no supervisado basado en la simulación de conexiones en grafos, en el que los nodos más próximos representan aquellas parejas de *contigs* con una mayor similitud. En nuestro caso, las similitudes se calcularon mediante *BLASTn* (*e-value* < $1x10^{-05}$) entre parejas de *contigs*. Sólo aquellos nodos con tres o más conexiones internodos fueron representados.

La representación gráfica de estas conexiones entre *contigs* se hizo utilizando *Cytoscape v3.3.0* (Shannon et al., 2003), que permite la visualización y edición de redes.

2.2.8. Estudio de la diversidad beta

2.2.8.1. Obtención de tablas *BIOM* (abundancia de especies)

En esta tesis se han generado dos tipos de tablas de abundancias de especies o tablas *BIOM* (*Biological observation matrix*):

1. Para el estudio de los sesgos introducidos durante los pasos de purificación y amplificación al azar se hicieron ensamblajes cruzados de viromas de saliva. A continuación, se alinearon submuestras con 1.2000.000 secuencias de cada uno de los metagenomas contra los *contigs* usando *Bowtie2* (Langmead y Salzberg, 2013) y parámetros estrictos (96% de similitud).
2. Para el estudio de los viromas de mucosa y placa dental, éstos se ensamblaron por separado y se agruparon en *clusters* mediante *Nucmer* tal y como se describe en el apartado anterior. A continuación, se alinearon submuestras de 850.000 lecturas de cada viroma contra los *contigs* virales con *Bowtie2* bajo parámetros estrictos.

En ambos casos los resultados generados en formato .*sam* se procesaron con *samtools* (Li et al., 2009). Esta herramienta permite transformar los archivos .*sam* a archivos .*bam* (*samtools view*), ordenar las secuencias alineadas del archivo .*bam (sort*), mostrar el número de lecturas que mapean a cada *contig* (*mpileup*) y calcular el perfil de cobertura por nucleótido para cada *contig* (*idxstats*). Las lecturas alineadas a los *contigs* de un mismo *cluster* se sumaron y el resultado se normalizó por el tamaño en kilobases de cada *contig* (o el tamaño medio de los *contigs* del *cluster*) y por millón de lecturas (*RPKM, reads per kilobase per million of reads*).

Debido a que conocemos la existencia de una contaminación cruzada entre las lecturas de distintas muestras, decidimos eliminar el posible fondo de aquellos *contigs* alineados con 1-2 lecturas, considerándolos como 0. Cada *contig* no agrupado o cada *cluster* se consideró como una "Unidad taxonómica operativa" (*Operational taxonomic unit, OTU*) y la tabla de abundancias de *OTUs* se trató como una tabla *BIOM*.

2.2.8.2. Cálculo de distancias entre viromas y sistemas de ordenación

A partir de las tablas de abundancias de *OTUs* virales aplicamos distintos métodos de cálculo de distancias y métodos estadísticos para estudiar el grado de conexión entre los viromas. En esta tesis hemos usado como medidas de distancias: la disimilitud de Sørensen-*Dice* (Sørensen, 1948) ($d = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$), que solo considera la presencia y ausencia de *OTUs* compartidas, y la disimilitud de Bray-Curtis (Bray y Curtis, 1957) ($d = 1 - \frac{2Cij}{Si + Sj}$), que tiene en cuenta las abundancias de *OTUs*. Estas distancias se representaron en un plano mediante el sistema de ordenación *Non-metric multidimensional scaling* (*NMDS*) del paquete *vegan* (Oksanen et al., 2011), de *R*. Las diferencias estadísticas entre

ambientes y condiciones de salud y enfermedad se evaluaron mediante PERMANOVA (*Permutation anova*) (Anderson y Walsh, 2013), que es adecuada para distribuciones de datos que no cumplen con los criterios de normalidad como la mayor parte de las distribuciones de especies estudiadas en ecología.

Por último, para explorar cuáles eran las *OTUs* que presentaban una distribución diferencial en función de una determinada condición ambiental (tipo de muestra o estado de salud) aplicamos dos métodos: *Wilcoxon* (también conocido como *Mann-Whitney*), que es un test no paramétrico que sirve para descartar la hipótesis nula de que dos grupos de datos son iguales; y *metagenomeSeq*, que es una función propia del paquete *Qiime (Caporaso et al., 2010)*, la cual incorpora el modelo *fitZIG* para distribuciones infladas con ceros. Este sistema modela la distribución de conteos como una mezcla de dos distribuciones, una centrada en cero y una distribución normal. Los parámetros para este modelo mixto se estiman con un algoritmo de esperanza-maximización acoplado a un test estadístico *t*.

2.2.8.3. Árbol proteómico de bacteriófagos

El árbol proteómico se elaboró a partir de aquellos *contigs* que se consideraron completos o casi completos. Los criterios seguidos para seleccionar estos *contigs* fueron:

1. *Contigs* cuyo tamaño guardara una relación de al menos un 70% al tamaño del genoma del virus asignado en el **apartado 2.2.6.2**.
2. *Contigs* de naturaleza circular. Para ello, se emplearon dos estrategias: el programa Minimus (AMOS v3.1.0) (Treangen et al., 2011) comprueba si los extremos del *contig* son solapantes, y el script *ClosetheCircle.pl* (desarrollado por David Abia del servicio de Bioinformática del CBMSO), que ejecuta dos *BLASTn* consecutivos entre las lecturas del metagenoma y los extremos de los *contigs* ensamblados para determinar si son circulares.

Una vez seleccionados los *contigs*, añadimos al estudio 233 genomas completos de bacteriófagos relacionados por secuencia con nuestros *contigs* virales y 49 genomas de bacteriófagos de subfamilias de *Caudovirales* de referencia como control interno de la agrupación. Comparamos dos a dos todos estos genonas mediante *tBLASTx* y aquellos resultados con un *e-value* < $1x10^{-05}$ y con una identidad mínima de un 35% sobre una longitud mínima de alineamiento de 45 aminoácidos se consideraron significativos. Con estos alineamientos calculamos una distancia basada en una variante de la distancia Sørensen-*Dice* (Sørensen, 1948) empleada en trabajos previos (López-Pérez et al., 2017; Mizuno et al., 2013), pero con modificaciones. Así, en lugar de utilizar los valores de *score* de un *BLASTn* entre los *contigs*, tuvimos en cuenta el logaritmo en base 10 de los nucleótidos alineados entre los *contigs* según la fórmula: $d = 1 - \frac{log10\ A}{log10\ B}$, donde A es el menor número de pares de bases alineadas en ambas direcciones, y B representa el tamaño del menor de los dos *contigs*. El proceso completo se automatizó mediante el *script* propio *Famio_Breadth.pl* (**Anexos - Script IV**). Una vez obtenida la tabla de distancias calculamos un

árbol proteómico mediante el algoritmo de *Neighbor joining*, el cual se encuentra contenido en el paquete *Phylip v3.6.9.7* (Plotree y Plotgram, 1989). El árbol proteómico se guardó en formato *.newick*. y la representación gráfica se hizo con *Dendroscope v3.5.7.* (Huson y Scornavacca, 2012).

Este método se comparó con un método similar recientemente publicado y disponible en el servicio online *ViPTree* (Nishimura et al., 2017) que se basa también en el cálculo de la distancia Sørensen-Dice. Esto justifica la agrupación correcta en ambos métodos de los miembros de referencia de varias subfamilias dentro del orden *Caudovirales* (**Fig. 7**).

Los *contigs* y genomas de referencia agrupados en las mismas ramas del árbol proteómico se juntaron en un nivel de agregación superior denominado *megacluster*.

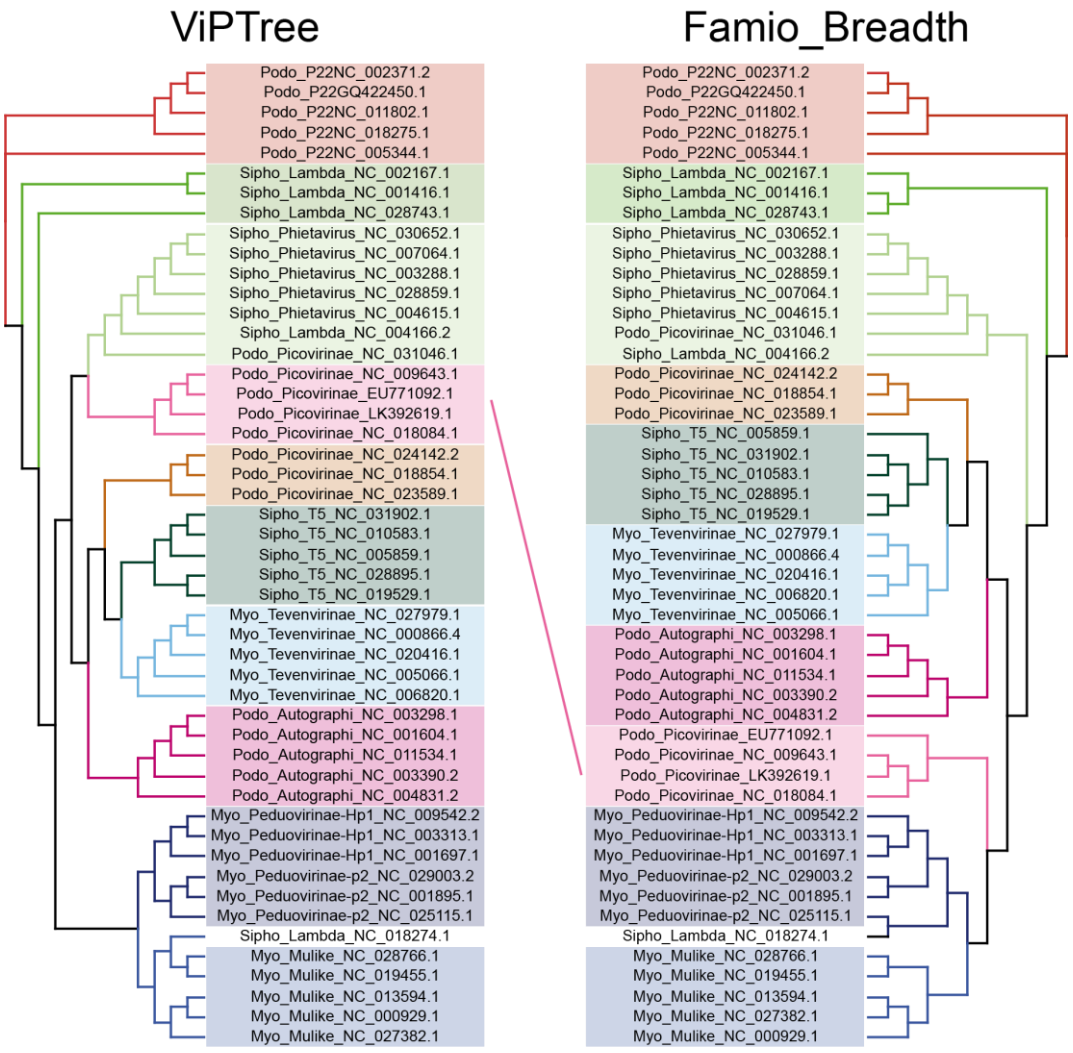

Figura 7. Comparación de los árboles proteómicos de bacteriófagos generados con *ViPTree* y el *script* propio *Famio_Breadth.pl*. Se muestra la agrupación coherente de ambos métodos de virus pertenecientes a varias subfamilias (indicados con el mismo color) dentro del orden *Caudovirales*.

2.2.9. Estudios filogenéticos basados en genes virales conservados

Los genes de interés contenidos en *contigs* virales junto con los más relacionados o representativos encontrados en las bases de datos mediante *BLAST* se alinearon con la herramienta de alineamiento múltiple *ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/)* (Sievers et al., 2011), que proporciona resultados de porcentaje de nucleótidos o aminoácidos alineados para cada pareja de secuencias y la información de alineamiento múltiple en archivos con formato *.aln*. El resultado se procesó con el programa *Jalview* (Waterhouse et al., 2009) que permite la visualización y edición manual de los alineamientos y genera un árbol filogenético mediante el algoritmo *Neighbor joining* en formato *.newick.* Alternativamente usamos el programa Mega 5 (Tamura et al., 2011) o el paquete *phangorn* (Schliep, 2011) de *R* para generar árboles de máxima verosimilitud. Los árboles filogenéticos se visualizaron en *Dendroscope v3.5.7.*

2.2.10. Estudios de sintenia

Varios representantes de alguno de los *megaclusters* se alinearon entre sí y con el genoma de referencia del bacteriófago más cercano utilizando la herramienta *BLASTn*. Los resultados con un *e-value* < $1x10^{-03}$ se visualizaron mediante la aplicación *EasyFig 2.2.3*. (Sullivan et al., 2011) utilizando parámetros de visualización por defecto. Esta herramienta aplica un código de color en función del porcentaje de identidad de cada alineamiento.

2.2.11. Búsqueda de lisinas y holinas

Utilizando los resultados obtenidos en la comparación con la base de datos *nr* del *GenBank* para cada *ORF* de cada *contig* viral (*e-value* < $1x10^{-03}$), se seleccionaron aquellos genes relacionados con lisinas y holinas. Para ello, se tomaron como positivos aquellos resultados que contuviesen dentro del nombre de las proteínas los términos *lysin*, *endolysin*, *cell wall hydrolase*, *endopeptidase* o *amidase* en el caso de las lisinas, y *holin* en el caso de las holinas.

2.2.12. Predicción de hospedador

Hemos utilizado cinco aproximaciones para predecir los potenciales hospedadores bacterianos de los *contigs* virales.

1- La primera asume que el hospedador anotado del virus más cercano por *BLASTx* contra *nr* coincide con el hospedador real del *contig* a un determinado nivel taxonómico. Se utilizó como virus más relacionado en las bases de datos el asignado en el **apartado 2.2.6.2**.

2- La segunda es la comparación de perfiles de tetranucleotídicos entre *contigs* virales y genomas de bacterias orales disponibles en la base de datos Human Oral Microbiome Database (*HOMD)* (http://www.homd.org/). Para ello, se calcularon las frecuencias tetranucleotídicas con el *script* propio

Calculo_frecuencias_nucleotidicas.R (**Anexos - Script V**) y a continuación se calculó la correlación de Pearson entre todos los perfiles de tetranucléotidos. Aquellas conexiones entre *contigs* virales y bacterias con una correlación de Pearson mayor de 0,96 fueron consideradas como positivas. La validación de este criterio se hizo comparando los perfiles de tetranucleótidos de bacteriófagos y sus hospedadores conocidos mediante correlaciones de Pearson. Para ello, seleccionamos 679 profagos descargados de ProphageDB en Marzo de 2017 y 1.926 genomas de bacteriófagos que infectan 439 bacterias bucales (cuyos genomas completos se descargaron del Human Oral Microbiome Database en Marzo de 2017) (**Fig. 8**). Tolerando un 10% de falsos positivos consideramos válida una correlación de Pearson >0.94 para la asignación de hospedador a nivel de familia, >.93 a nivel de Orden, >0.89 a nivel de Clase y >0.84 a nivel de Filo. El cálculo se hizo utilizando un *script* propio en *R*.

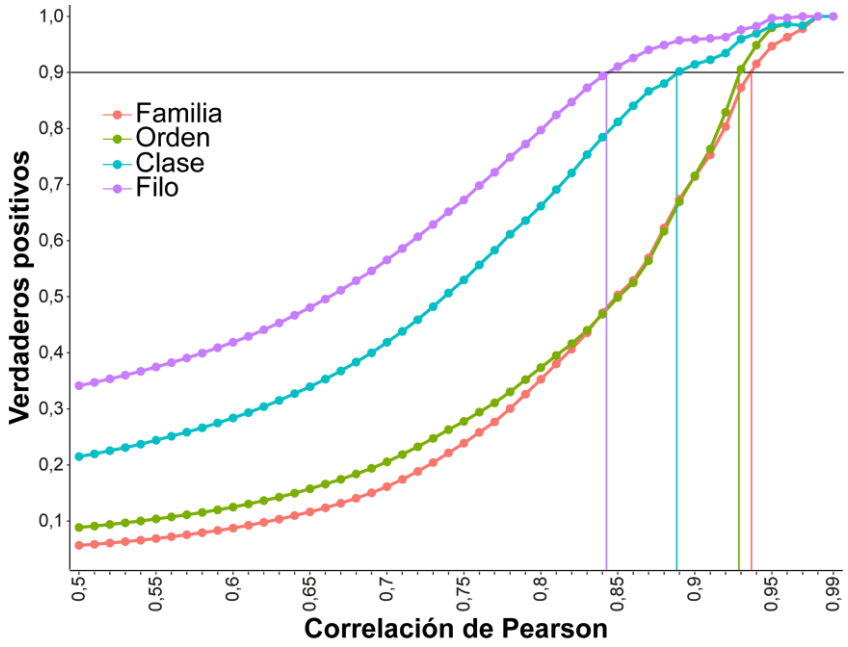

Figura 8. Sistema de evaluación de la correlación entre el número de asignaciones taxonómicas idénticas y la correlación de Pearson entre parejas de genomas de bacteriófagos-profagos y las bacterias que se sabe que infectan.

3- La tercera aproximación se basa en la búsqueda de genes metabólicos propios de genomas bacterianos en los *contigs* virales, asumiendo que estos genes podrían tener su origen en bacterias relacionadas con sus hospedadores. Para ello, se estudió la función de los genes a través de modelos de Markov ocultos (*Hidden Markov Models*: HMM), mediante la herramienta *eggNOG* (http://eggnogdb.embl.de/#/app/emapper) (Huerta-Cepas et al., 2017, 2016), que engloba a las bases de datos de matrices de resultados de *HMM:* pfam, *KEGG* y *COG* (Finn et al., 2016; Kanehisa et al., 2017, 2016; Kanehisa y Goto, 2000; Tatusov et al., 2000) y proporciona las asignaciones más probables. Siguiendo un método de puntuación propio basado en la cantidad de resultados logrados para cada función y grupo funcional, el sistema predijo una serie de funciones metabólicas. Los genes cuya función asignada fue "genes metabólicos" se compararon mediante *BLASTx* contra la base de datos *nr* (*e-value* < $1x10^{-50}$) para identificar la bacteria más relacionada en la base de datos y probable hospedador del *contig* que contenía el gen metabólico.

4- La cuarta aproximación se basa en el estudio de las secuencias de integración de los bacteriófagos lisogénicos: attPs. Las regiones attP guardan similitud de secuencia con algunas secuencias de ARN transferente bacteriano conocidas como attB, ya que son los sitios donde se integran. Utilizando la aplicación *tRNAscan-SE v1.23* (Lowe y Eddy, 1997) se identificaron algunas de estas posibles secuencias de integración, desvelando a la vez el tipo de hospedador en el que podrían eventualmente integrarse. La taxonomía asociada a cada attP se determinó mediante *BLASTn* contra la base de datos de nucleótidos *nt*, requiriendo un alineamiento mínimo de 35 nucleótidos y una identidad del 90%.

5- La última estrategia consistió en el análisis de secuencias *CRISPRs* presentes en los *contigs* de los microbiomas de *HiSeq* y PacBio®. Utilizamos la herramienta *CRISPRCasFinder* (Abby et al., 2014; Couvin et al., 2018; Grissa et al., 2007), bajo parámetros por defecto. Esta herramienta *online* busca secuencias repetidas contiguas de tamaño variable separadas por diferentes secuencias con un tamaño medio de 35 nucleótidos (espaciadores). La secuencia de estos espaciadores puede proporcionar un registro histórico de los virus con los que esa bacteria ha tenido contacto en el pasado y nos permite por tanto establecer relaciones virus-hospedador. Solo tuvimos en cuenta aquellos *CRISPRs* con al menos 4 espaciadores, ya que un número menor pueden resultar falsos positivos (Kupczok et al., 2015). La secuencia de los espaciadores se comparó con la colección de *contigs* virales para identificar cuáles habían podido infectar en el pasado las bacterias secuenciadas mediante *HiSeq* o PacBio®. Se consideraron positivos los alineamientos entre los *contigs* virales y los separadores de los *CRISPRs* de cada metagenoma que tenían una cobertura mínima de alineamiento del 95% de la longitud del separador y una identidad del 95% (Shmakov et al., 2017). La asignación taxonómica del *contig* bacteriano que contenía el CRISPR se hizo por *BLASTx* frente a la base de datos *nr* con un *e-value* < $1x10^{-03}$).

2.2.13. Análisis metagenómico basado en el gen marcador para ARNr 16S

Los amplicones 16S procedentes de las muestras de sedimento bacteriano fueron analizados con el paquete de herramientas *Qiime* (Caporaso et al., 2010). Las secuencias R1 y R2 se asociaron a las distintas muestras en base a sus etiquetas o *barcodes* identificativos. A continuación, se ensamblaron en una sola secuencia a partir de sus extremos solapantes con el script *fastq-join* (https://github.com/brwnj/fastq-join) (Aronesty, 2013), permitiendo un porcentaje máximo de diferencia del 20% (-p 20), y se elaboró un archivo *map_file* con la información necesaria para que *Qiime* ejecutara sus funciones. Este archivo se validó con el script *validate_mapping_file.py*. Posteriormente, a través del script *pick_open_reference_otus.py*, se generó una tabla *BIOM* con la información de todas las secuencias agrupadas en *OTUs* mediante el sistema *uclust* (Edgar, 2010) y la información de referencia de todos los genomas agrupados al 97% (archivo *.biom*). Esta tabla se cargó en *R* y se analizó con el paquete *phyloseq* (McMurdie y Holmes, 2013) para calcular disimilitudes Bray-Curtis y representar los metagenomas en un sistema de ordenación *NMDS*.

RESULTADOS

1. Estudio de los sesgos introducidos durante el enriquecimiento de partículas virales y la amplificación inespecífica de sus genomas

La introducción de sesgos durante los estudios metagenómicos de las comunidades de virus es un problema conocido que compromete la extracción de datos cuantitativos y limita la obtención de conclusiones biológicas. Para comprender mejor las causas de estos sesgos y el impacto que pudieran tener en los estudios de comparación entre comunidades, en esta tesis doctoral nos propusimos estudiar una comunidad sintética de composición conocida y una comunidad natural de saliva humana a lo largo de las distintas etapas de un protocolo sencillo de enriquecimiento de partículas virales y amplificación inespecífica de sus genomas.

1.1. Evaluación de los sesgos en comunidades sintéticas de virus de ADN

Para este estudio se prepararon dos comunidades sintéticas (en adelante comunidades control) formadas por siete virus de ADN con características morfológicas y tipos de genomas diferentes, tratando de reflejar la diversidad natural de virus de ADN asociados a ecosistemas humanos (**Tabla 1**).

		Morfología		Genoma		Proporción teórica (%)	
Virus	**Familia**	**Estructura**	**Diámetro (nm)**	**Tipo**	**Tamaño (kpb)**	**Comunidad sintética 1**	**Comunidad sintética 2**
Virus Vaccinia Western Reserve (WR)	*Poxviridae*	Envuelta, virión en forma de ladrillo	250x360	ADNbc lineal	194,7	14,28	16,65
Bacteriófago λ (Lambda)	*Siphoviridae*	Sin envuelta, cabeza-cola	60	ADNbc lineal	48,5	14,28	16,65
Adenovirus humano 5 (AdenoV)	*Adenoviridae*	Sin envuelta, pseudo-cápsida T=25	90	ADNbc lineal	35,9	14,28	16,65
Bacteriófago Φ29 (Phi29)	*Podoviridae*	Sin envuelta, cabeza-cola	54	ADNbc lineal	19,3	14,28	16,65
Fago M13 (M13)	*Inoviridae*	Sin envuelta, filamentoso	7x700-2.000	ADNmc circular	6,4	14,28	16,65
Virus diminuto del ratón p (MVMp)	*Parvoviridae*	Sin envuelta, cápsida T=1	23	ADNmc lineal	5,1	14,28	16,65
Circovirus porcino 2a (PCV2a)	*Circoviridae*	Sin envuelta, cápsida T=1	17	ADNmc circular	1,8	14,28	0,075

Tabla 1. Características de los virus incluidos en las comunidades sintéticas. ADNmc: ADN monocatenario, ADNbc: ADN bicatenario.

Un aspecto importante en el diseño de protocolos de enriquecimiento de partículas virales es la eliminación de células y de material genético libre. Sin embargo, decidimos no incluir bacterias en nuestras comunidades control, al observar en experimentos previos que la centrifugación a baja velocidad y filtración en filtros de 0,45μm de cultivos puros de *Escherichia coli*, *Staphylococcus aureus* y *Roseobacter litoralis* provocaba una reducción de 7-8 órdenes de magnitud en el número de colonias viables.

Para preparar comunidades control con cantidades equivalentes de material genético de cada uno de los siete virus de ADN se estimó el número de genomas protegidos de la acción de nucleasas por *PCR* cuantitativa. De esta manera evitamos la sobrestimación de genomas de virus cuyas estructuras pudieran verse dañadas durante los pasos de purificación o preservación del preparado viral. Por otro lado, la *PCR* cuantitativa permite cuantificar de forma eficiente genomas de virus de ADN de cadena sencilla, los cuales se detectan mal mediante técnicas de tinción (Holmfeldt et al., 2012) frecuentemente empleadas en la cuantificación de los componentes de comunidades sintéticas. Estas comunidades control con cantidades balanceadas de los siete genomas de virus de ADN se alicuotearon y duplicados o triplicados biológicos de éstas se analizaron mediante *PCR* cuantitativa (cada uno de ellos con triplicados técnicos) antes y después de ser sometidas por separado a varios protocolos de enriquecimiento viral y amplificación al azar de sus genomas. En algunos experimentos se evaluó el efecto de varios de estos protocolos combinados.

Como se aprecia en la **Figura 9**, las proporciones de los genomas de los siete virus protegidos de la acción de las nucleasas eran parecidas en términos generales. Sin embargo, la comunidad control 1 mostró una inesperada baja proporción de genomas virales de Vaccinia WR con respecto al resto de virus (0,26% de media), probablemente debido a problemas durante su conservación a 4ºC, que podrían haber afectado a la integridad de su estructura. De igual modo, la comunidad control 2 mostró una esperada baja representación de PCV2a debido al agotamiento del único preparado disponible de este virus (0,11% de media) (**Fig. 9A,B**). De forma general, las centrifugaciones a baja velocidad y las filtraciones redujeron la cantidad de algunos genomas virales protegidos (**Fig. 9C,D**). Entre ellos destaca la caída de 27-150 veces en el número de genomas del virus de gran tamaño Vaccinia WR tras la centrifugación a baja velocidad, y de más de 500 veces en dos de las tres réplicas tras la filtración en 0,22μm en el caso de la comunidad control 2. Estos sesgos negativos causan una reducción drástica en la abundancia relativa de este virus (22,1-41,8% a 1,4-2,0%) (**Fig. 9B**). De igual manera, el número de genomas del virus Vaccinia WR cayó a niveles casi indetectables durante las centrifugaciones y filtración en 0,22μm de la comunidad control 1 (**Fig. 9C**). También detectamos que los virus de pequeño tamaño (M13, MVMp y PCV2a) se vieron más afectados de forma general por los pasos de centrifugación, y en menor medida, por las filtraciones y el colchón de iodixanol, que el resto de los virus de mayor tamaño (Lambda, Phi29 y AdenoV). Las diferencias en el efecto de la centrifugación entre estos dos grupos de virus resultaron estadísticamente significativas (*p-value* = 0.00082, Mann-Whitney), y la combinación de algunos de estos protocolos de purificación en la comunidad control 2 redujo la cantidad de virus pequeños en un rango de 6,2 a 10 veces. De entre todos los protocolos de enriquecimiento utilizados, la concentración de partículas virales mediante colchón de iodixanol fue el paso del protocolo que mejor preservó la composición original de las comunidades control, incluido el virus Vaccinia WR.

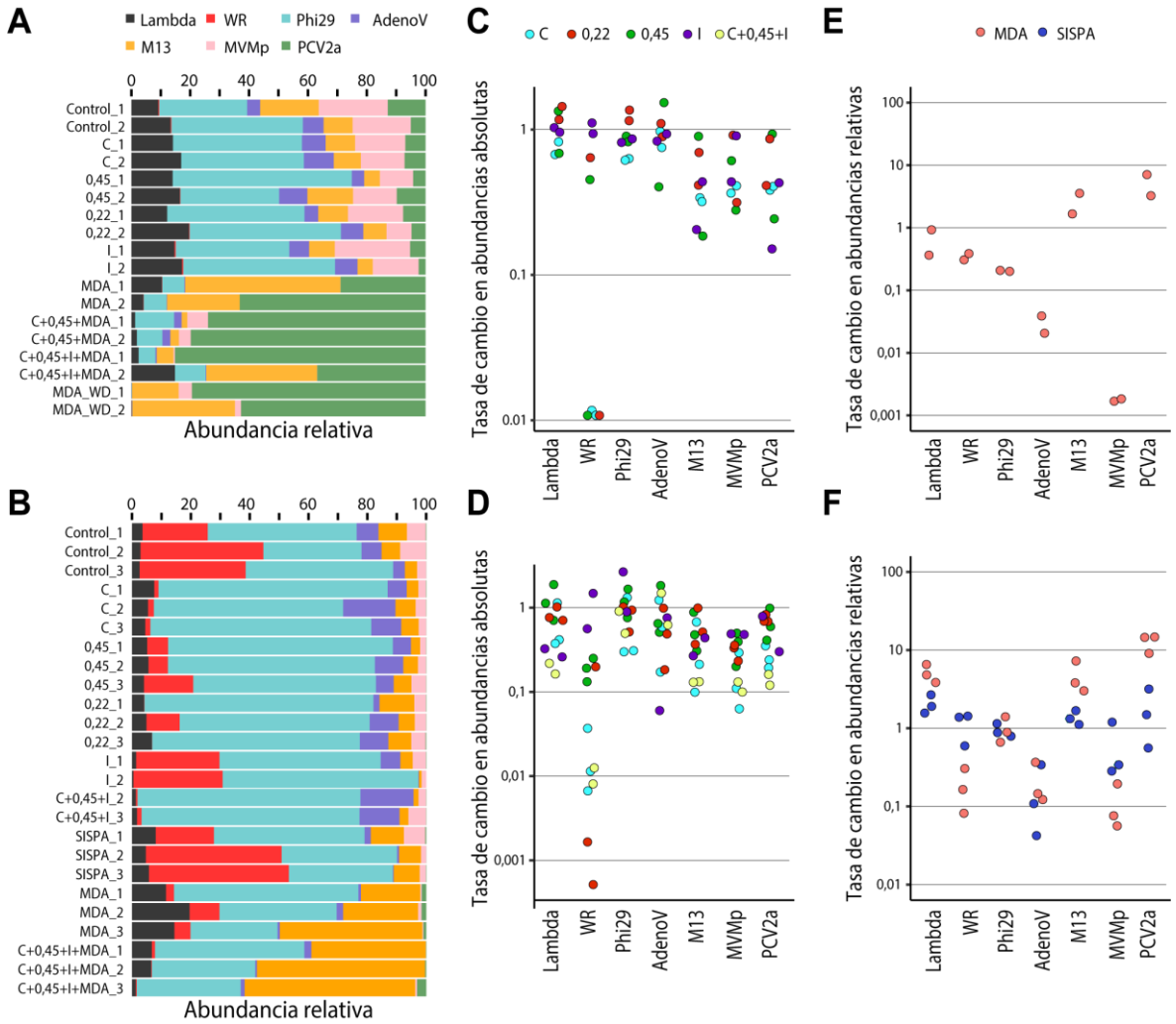

Figura 9. Impacto del enriquecimiento viral y la amplificación al azar de los genomas virales en la composición de las comunidades control. Las proporciones relativas de los siete virus de ADN (Lambda = bacteriófago λ; WR = Vaccinia WR; Phi29 = bacteriófago Φ29; AdenoV = adenovirus humano 5; M13 = bacteriofago M13; MVMp = Parvovirus diminuto de ratón cepa p; PCV2a = Circovirus porcino 2a) de la comunidad control 1 (A) y la comunidad control 2 (B) fueron evaluadas por *PCR* cuantitativa antes y después de cada tratamiento de forma independiente o combinados. Dos o tres réplicas independientes se evaluaron para cada muestra y se anotaron con números del 1 al 3. Las cantidades de genomas relativos a la muestra control sin tratar se muestra para la comunidad control 1 (C) y 2 (D). Las proporciones de genomas virales después de las amplificaciones al azar relativizadas a las proporciones de la comunidad sin tratar se muestran para la comunidad control 1 (E) y 2 (F). Los siguientes identificadores se utilizaron para designar el tratamiento empleado: Control: comunidad viral control sin tratar; C: dos pasos de centrifugación consecutivos a 3.000 g 10 min; 0,45 y 0,22: tamaño de poro expresado en µm usado durante la filtración con jeringa; I: colchón de iodixanol; *MDA*: amplificación por desplazamiento múltiple de banda con el kit de GenomiPhi™ V2; MDA_WD, amplificación por desplazamiento múltiple de banda sin paso de desnaturalización; SISPA, Amplificación independiente de secuencia con un oligonucleótido único. En todos los casos los genomas virales de las comunidades tratadas y no tratadas se extrajeron tras incubación con nucleasas que eliminan material genético no protegido por cápsidas o envueltas.

Tal y como se había publicado previamente (Kim y Bae, 2011), la amplificación al azar mediante *MDA* provocó la sobrerrepresentación de genomas pequeños circulares de ADN de cadena sencilla (M13 y PCV2a). Así el genoma del bacteriófago M13 aumentó su abundancia relativa de 1,7-3,6 a 3-7,2 veces

en las comunidades control 1 y 2, respectivamente, mientras que PCV2a tuvo una sobreamplificación de 3,2-7,1 y 9,1-14,7 veces en las comunidades control 1 y 2, respectivamente. Como era de esperar, la ausencia de desnaturalización durante el protocolo de *MDA* evitó la hibridación de oligonucleótidos a las moléculas de cadena doble de ADN, aumentando los sesgos hacia genomas de cadena sencilla circulares (MDA_WD, **Fig. 9A**). Por el contrario, el genoma lineal de cadena sencilla del virus MVMp mostró una bajada pronunciada de hasta 500 veces en su abundancia relativa, lo que ocurría de forma consistente en los cinco experimentos de amplificaciones *MDA* ensayados. Este sesgo negativo también fue detectado para el genoma del virus AdenoV aunque en menor medida.

A diferencia de las amplificaciones *MDA*, *SISPA* preserva bastante bien la composición de las comunidades control. La única excepción fue AdenoV, el cual presentó una disminución en sus proporciones relativas en las tres réplicas experimentales en un rango de 2,94 a 23,6 veces. La pérdida del número de copias de AdenoV en las dos estrategias alternativas de amplificación empleadas, y de MVMp durante la amplificación mediante *MDA*, requiere de un estudio en mayor profundidad.

1.2 Evaluación del sesgo introducido por protocolos de amplificación al azar en viromas de saliva humana

A partir de una muestra formada por la mezcla de saliva de varias personas evaluamos los sesgos introducidos por los diferentes protocolos de amplificación al azar de ADN. Mediante secuenciación masiva en equipos *MiSeq* (Illumina®) obtuvimos nueve viromas procedentes de esta mezcla, incluyendo un viroma sin amplificar (Unamp1), seis viromas generados mediante amplificación con dos kits comerciales de *MDA* (GenomiPhi™: MDA_G1-4; y TruePrime™: MDA_T1-2) y otros dos con *SISPA* (SISPA1-2). Estos viromas contenían un promedio de 1.566.548 lecturas, que tras el filtrado por calidad se quedaron en 1.490.980 lecturas de media por viroma (1.097.629-2.011.102 lecturas).

1.2.1. Sesgo estocástico en la amplificación por *MDA* a partir de picogramos de ADN molde

En la **Figura 10A** se muestran 277 *contigs* obtenidos a partir de un ensamblaje cruzado de medio millón de lecturas de cada viroma y cuya abundancia relativa había aumentado o disminuído al menos 50 veces en los viromas amplificados en comparación con Unamp1. Los viromas de *MDA* amplificados desde 1 ng mostraron patrones de *contigs* sesgados similares entre sí, con una influencia mínima del factor tiempo de extensión durante la amplificación de GenomiPhi™ (2,5 h y 10 h en MDA_G1 y MDA_G2, respectivamente) o de la estrategia elegida para el cebado de la polimerasa (GenomiPhi™: MDA_G1, y TruePrime™: MDA_T1). Por el contrario, en las muestras amplificadas desde 10 pg de molde (MDA_G3, MDA_G4, y MDA_T2) se registró un incremento notable en el número de *contigs* sesgados, que mostraban además patrones muy diferentes entre sí. También observamos que el perfil de *contigs* sesgados en las muestras de *MDA* difiere notablemente de aquellos encontrados en los viromas de *SISPA*. Estos resultados muestran que las amplificaciones desde bajas cantidades de ADN no solo

introducen más sesgo, si no que incrementan su variabilidad entre réplicas obtenidas con procedimientos muy parecidos.

			Mejor resultado *BLASTx*			Extremos solapantes	
Contig	**Tamaño (pb)**	**Tasa de cambio**	**Especies**	**Familia**	***e-value***	***Close_the_circle***	***Minimus2***
1.473	3.117	2.965	*Enterobacteria phage I2-2*	*Inoviridae*	$1x10^{-9}$	SI	SI
917	4.832	501	*Microviridae Fen7918_21*	*Microviridae*	$4x10^{-84}$	SI	SI
640	6.738	356	*Microviridae Fen685_11*	*Microviridae*	$3x10^{-24}$	SI	SI
732	5.884	277	*Microviridae IME-16*	*Microviridae*	0	SI	SI
1.041	4.332	253	*Microviridae IME-16*	*Microviridae*	0	NO	NO
1.084	4.182	205	*Vibrio phage fs2*	*Inoviridae*	$2x10^{-21}$	SI	SI
45	39.552	168	*Dickeya phage Limestone*	*Myoviridae*	$5x10^{-43}$	NO	NO
781	5.536	153	*Ralstonia phage p12J*	*Inoviridae*	$2x10^{-12}$	SI	SI
674	6.397	140	*Parabacteroides phage YZ-2015a*	*Microviridae*	$4x10^{-31}$	SI	SI
211	18.180	130	*Mycobacterium phage DrDrey*	*Siphoviridae*	$2x10^{-21}$	NO	NO
218	17.800	114	*Bacillus phage AR9*	*Myoviridae*	$3x10^{-18}$	NO	NO
1.431	3.182	86	*Porcine stool-associated circular virus 5*	*Circoviridae*	$7x10^{-131}$	SI	SI
413	10.049	55	*Enterobacteria phage Min27*	*Podoviridae*	$1x10^{-18}$	NO	NO
1.465	3.125	52	*Enterobacteria phage I2-2*	*Inoviridae*	$6x10^{-9}$	SI	SI
977	4.555	50	*Gokushovirus WZ-2015a*	*Microviridae*	$3x10^{-7}$	NO	SI

Tabla 2. Naturaleza circular de los *contigs* más sobrerrepresentados en los viromas MDA_G1 y MDA_G2. Solo se muestran aquellos *contigs* con una tasa de cambio >50x en MDA_G1 y MDA_G2.

MDA amplifica los plásmidos pequeños y los genomas virales circulares de manera más eficiente que las moléculas de ADN lineal (Kim y Bae, 2011) (**ver sección anterior**). De acuerdo con estos resultados, 10 de los 15 *contigs* con mayor sesgo positivo en las muestras MDA_G1 y MDA_G2 eran *contigs* pequeños con extremos solapantes, sugiriendo su naturaleza circular o mostraron mejor similitud de secuencia por *BLAST* con miembros conocidos de la familia *Microviridae*, formada por virus de genoma pequeño y circular (**Tabla 2**). Este sesgo sistemático hacia genomas circulares pequeños también podría explicar la sobreamplificación de muchos *contigs* en el viroma MDA_T1, pero no la enorme variabilidad de *contigs* con sesgo positivo observado cuando se usaron 10 pg como molde (MDA_G3, MDA_G4 y MDA_T2) ya que muchos de sus *contigs* con sesgos positivos no eran pequeños ni circulares (**Fig. 10C**).

1.2.2. La amplificación mediante *SISPA* y *MDA* de viromas de saliva introduce un sesgo sistemático asociado a regiones de contenido extremo de CG

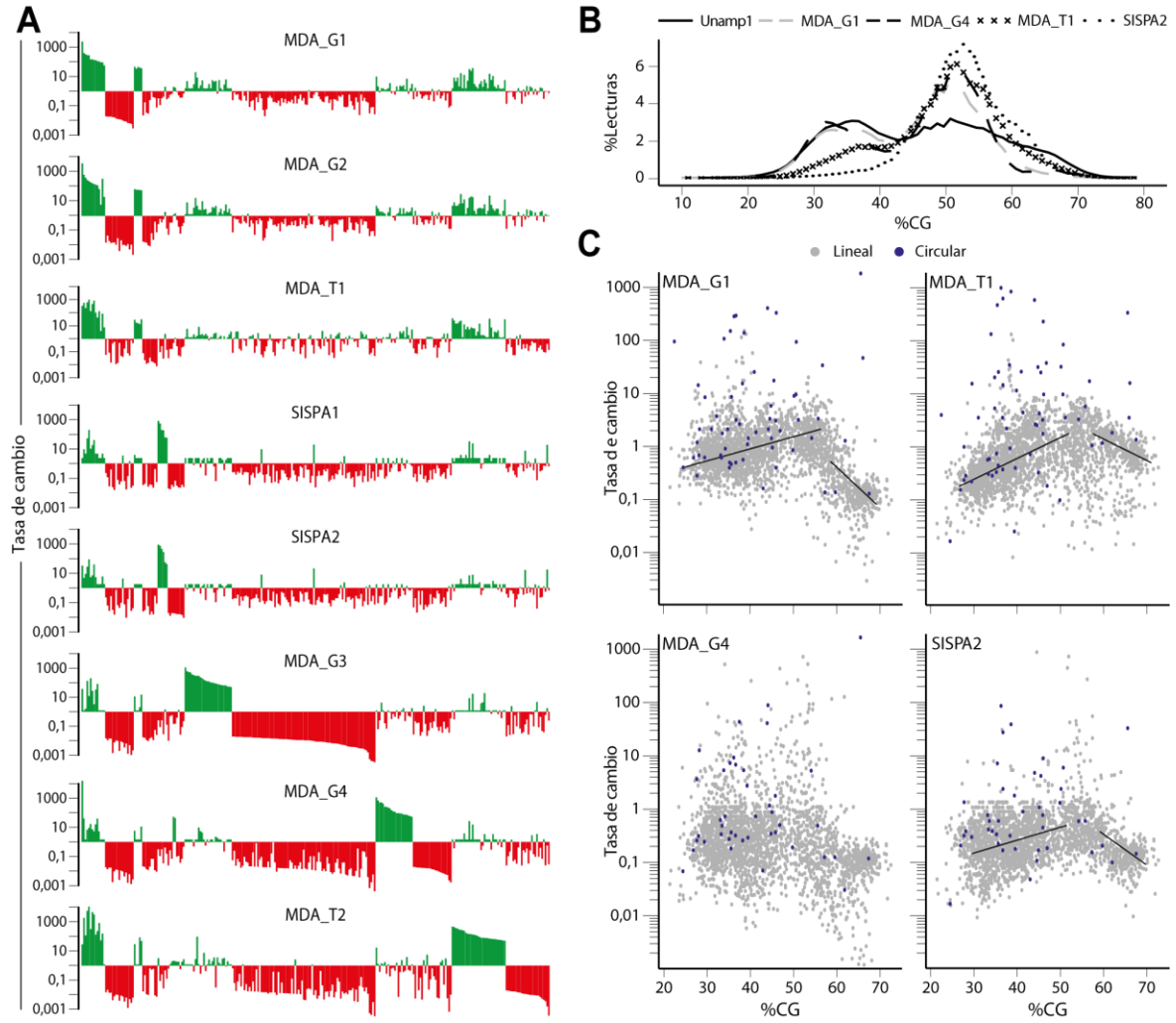

Figura 10. Impacto de los sesgos inducidos por la amplificación al azar en lecturas y *contigs* de viromas de saliva. (A) Tasa de cambio de la abundancia normalizada de los *cross-contigs* (*RPKM*) entre los viromas amplificados y el viroma sin amplificar. Solo aquellos *contigs* > 2 kpb con una tasa de cambio >50x (color verde) o < 0,02x (color rojo) están representados. Se llevaron a cabo cuatro amplificaciones usando el kit GenomiPhi™ con dos cantidades diferentes de ADN molde y tiempos de extensión: 1 ng 2,5 h (MDA_G1); 1 ng 10 h (MDA_G2); 10 pg 3,5 h (MDA_G3); y 10 pg 10 h (MDA_G4). Las amplificaciones con el kit TruePrime™ se llevaron a cabo desde 1 ng durante 2,5 h (MDA_T1) y 10 pg 3,5 h (MDA_T2). En las amplificaciones *SISPA* se usó un solo oligonucleótido (FR26RV-12N; SISPA1) o la mezcla de tres oligonucleótidos diferentes (FR26RV-12N, K-12N, y 454-A-12N; SISPA2). (B) Abundancia relativa de las lecturas en función de su contenido medio de CG para el viroma sin amplificar y algunos viromas amplificados al azar. (C) Tasa de cambio de los 2.577 *cross-contigs* en función de su contenido medio de CG. Los *cross-contigs* pequeños (<12 kpb) y circulares se representan como círculos azules y los *cross-contigs* lineales (o >12 kpb) como puntos grises. Las líneas de tendencia representan la regresión lineal en dos rangos de CG: 30-55 y 55-70%.

El estudio del contenido medio en CGs de las lecturas y *contigs* de cada viroma nos permitió detectar un sesgo en todos viromas amplificados relacionado con regiones de bajo y alto contenido en CGs. Así las lecturas de Unamp1 mostraron dos picos de abundancia a 36 y 51% de CGs, mientras que las lecturas procedentes de los viromas amplificados presentaban un mayor número de lecturas acumuladas en el

segundo pico (**Fig. 10B**). Por el contrario, los protocolos de amplificación con GenomiPhi™ presentaban una menor proporción de lecturas con porcentajes medios de CGs por encima del 60%, mientras que en la amplificación con MDA_T1 presentaba una menor proporción de lecturas con %CGs inferiores al 40%. Este último sesgo negativo estaba acentuado en los viromas amplificados por *SISPA*. Los 2.557 *contigs* originados durante el ensamblaje cruzado reprodujeron a la perfección los patrones de sesgo descritos para las lecturas (**Fig. 10C**). En este sentido, un análisis de regresión lineal de la abundancia relativa de los *contigs* no circulares de los viromas amplificados con respecto al obtenido sin amplificación en función del contenido de CG medio (en el rango de 30-65%) mostró pendientes positivas más pronunciadas en los viromas MDA_T1 y SISPA2 que en MDA_G1. También, de acuerdo con lo observado para las lecturas, observamos el efecto contrario en el rango de CGs de 60-70% (**Fig. 10B**).

Los viromas MDA_T1 y MDA_G1 mostraron el menor número de *contigs* con sesgos de ± 10 veces en relación a Unamp1 (6,2 y 7,6% respectivamente). Estos porcentajes de *contigs* altamente sesgados cayeron a 4,5 y 6% respectivamente cuando solo se analizaron los *contigs* >12 kpb o no circulares en el rango de CGs de 35-65%. La proporción de *contigs* con alto sesgo en los viromas amplificados con *SISPA* fue algo mayor (10,75-16,11%) y no se vio afectada por la eliminación de *contigs* circulares pequeños (10,82-16,30%), pero mostró una reducción similar a la de los viromas de *MDA* cuando no se tuvo en cuenta aquellos *contigs* con un contenido extremo de CGs (7,91-12,48%). Las muestras amplificadas por *MDA* desde 10 pg de molde mostraron alrededor de un 30% de los *contigs* con sesgos de ± 10 veces, lo que está de acuerdo con los sesgos estocásticos propuestos anteriormente. Además, este número no se redujo apenas tras la eliminación de *contigs* circulares pequeños o con contenido extremo de CGs.

1.2.3. La cobertura de los *contigs* en los viromas obtenidos mediante *MDA* es más uniforme que la obtenida mediante *SISPA*

La uniformidad de la cobertura de los *contigs* se analizó para los 38 *contigs* con coberturas medias >50x tanto en el viroma sin amplificar (Unamp1), como en los viromas amplificados a partir de 1 ng del mismo ADN molde (**Fig. 11**). Como se ejemplifica para los *contigs* 16 y 624, *MDA* proporciona una distribución de lecturas más uniforme que *SISPA*, pero peor que la que se obtiene en el viroma sin amplificar. Estos resultados se observan mejor dibujando curvas de Lorenz que representan la proporción de lecturas acumuladas a lo largo de la longitud de cada *contig* (**Fig. 11B**). Las curvas que mostraban menos diferencia con la distribución teórica perfecta fueron las correspondientes a los viromas sin amplificar, seguidas de los viromas de *MDA* y *SISPA*, en este orden. Para cuantificar la uniformidad de la cobertura sobre un número representativo de *contigs*, calculamos el coeficiente de variación y los índices de correlación de Pearson de estos 38 *contigs* (**Fig. 11C**). Los coeficientes de variación más altos correspondían al viroma SISPA2, con un valor promedio por encima de uno, y con

diferencias estadísticamente significativas respecto a otros viromas (*p-value* < 4,9x10^{-12}; Mann-Whitney prueba de dos colas). Las diferencias entre los *contigs* de viromas amplificados por *MDA* y no amplificados también fueron estadísticamente significativas (*p-value* = 0,002 para MDA_G1 y *p-value* = 0,0009 para MDA_T1), pero sus coeficientes medios de variación fueron inferiores a 0,5 en ambos casos. Además, los valores de correlación de Pearson entre los perfiles de cobertura de *contigs* amplificados y no amplificados (**Fig. 11D**) fueron más bajos para SISPA2 que para los viromas *MDA*, mostrando diferencias estadísticamente significativas en las pruebas de Mann-Whitney de dos colas (*p-value* < 6,4x10^{-13}). Estos resultados demuestran un mejor rendimiento de *MDA* sobre *SISPA* en términos de uniformidad en la cobertura del genoma.

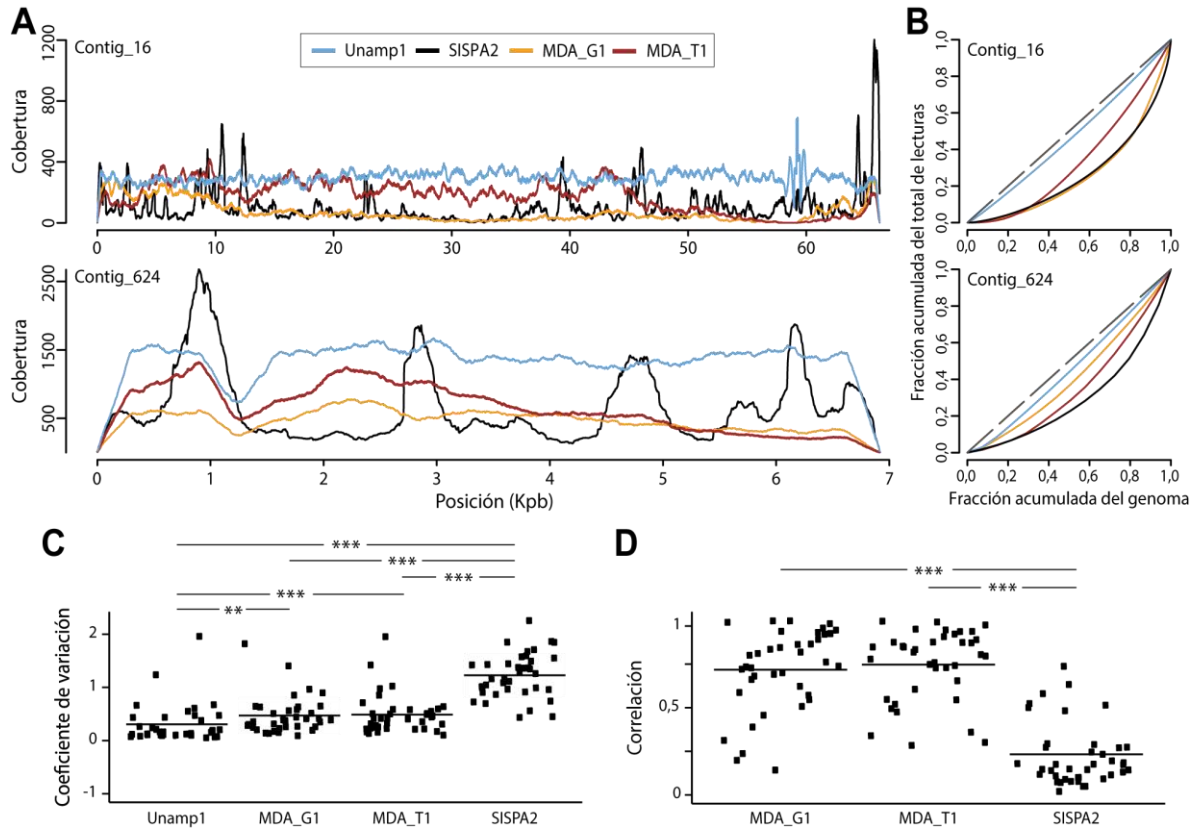

Figura 11. Cobertura de los *contigs* de viromas de saliva obtenidos mediante diferentes técnicas de amplificación al azar. (A) Perfiles de cobertura en dos de los *cross-contigs* más abundantes. (B) La homogeneidad de la distribución de las lecturas según la posición del *contig* se muestra mediante curvas de Lorenz. La línea discontinua representa una cobertura teórica perfecta. (C) Coeficientes de variación de cobertura para los 38 *cross-contigs* más abundantes que presentaban coberturas >50x en todos los viromas analizados. (D) Correlación de Pearson entre los perfiles de cobertura de los viromas sin amplificar y amplificados para el mismo grupo de *cross-contigs*. La línea horizontal en C y D representa el valor medio. * $p < 0,01$; ** $p < 0,005$; y *** $p < 0,001$.

1.2.4 El perfil irregular de cobertura de los *contigs* obtenidos mediante *SISPA* se debe en parte a picos de alta cobertura en las regiones con alta complejidad lingüística

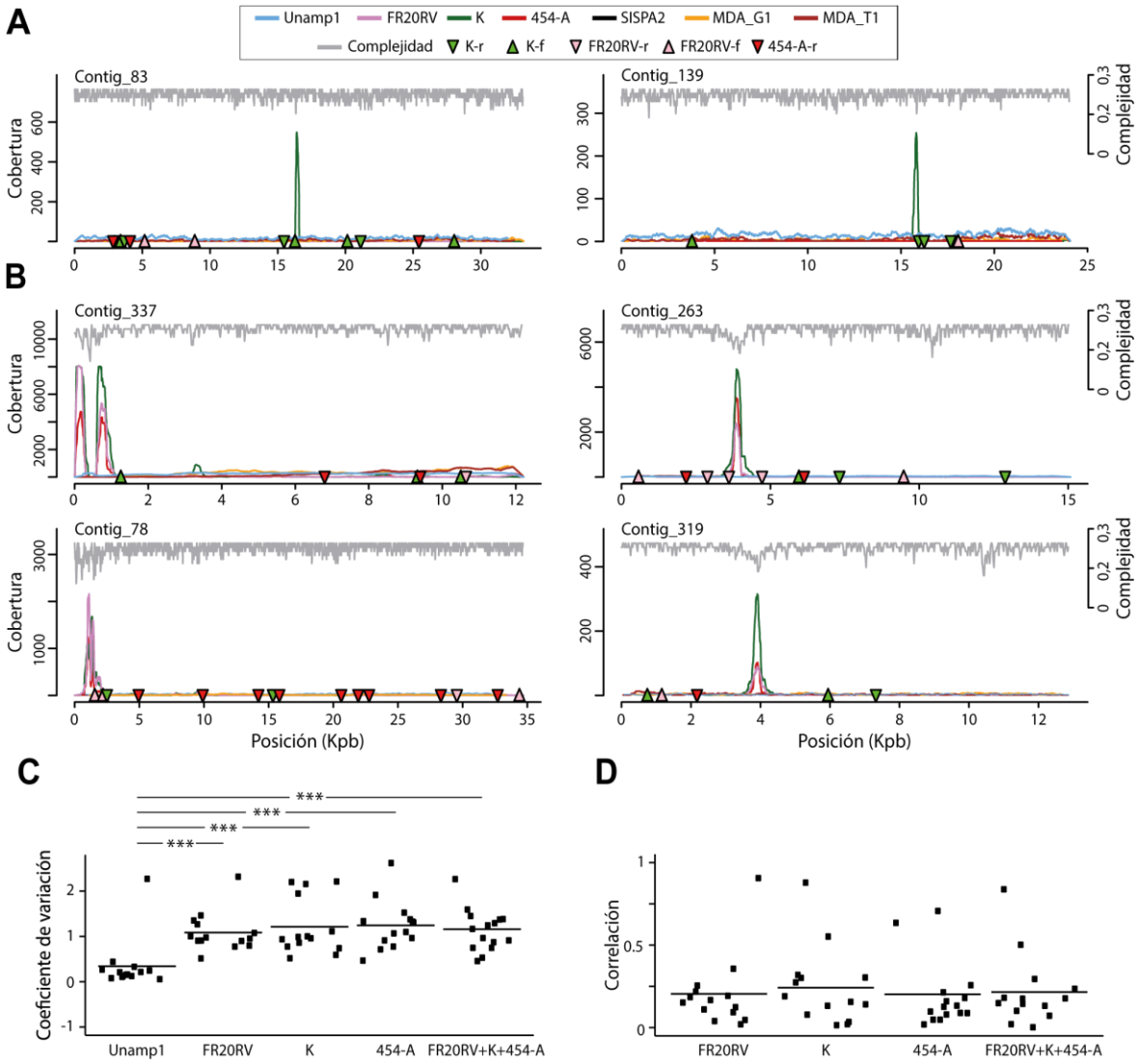

Figura 12. Perfil de cobertura en viromas de saliva obtenidos por *SISPA*. (A) Se muestran dos *cross-contigs* representativos con picos de alta cobertura flanqueados por secuencias con similitud a la región constante de los oligonucleótidos (triángulos coloreados) usados durante *SISPA*. (B) *Cross-contigs* representativos de viromas *SISPA* con una alta cobertura en regiones con baja complejidad lingüística de secuencia (línea gris). (C) Coeficientes de variación de cobertura para los 14 *cross-contigs* más abundantes que presentaban valores de cobertura media >50x entre los viromas analizados. (D) Correlación de Pearson entre los perfiles de cobertura de los viromas sin amplificar y los viromas amplificados por *SISPA* demultiplexados. La línea horizontal en C y D representa el valor medio. * $p < 0{,}01$; ** $p < 0{,}005$; y *** $p < 0{,}001$.

A continuación, quisimos estudiar la causa de la formación de picos de alta cobertura en los viromas de *SISPA*. Se había descrito previamente que estos picos pueden deberse a la hibridación preferente de la región constante 5' del oligonucleótido empleado, apuntando como solución la combinación de varios oligonucleótidos (Rosseel et al., 2013). En nuestros viromas encontramos que sólo un 20% de los picos de alta cobertura eran específicos de un oligonucleótido en particular, y estaban flanqueados por

secuencias con elevada identidad de secuencia con la región conservada del oligonucleótido empleado (**Fig. 12A**). Además, no encontramos diferencias estadísticamente significativas ni en los coeficientes de variación ni en los índices de correlación para la cobertura obtenida a partir de un único oligonucleótido (SISPA1) o la combinación de tres (SISPA2; **Fig. 12C,D**). Esto es coherente con el hecho de que muchos de estos picos estuvieran formados por lecturas obtenidas usando los tres oligonucleótidos. También observamos que en torno al 30% de los picos de alta cobertura no eran específicos de oligonucleótido y se presentaban en regiones de baja complejidad de secuencia como se ejemplifica en la **Figura 12B**. Estos resultados indican que el sesgo en la cobertura que induce *SISPA* es el resultado de la convergencia de múltiples factores, incluyendo la unión preferencial de la parte constante del oligonucleótido y una amplificación preferente de secuencias de ADN con baja complejidad lingüística.

1.3. Los sesgos introducidos durante la amplificación al azar tienen un impacto mínimo en estudios de diversidad beta de viromas de saliva

La amplificación al azar altera la abundancia relativa de ciertos miembros de las comunidades virales sintéticas y naturales. Para evaluar el impacto de este sesgo a nivel de comparaciones entre comunidades completas de virus, calculamos las disimilitudes de Bray-Curtis entre viromas en función de la abundancia normalizada de los *contigs* compartidos generados mediante ensamblaje cruzado. Para ello, la abundancia se expresó en *RPKMs* (lecturas alineadas por kilobase y por millón de lecturas). Las disimilitudes resultantes se representaron en sistemas de ordenación *NMDS*. De acuerdo con los espectros de los *contigs* más sesgados (**Fig. 10A**), las gráficas de ordenación mostraron que los viromas obtenidos a partir de *MDA* desde 1 ng de molde estaban localizados más próximos al viroma sin amplificar que los amplificados a partir de 10 pg (**Fig. 13A**). Además, se observaron correlaciones de Pearson de 0,55-0,65 entre los perfiles de *contigs* de Unamp1 y los amplificados a partir de 1 ng, incluidos los viromas *SISPA*, mientras que las correlaciones con viromas amplificados a partir de 10 pg variaron en el rango 0,24-0,49 (**Tabla 3**). Un resultado similar se obtuvo cuando en lugar de disimilitudes Bray-Curtis, se utilizaron índices de Sørensen (Sørensen, 1948), que tienen en cuenta sólo la presencia o ausencia de los *contigs*, y es por tanto una medida de disimilitud más sensible a variaciones en la detección de virus poco abundantes (**Fig. 13C**).

Es importante destacar que la inclusión de dos nuevos viromas de saliva (SaC25 y Sa33), procedentes de sujetos que no habían contribuido a la muestra Unamp1 en un segundo ensamblaje cruzado, dio lugar a una superposición perfecta de Unamp1 y todos los viromas obtenidos por amplificación al azar de esta misma muestra (**Fig. 13B**). Por el contrario, los dos nuevos viromas no relacionados mostraron una gran separación entre ellos, con valores de disimilitud de Bray-Curtis por encima de 0,98 e índices de Sørensen por encima de 0,65, y una ausencia completa de correlación de Pearson (**Tabla 3**) que refleja la singularidad de los viromas en saliva humana.

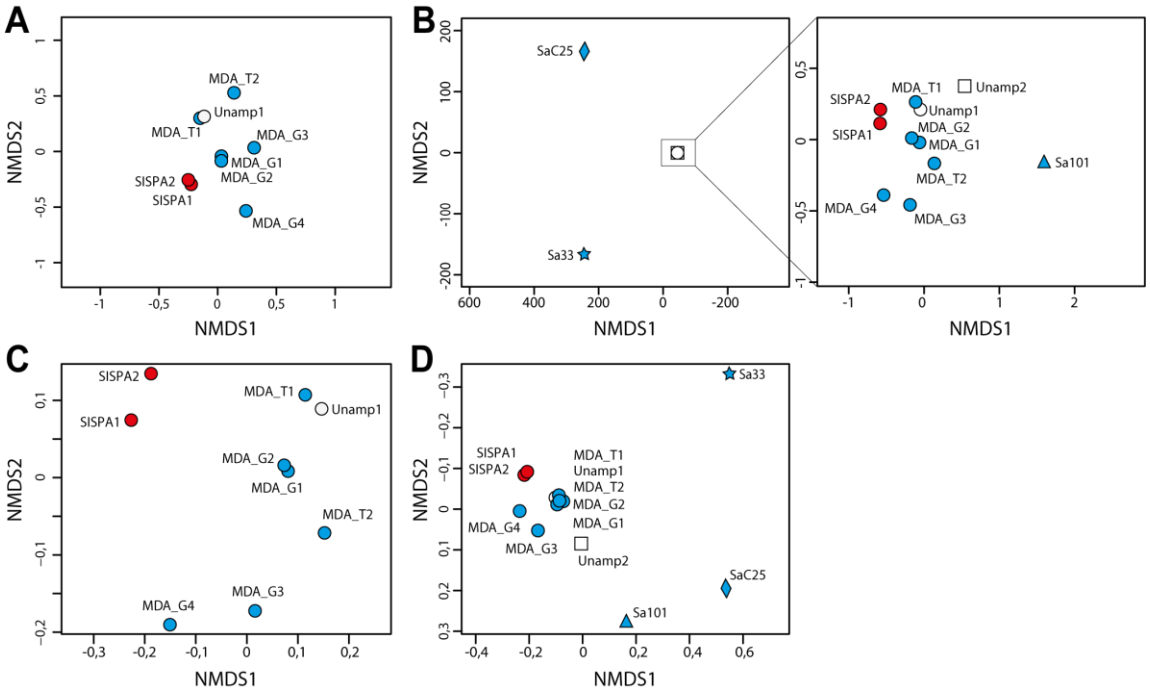

Figura 13. Sistemas de ordenación de viromas basados en los perfiles de abundancia de *cross-contigs*. Las abundancias normalizadas de los *cross-contigs* (*RPKM*) se usaron para calcular las disimilitudes de Bray-Curtis (A y B) y los índices de Sørensen (C y D) entre los viromas. (A y C) La matriz de disimilitudes entre el viroma sin amplificar (Unamp1) y ocho viromas derivados de este pero obtenidos por amplificación al azar (MDA_G1-4, MDA_T1-2 y SISPA1-2) se representaron siguiendo el sistema de ordenación *NMDS*. (B y D) Las matrices de disimilitud de los *cross-contigs* obtenidos con los anteriores nueve viromas, junto con dos viromas adicionales de saliva parcialmente relacionados (Unamp2 y Sa101), y dos viromas no relacionados procedentes de individuos que no contribuyeron a ninguno de las dos mezclas de saliva evaluadas (SaC25 y Sa33) también se representaron por *NMDS*. La distribución *NMDS* a la derecha del panel (B) representa la disimilitud entre todos los viromas, excluyendo SaC25 y Sa33 (nótense las diferencias de magnitud de los ejes). El tipo de símbolo indica el origen distinto de los viromas. Los colores blanco, azul y rojo indican viromas obtenidos sin amplificación, o amplificados al azar mediante *MDA* o *SISPA* respectivamente.

	Unamp1	MDA_G1	MDA_G2	MDA_G3	MDA_G4	MDA_T1	MDA_T2	SISPA1	SISPA2	Unamp2	Sa101	SaC25
MDA_G1	0,65											
MDA_G2	0,64	0,98										
MDA_G3	0,31	0,76	0,7									
MDA_G4	0,24	0,52	0,57	0,2								
MDA_T1	0,59	0,76	0,76	0,61	0,31							
MDA_T2	0,49	0,41	0,36	0,2	0,19	0,28						
SISPA1	0,55	0,42	0,41	0,16	0,1	0,48	0,12					
SISPA2	0,55	0,44	0,44	0,21	0,11	0,53	0,12	0,96				
Unamp2	0,49	0,22	0,25	0,07	0,04	0,27	0,03	0,32	0,33			
Sa101	0,33	0,28	0,28	0,1	0,06	0,4	0,06	0,49	0,45	0,21		
SaC25	-0,05	-0,04	-0,04	-0,02	-0,02	-0,03	-0,03	-0,02	-0,02	-0,02	-0,02	
Sa33	-0,04	-0,03	-0,03	-0,01	-0,02	-0,02	-0,03	-0,01	-0,02	-0,02	-0,02	-0,02

Tabla 3. Correlación de Pearson entre los perfiles de abundancia normalizados de los *cross-contigs* obtenidos de los nueve viromas derivados de la misma muestra de saliva, de dos viromas parcialmente relacionados y otros dos viromas no relacionados también de saliva.

Por otra parte, el *cluster* formado por Unamp1 y los viromas amplificados desde la misma muestra también incluía dos viromas adicionales: uno de un individuo que había sido un donante también para Unamp1 (Sa101), y otro obtenido de una mezcla de salivas de siete individuos (Unamp2), seis de los cuales también habían contribuido a Unamp1 (**Fig. 13B,D**). Este resultado sugiere que la casi ausencia de *contigs* compartidos entre viromas de saliva de individuos no relacionados es suficiente para forzar la agrupación de viromas relacionados, aunque solo compartan unos pocos virus. Finalmente, modelamos la distribución de lecturas compartidas entre viromas de individuos no relacionados (MDA1_G1, SaC25 y Sa33) mediante comparaciones por *BLASTn* de submuestras generadas al azar de 10.000 lecturas en cada caso), como una medida alternativa de distancia entre viromas. Esta distribución mostró una media de 1.391,49 ± 86,28 *SD* lecturas compartidas, que fue significativamente más baja (test de Mann-Whitney con *p-value* < $2{,}2 \times 10^{-16}$) que las compartidas entre viromas procedentes de la misma muestra (Unamp1 y los dos viromas más sesgados: MDA_G4 y MDA_T2; valor medio de 4.790,81 ± 63,61 *SD*).

Estos resultados sugieren que el sesgo introducido por cualquiera de los dos métodos de amplificación al azar empleados en esta tesis, incluso desde picogramos de ADN molde, tienen una incidencia mínima en los estudios de diversidad beta entre viromas de saliva de individuos no relacionados.

2. Estudio de las comunidades de virus en muestras bucales de individuos sanos y de pacientes con caries o estomatitis aftosa recurrente

Con el objetivo de tener una imagen más real y completa de la comunidad de virus de ADN de la boca humana y estudiar posibles cambios en su composición en situaciones de enfermedad, se tomaron, tras la firma de un consentimiento informado, un total de 59 muestras de la cavidad bucal de 55 individuos distintos con edades comprendidas entre los 20 y 42 años: 13 muestras de placa dental de individuos sin caries y 13 de individuos con caries activas, 16 muestras de mucosa bucal de individuos sanos y 13 de individuos con aftas activas. Además, se recogieron muestras de saliva de cuatro de estos donantes para realizar comparaciones entre viromas del mismo individuo.

2.1. Procesamiento y selección de muestras de la cavidad bucal para su estudio metagenómico

Teniendo en cuenta los resultados obtenidos en el apartado anterior elegimos un protocolo sencillo de enriquecimiento de partículas virales basado en una centrifugación a baja velocidad, filtración en 0,45μm, degradación de ADN libre con nucleasas y extracción del material genético viral con fenol/cloroformo (**Fig. 5**). Decidimos no incluir en este protocolo el paso por colchón de iodixanol debido a que el volumen manejado no exigía concentrar las muestras. Se obtuvieron entre 150 pg y 32 ng de ADN de genomas virales y en general se recuperó más ADN desde las muestras de mucosa y saliva que de placa dental debido a la limitada cantidad de material de partida de estas últimas muestras. Aun así, la cantidad de todas ellas fue insuficiente para preparar de librerías de secuenciación masiva.

Por ello, decidimos recurrir a la amplificación al azar mediante *MDA* que, como vimos en el apartado anterior, introduce sesgos estocásticos y sistemáticos pero sin incidencia en estudios de comparaciones entre viromas de individuos no relacionados. Como molde de estas reacciones se emplearon cantidades de genomas virales de ADN de 20 pg-5,6 ng, y se obtuvieron entre 1,33-63,2 μg de ADN amplificado. De las 59 muestras procesadas, tres no se lograron amplificar en *MDA* por encima de los niveles de amplificación observados en el control negativo.

Debido a que durante el proceso de purificación empleamos filtros de 0,45μm que podrían permitir el paso de bacterias de tamaño pequeño, estudiamos el grado de contaminación bacteriana mediante *PCR* semicuantitativa del gen que codifica por ARNr 16S. Establecimos como criterio para descartar la secuenciación masiva de una muestra que ésta tuviera un nivel de contaminación igual o superior al de una muestra de virus previamente secuenciada en nuestro laboratorio con aproximadamente un 15% de lecturas de origen bacteriano (**Fig. 14**).

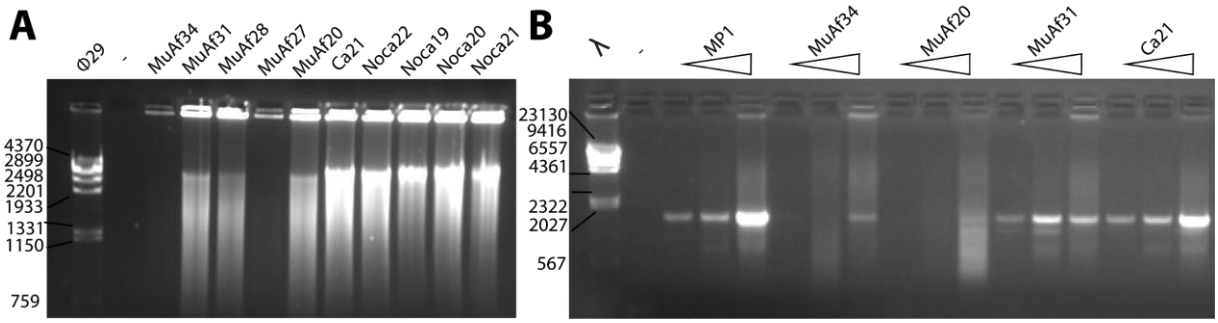

Figura 14. Control de la amplificación por *MDA* y determinación del grado de contaminación bacteriana. (A) Resultados representativos de la amplificación por *MDA* del ADN viral extraído de las muestras bucales. Una veinteava parte del producto de amplificación se cargó en un gel de agarosa al 0,8%. (B) Evaluación del grado de contaminación bacteriana mediante *PCR* semicuantitativa del gen del ARNr 16S en el ADN viral extraído y amplificado. La muestra de referencia MP1 contiene alrededor de un 15% de lecturas de origen bacteriano. Una quinta parte de los productos de *PCR* obtenidos desde 1 ng, 10 ng y 100 ng de molde estimado por *Picogreen* se cargaron en un gel de agarosa al 1%. Los marcadores se corresponden a los genomas de los bacteriófagos Φ29 y λ digeridos con HindIII.

En general detectamos un mayor grado de contaminación bacteriana en las muestras procedentes de placa dental que en las de mucosa (**Tabla 4**). Identificamos como muestras contaminadas con ADN bacteriano ocho muestras de placa de individuos sanos, dos de placa de individuos con caries, dos de mucosa de individuos sanos, dos de mucosa de individuos con afta y una de saliva. Estas muestras fueron descartadas para secuenciación masiva, con la excepción de tres procedentes de individuos sin caries (Noca1, Noca14 y Noca16) debido al pequeño número de muestras disponibles sin contaminación y a que éstas presentaban niveles de contaminación próximos al 15%. Así, las muestras finalmente secuenciadas fueron 38 (**Tabla 4**).

Tipo	Muestras	Amplificadas (*MDA*)	*PCR**	Secuenciadas (*Miseq*)	ARNr 16S$	Viromas estudiados#
Placa sana	13	13	5	8	5	5
Placa caries	13	13	11	9	7	7
Mucosa sana	16	15	13	10	9	9
Mucosa afta	13	11	9	8	8	6
Saliva	4	4	3	3	3	3
Total	59	56	41	38	32	30

Tabla 4. Número de muestras procesadas en cada etapa del protocolo. *PCR**: Muestras con baja contaminación bacteriana estimada por *PCR* del gen que codifica por ARNr 16S. ARNr 16S$: Viromas con baja contaminación bacteriana estimada por la presencia en el viroma de secuencias del gen para ARNr 16S. Viromas estudiados#: Viromas con un alto porcentaje de secuencias ensambladas en *contigs* largos.

2.2. Secuenciación masiva y preprocesado de las secuencias obtenidas

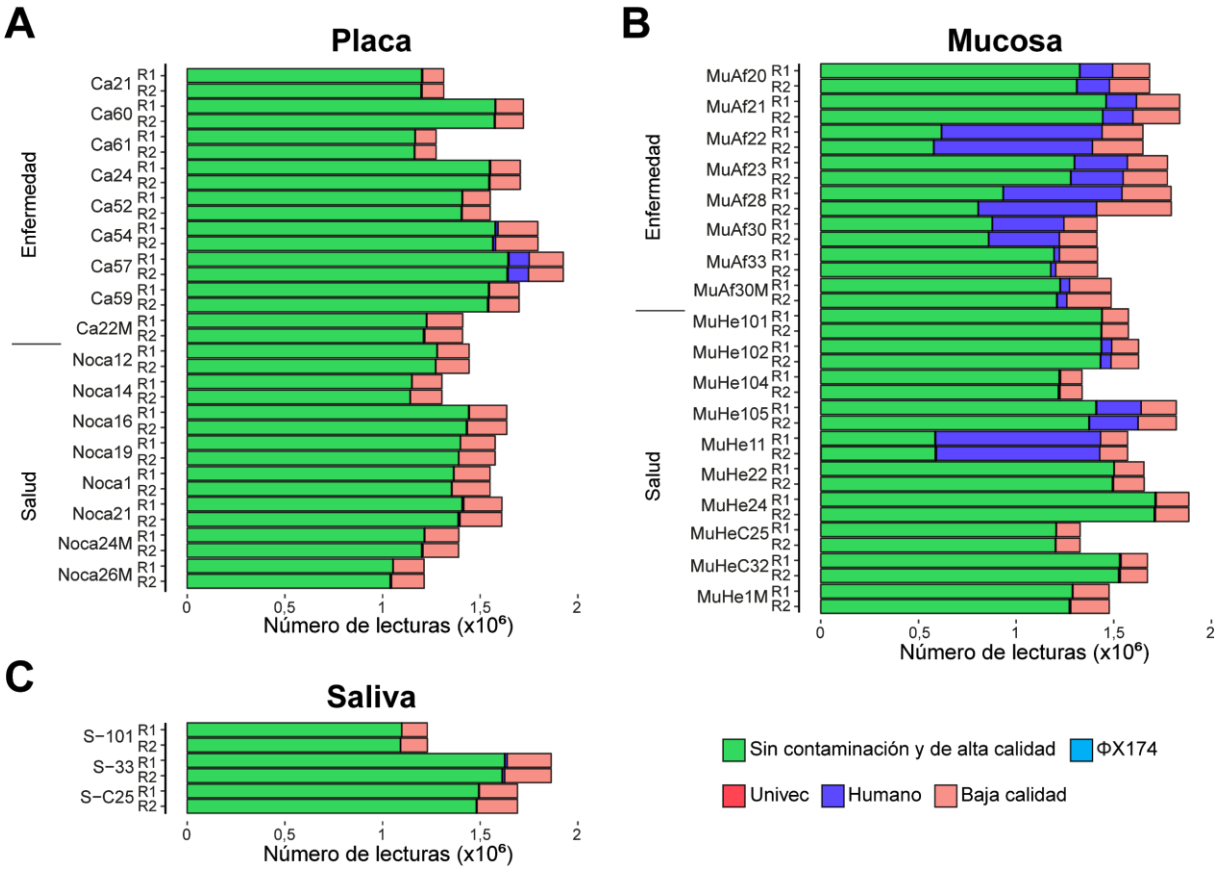

Figura 15. Filtración de secuencias por calidad y eliminación de secuencias contaminantes de origen conocido. Se muestra el porcentaje de secuencias de baja calidad determinadas con el programa *Prinseq* y las secuencias contaminantes identificadas mediante alineamiento con *Bowtie2* contra el genoma humano, el genoma del bacteriófago *ΦX174* y la base de datos de plásmidos *Univec*.

La secuenciación mediante la tecnología *MiSeq* de Illumina® nos permitió obtener una media de 1.576.263 lecturas por viroma de las que un 8,09-21,3% se eliminaron por tener baja calidad (**Fig. 15**). También se eliminaron aquellas secuencias con una elevada similitud con el genoma humano, con una base de datos de plásmidos (*Univec*) y con el genoma del bacteriófago *ΦX174* que se usa como control interno de la secuenciación de Illumina®. El nivel de contaminación con secuencias de origen humano

fue mayor en términos generales en los viromas de mucosa y tres de ellos presentaban >33,72% de sus secuencias de origen humano. El porcentaje de secuencias de *ΦX174* y de plásmidos conocidos no superó el 1,84% en ninguno de los viromas estudiados. El promedio final de secuencias de alta calidad y libres de contaminación conocida fue de 1.299.533 y la mediana de las secuencias descartadas en los pasos de preprocesado fue de un 12,44%.

2.3. Asignación taxonómica de las lecturas de alta calidad

Para tener una imagen preliminar de la composición de los viromas de la cavidad bucal humana realizamos alineamientos contra la base de proteínas *nr* del *GenBank* (**Fig. 16A**). Una parte importante de las secuencias no presentaba similitud con ninguna secuencia de la base de datos (32,98-72,57%) y el dominio mejor representado fue el de bacteria (19,59-60,64%). Los viromas de mucosa presentaron un mayor porcentaje de lecturas relacionadas con eucariotas (en su mayoría procedentes del genoma humano) y en alguno de ellos, como MuAf28, estas secuencias alcanzaban el 31,02% de las lecturas. Finalmente, el porcentaje de secuencias asignadas a virus fue de un 2,28-29,36% (mediana en 10,58%), lo que coincide con resultados previos de viromas humanos (Minot et al., 2011) o de ecosistemas naturales (López-Bueno et al., 2009).

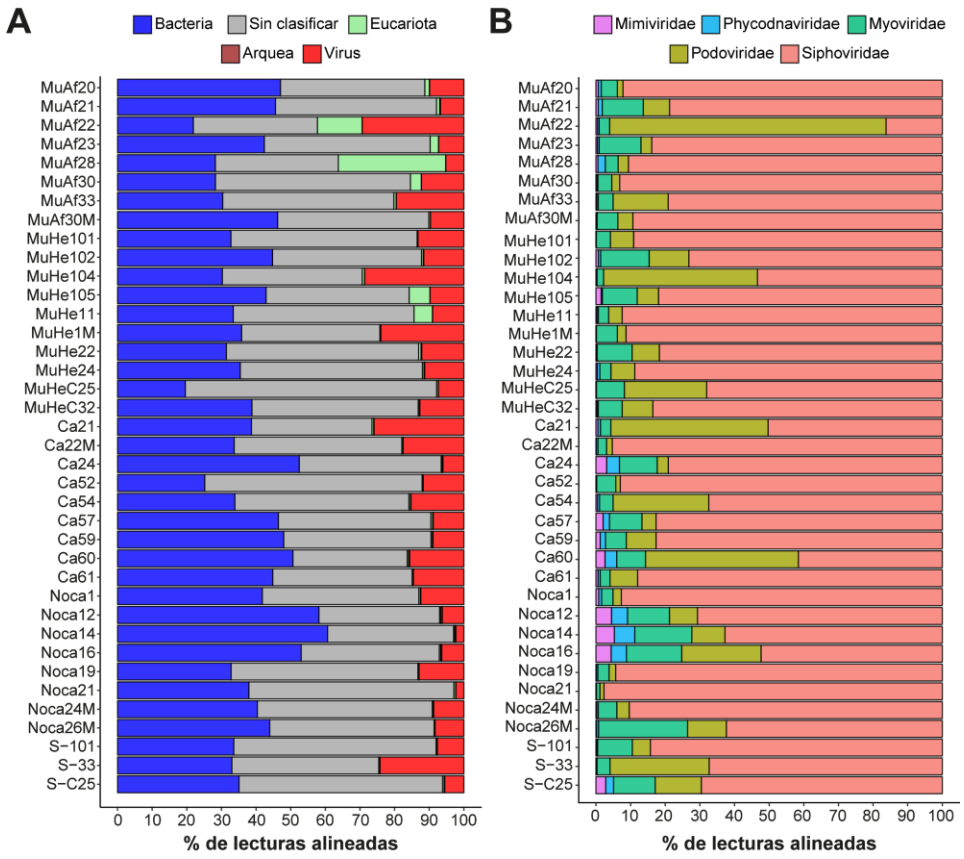

Figura 16. Clasificación taxonómica de los viromas de la cavidad bucal. (A) Proporciones relativas de secuencias asignadas a cada dominio mediante comparación (*BLASTx e-value* < $1x10^{-03}$) con la base de datos de proteínas *nr*. (B) Proporciones relativas de las secuencias asignadas a familias virales más abundantes mediante comparación (*BLASTx e-value* < $1x10^{-03}$) con la base de datos de proteínas virales *PHAST*.

Una segunda comparación de los viromas con una base de datos de proteínas exclusivamente virales (**Fig. 16B**) permitió ver un predominio de secuencias asignadas a bacteriófagos del orden *Caudovirales*, siendo los virus de la familia *Siphoviridae* los más abundantes con una mediana del 81,76%, seguido por *Podoviridae* (7,13%) y *Myoviridae* (5,73%). También observamos asignación de secuencias a virus grandes eucarióticos como *Phycodnaviridae* (0,49%) o *Mimiviridae* (0,38%) aunque en general con peores *e-values* que las asignaciones a *Caudovirales*.

2.4. La mayoría de las secuencias de alta calidad de los viromas se ensamblan en *contigs* virales de gran tamaño y cobertura

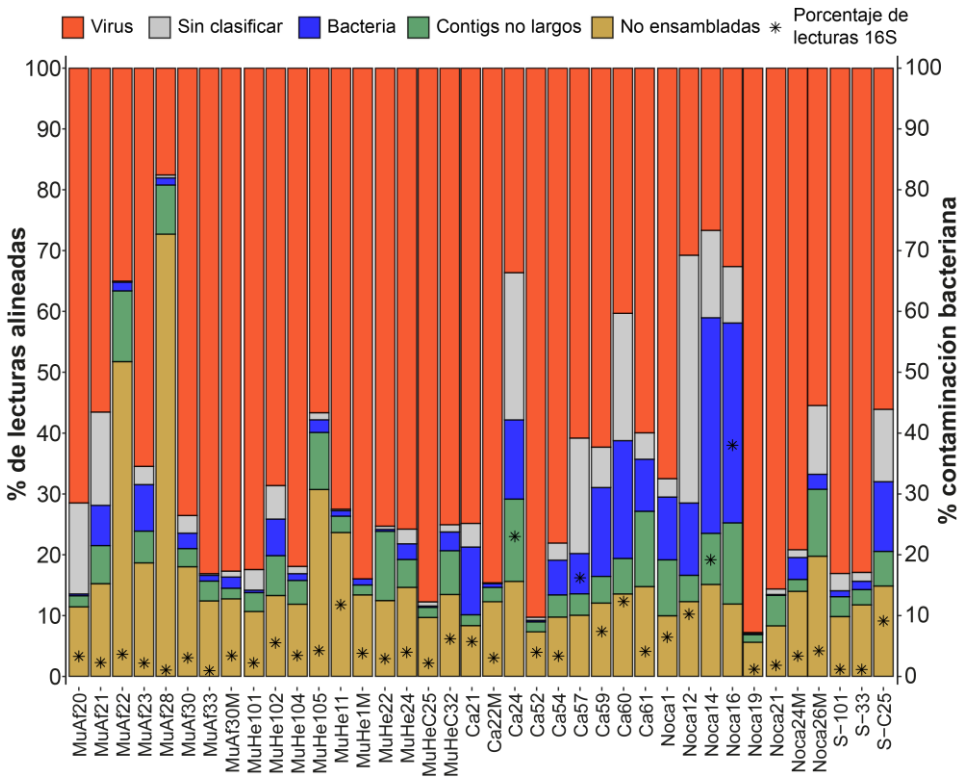

Figura 17. Composición de los viromas en función del porcentaje de lecturas alineadas en *contigs* asignados o no a los distintos dominios. Se muestra el porcentaje de secuencias alineadas mediante *Bowtie2* a *contigs* largos clasificados como virales, bacterianos o sin clasificar, así como el % de secuencias no ensambladas o ensambladas en *contigs* pequeños o de baja cobertura. Los asteriscos muestran el porcentaje de contaminación bacteriana estimado en función del número de lecturas con similitud por el gen que codifica por ARNr16S (*BLASTn* contra *SILVA*, *e-value* < $1x10^{-10}$). Se consideró como 100% de contaminación bacteriana (eje secundario) el valor medio del porcentaje de secuencias del gen 16S encontrado en seis microbiomas de placa dental (0,614%; metagenomas depositados en la base de datos pública *MG-RAST* bajo los números de acceso: 4447192.3, 4447102.3, 4447103.3, 4447101.3, 4447943.3, 4447903.3, 4447971.3 y 4447970.3 (Belda-Ferre et al., 2012))

Las lecturas de alta calidad apareadas y huérfanas de cada metagenoma se ensamblaron *de novo* por separado. Esto generó entre 812 y 8.730 *contigs* mayores de 500 pb por viroma, con longitudes medias de 931-3.794 pb y un N50 de 883-16.825 pb**.** Con el objetivo de reducir la contaminación de *contigs* bacterianos decidimos aplicar un filtro adicional de eliminación de *contigs* de baja cobertura y pequeño

tamaño (< 3 kpb y cobertura < 15x ó < 10 kpb y cobertura < 4x), ya que, debido al mayor tamaño de los genomas bacterianos, la cobertura de sus *contigs* debería ser baja y tener un alto grado de fragmentación. Los 6.544 *contigs* de gran tamaño o cobertura resultantes (de aquí en adelante "*contigs* largos"), pese a representar sólo un 5,37% del total de *contigs*, estaban ensamblados por un 79,96% de las secuencias de todos los viromas. A continuación, estos *contigs* se clasificaron en tres grupos definidos: Virales, Bacterianos y Sin clasificar en función del porcentaje de pautas de lectura abiertas que presentaban similitud de secuencia con proteínas virales o bacterianas (**sección 2.2.6.1. de Materiales y Métodos**). Las medianas de los porcentajes de *contigs* asignados a virus, bacterias o no clasificados fueron respectivamente 42,71%, 42,57% y 14,64%. Sin embargo, para tener una imagen más real de la abundancia de genomas virales en nuestros viromas calculamos la cantidad de lecturas que alinean contra estos *contigs* mediante *Bowtie2*. Entre un 17,57-92,78% (mediana del 73,04%) de las lecturas de los viromas estaban contenidas en *contigs* largos categorizados como virales (**Fig. 17**), frente a un 0,11-35,45% (mediana de 2,49%) de lecturas en *contigs* bacterianos y un 0,016-40,72% (mediana del 2,83%) en *contigs* sin clasificar. El promedio de la longitud de los *contigs* largos asignados a virus era 22.119 pb frente a 2.172 pb de los *contigs* bacterianos y 16.966 pb de los no clasificados. Estos resultados sugieren que la contaminación bacteriana es baja y estas secuencias se ensamblan en *contigs* de pequeño tamaño. Sin embargo, en los viromas con mayor contaminación hemos podido identificar varios *contigs* de gran tamaño relacionados con bacterias del filo *TM7*, incluyendo un genoma posiblemente completo de 734 kpb. A continuación, estimamos el grado de contaminación bacteriana en función del contenido en secuencias del gen que codifica por ARNr 16S, teniendo en cuenta que la proporción media de secuencias de este gen en seis microbiomas de placa dental es del 0,614% (Belda-Ferre et al., 2012). Como cabía esperar, los viromas con una contaminación mayor de genomas bacterianos eran las que presentaban mayor porcentaje de secuencias en *contig* clasificados como bacterianos o sin clasificar. De hecho, ambos parámetros mostraron una buena correlación (R^2 = 0,82). Teniendo en cuenta ambas estrategias de estimación de la contaminación bacteriana y el porcentaje de secuencias en *contigs* virales en cada viroma decidimos descartar para posteriores análisis de esta tesis doctoral los siguientes seis viromas: MuHe11, Caries24, Caries60, Noca12, Noca14 y Noca16. Además, descartamos también otros dos viromas de mucosa con aftas (MuAf22 y MuAf28) ya que mostraban una alta proporción de secuencias de ADN humano y un número muy elevado de secuencias no ensambladas o ensambladas en *contigs* de tamaño pequeño y baja complejidad (dato no mostrado).

Los 30 viromas restantes (**Tabla 4**) estaban formados por una media de 61 *contigs* virales largos (7-149 *contigs*) ensamblados por un 74,33% de las lecturas originales (valor medio) de cada viroma. Además, el hecho de no detectar ni una sola secuencia relacionada con el gen que codifica por ARNr 16S en estos *contigs* avala la fiabilidad de la metodología empleada para su asignación al dominio de virus.

Finalmente quisimos comparar nuestro sistema de identificación de *contigs* virales con la herramienta *VirSorter* (Roux et al., 2015a), la cual detecta bacteriófagos libres y profagos en metagenomas

ensamblados en *contigs*. Observamos que ambos métodos coincidían en identificar 1.031 *contigs* como virales (con un 62,97% de las lecturas de todos los viromas alineadas; **Fig. 18**). *VirSorter* identificó 181 *contigs* virales adicionales que eran poco abundantes (1,04% de lecturas alineadas) y que nuestro *script* no pudo identificar. Estos *contigs* en muchos casos podrían corresponderse con profagos ya que presentaban pautas abiertas de lecturas con similitud por proteínas relacionadas con virus (holinas, colas de bacteriófagos, etc), pero también con genes de genomas bacterianos. Por otro lado, nuestro método de identificación de *contigs* virales mostró una mayor sensibilidad al detectar 797 *contigs* virales no identificados por *VirSorter* (con un 11,36% de las secuencias de todos los viromas alineadas). Aunque no tenemos un sistema de validación de falsos positivos, ninguno de estos *contigs* contenía secuencias relacionadas con el gen que codifica por ARNr 16S y presentaban, en muchos casos, valores de cobertura muy altos, incompatibles con un posible origen bacteriano.

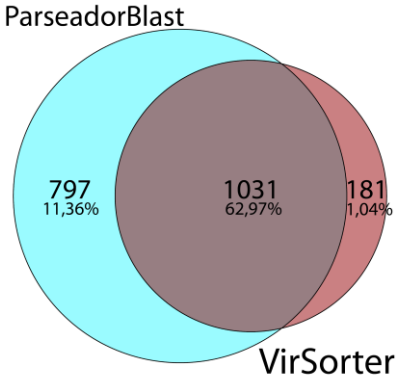

Figura 18. Comparación de dos métodos de identificación de *contigs* virales: *VirSorter* (Roux et al., 2015a) **y *ParseadorBLAST.pl v1.21* (Anexos - Script III).** Se muestran los 4.418 *contigs* largos obtenidos en los 30 viromas analizados identificados como virales por alguno de los dos métodos y entre paréntesis el porcentaje de lecturas alineadas en estos *contigs*.

2.5. Los viromas de mucosa bucal y placa dental están formados por cientos de virus distintos, dominados virus emparentados con bacteriófagos que no se han descrito previamente en este ambiente

Las comunidades de virus de la mucosa bucal y de la placa dental presentan una distribución de especies caracterizada en muchos casos por la presencia de unos pocos virus muy abundantes y una gran diversidad de virus poco numerosos. Así, 26 de los 27 viromas estudiados presentaban al menos un *contig* viral ensamblado con más de medio millón de lecturas (**Fig. 19**), y en el caso extremo de los viromas MuHe24, Noca21 y Ca52 el genoma del virus dominante estaba ensamblado por un 49,2%, 73,67% y un 73,79% de las lecturas de sus viromas respectivamente. La estructura poblacional de estos viromas con un virus dominante sugiere una baja diversidad alfa y contrasta con otros viromas como MuHeC32, MuHe104 o MuAf21, cuya estructura poblacional más uniforme sugiere una mayor diversidad alfa.

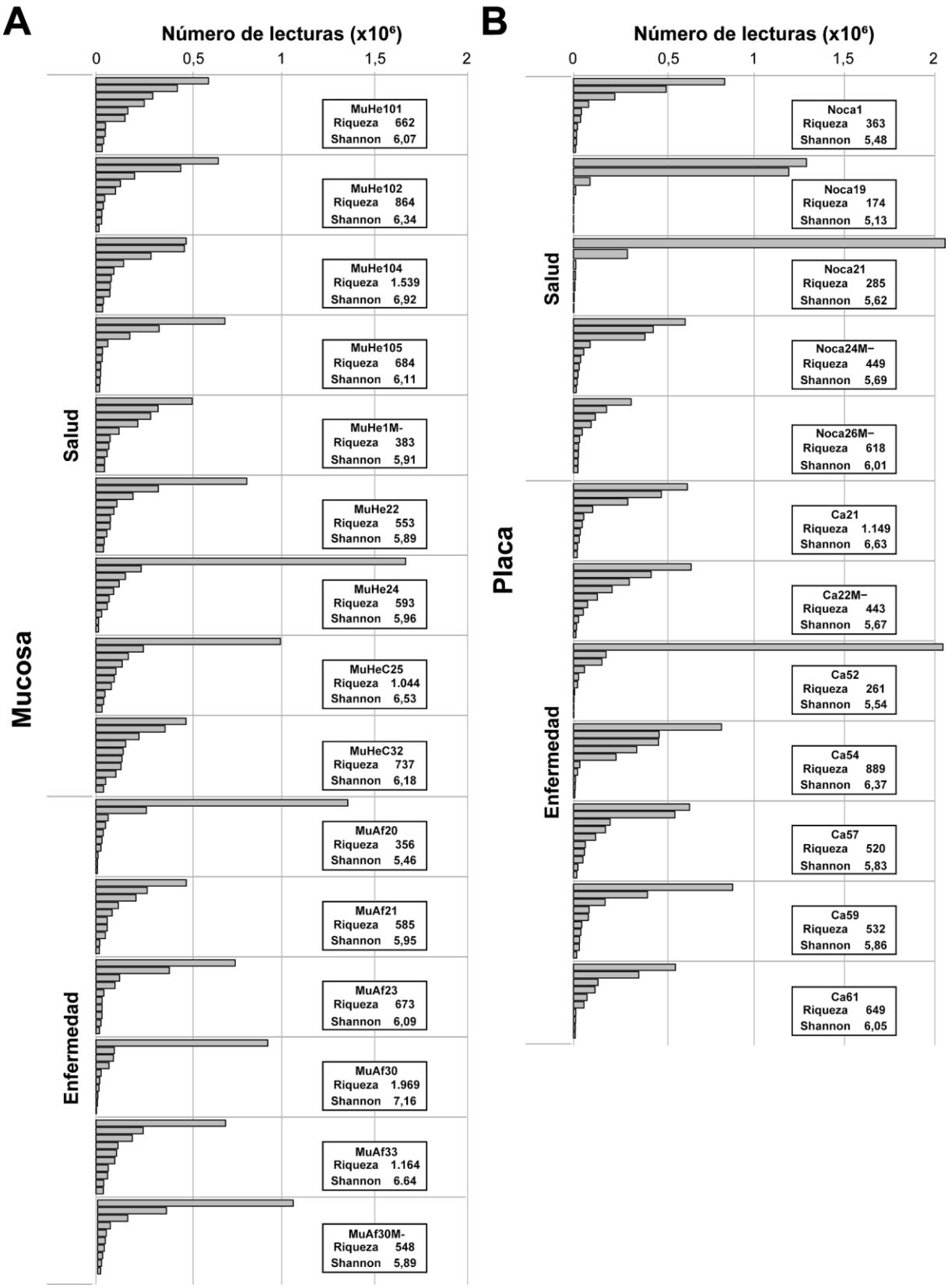

Figura 19. Curvas de rangos de abundancia (curvas *Whittaker*) de los viromas bucales. Se representan sólo los 10 *contigs* más abundantes de los viromas de mucosa (A y de placa dental (B)) expresados en función del número de lecturas R1+R2 alineadas con cada *contig*. Los valores de riqueza de especies e índice de Shannon se estimaron utilizando la herramienta *PHACCS* (Angly et al., 2005) y se muestran en el interior de cada gráfico.

Los índices de diversidad alfa de Shannon calculados con la herramienta *PHACCS* variaron entre 5,13 y 7,16 (mediana de 5,96) y la riqueza estimada de especies entre 174-1.969 (mediana 593). Como era de esperar, aquellos viromas con estructuras poblacionales más uniformes presentaban índices de Shannon y estimaciones de riqueza de especies mayores que los viromas con estructuras poblacionales menos uniformes. Por último, un análisis estadístico basado en un test no paramétrico de Mann-Whitney mostró que la mayor diversidad alfa de los viromas de la mucosa con respecto a los viromas de placa dental era estadísticamente significativa (*p-value* = 0,0204). Sin embargo, no detectamos diferencias significativas cuando se compararon condiciones de salud y enfermedad (Mucosa sana y afta: *p-value* = 0,723; Placa sana y placa caries *p-value* = 0,106).

La asignación taxonómica de los 1.557 *contigs* virales ensamblados desde los viromas de la cavidad se hizo considerando el resultado más frecuente de la comparación por *BLAST* contra la base de datos de proteínas *PHAST* de todas pautas de lectura abiertas de cada *contig.* Los *contigs* virales con mayor cobertura estaban relacionados con las especies virales *Arthrobacter phage Mudcat*, *Rhodococcus phage ReqiPoco6*, *Mycobacterium phage Muddy*, *Rhodococcus phage ReqiPepy6* y *Gordonia phage Emalyn*. En su conjunto, más de un 50% del total de lecturas de todos los viromas se alineaban con estos *contigs* virales, que dominaban en muchos casos las comunidades al estar entre los 10 más abundantes de cada viroma (**Tabla 5**).

Especies virales	**Lecturas totales**	***Contigs***	***Contigs* en *top* 10**
Arthrobacter phage Mudcat	11.657.302	146	27
Rhodococcus phage ReqiPoco6	5.194.984	76	16
Mycobacterium phage Muddy	4.300.551	13	7
Rhodococcus phage ReqiPepy6	3.955.964	66	13
Gordonia phage Emalyn	2.255.229	15	7
Gordonia phage Jswag	2.149.025	2	2
Actinomyces virus Av1	1.885.769	86	12
Microbacterium phage vB_MoxS-ISF9	1.730.763	15	8
Xanthomonas phage Xp15	1.249.733	11	5
Enterococcus phage phiFL4A	1.224.874	19	8

Tabla 5. Virus más representados en la cavidad bucal humana. Se muestra el número de *contigs* (y las lecturas alineadas con ellos) relacionadas por similitud de secuencia con las especies de virus de referencia más representadas en los viromas de la cavidad bucal.

2.6. Un número alto de *contigs* virales se corresponden con genomas completos o casi completos

La distribución de los *contigs* virales en función de su tamaño y abundancia relativa mostró que el pico más importante de *contigs* virales de la boca (42-46 kpb) coincidía con el pico principal de tamaños de los bacteriófagos alojados en la base de datos del *GenBank* (42 kpb) (**Fig. 20**). El tercero de los picos de *contigs* virales más abundante, centrado en torno a las 17 kpb, presentaba similitudes de secuencia con virus del mismo tamaño como *Actinomyces phage Av1* (17.171 pb). Estos resultados sugieren que buena parte de los *contigs* virales ensamblados podrían estar completos o casi completos. En este sentido, cabe destacar seis *contigs* virales de gran tamaño (108-202 kpb) relacionados con especies de bacteriófagos también de gran tamaño: *Prochlorococcus phage P-SSM2*, *Erwinia phage PhiEaH1*, *Pelagibacter phage HTVC008M*, *Campylobacter virus CPt10* o *Ralstonia phage RSF1*. Los tres *contigs* más largos presentaban en un tamaño similar al genoma de los virus de referencia más relacionados por secuencia y presentaban extremos solapantes, indicando que estaban completos (**Fig. 21**).

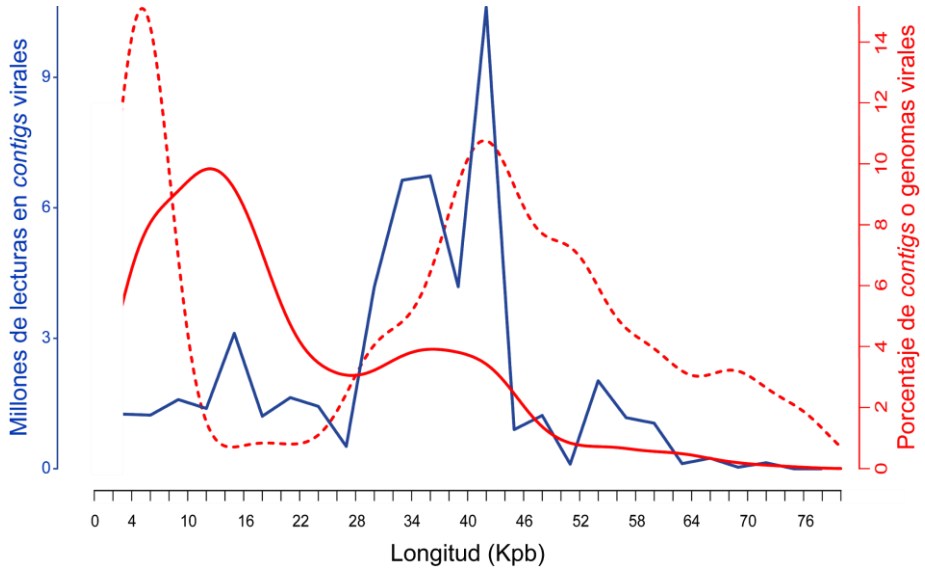

Figura 20. Distribución de *contigs* virales según su tamaño. La línea roja hace referencia a la frecuencia de *contigs* virales de la cavidad bucal (continua) o genomas de bacteriófagos de ADN en el *GenBank* (discontinua), mientras que la línea azul indica la abundancia de lecturas alineadas con los *contigs* virales. Los datos se agruparon en bloques de 3 kpb. Los genomas de los bacteriófagos de ADN del *GenBank* utilizados en este estudio incluyen 5.199 genomas de virus del orden *Caudovirales*, 86 de la familia *Inoviridae* y 1.536 de la familia *Microviridae*.

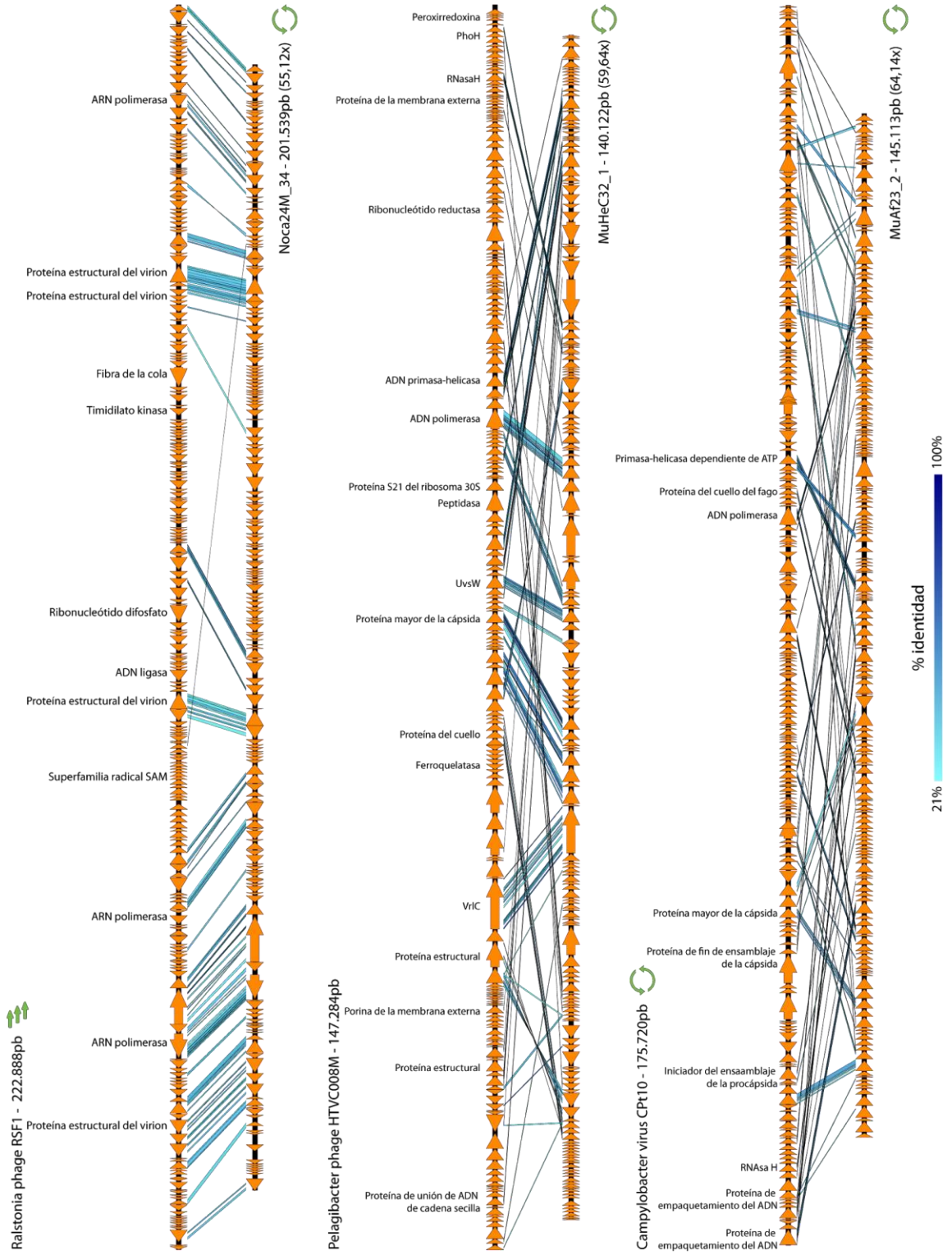

Figura 21. Estudio de la sintenia de algunos *contigs* de gran tamaño (>145 kpb) con los bacteriófagos más relacionados disponibles en las bases de datos. Alineamiento de los *contigs* Noca24M_34, MuHeC32_1 y MuAf23_2 con los genomas de referencia de las especies *Ralstonia phage RSF1*, *Pelagibacter phage HTVC008M*, *Campilobacter virus CPt10*, respectivamente. Los genes predichos con *Prodigal* y anotados por comparación con la base de datos de proteínas *nr* del *GenBank* se indican mediante flechas naranjas. Se muestran sólo aquellos alineamientos (*tBLASTx*) con un *e-value* < $1x10^{-10}$. Las flechas verdes circulares indican la naturaleza circular de los genomas virales de referencia y de los *contigs*, y las flechas verdes parcialmente solapantes indican la naturaleza de permutación cíclica de alguno de los genomas de referencia.

2.7. Los viromas de placa dental y mucosa oral se agrupan por ambiente, pero no por estado de salud, en estudios de diversidad beta

Con el fin de estudiar la influencia de la localización dentro de la cavidad bucal en la composición de los viromas, estudiamos en primer lugar la información compartida dos a dos entre viromas del mismo o distinto ambiente.

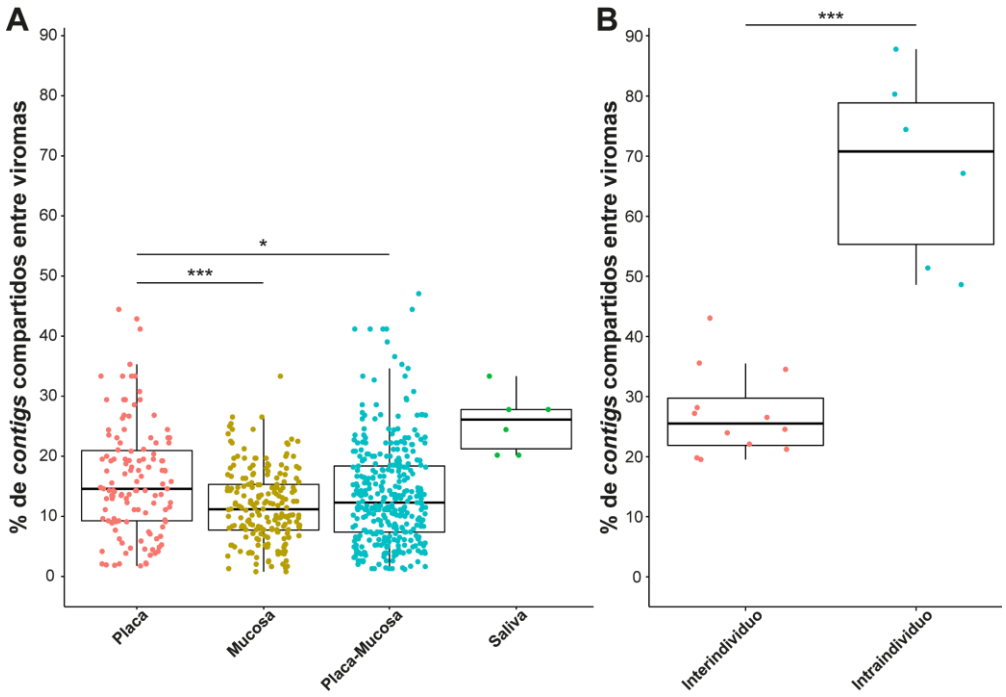

Figura 22. *Contigs* compartidos entre viromas diferentes de la cavidad bucal. Cada punto representa el porcentaje de *contigs* compartidos de un viroma con otro viroma del mismo o distinto ambientes pero de individuos distintos (A), o entre viromas de saliva y mucosa del mismo individuo (intraindividuo; 3 individuos) y saliva y mucosa de individuos distintos (interindividuo) (B). Las comparaciones se llevaron a cabo por parejas de viromas (Nucmer; >80% de identidad en al menos 1 kpb) considerando el porcentaje de *contigs* compartidos en ambas direcciones. El análisis estadístico de las diferencias entre grupos se hizo mediante un test de Mann-Whitney. *p-value* (Mucosa/Placa) = $1{,}45x10^{-05}$, *p-value* (Mucosa/Placa-Mucosa) = $3{,}09x10^{-02}$, *p-value* (Placa/Placa-Mucosa) = $6{,}81x10^{-03}$ y *p-value* (interindividuo/intraindividuo) = $4{,}73x10^{-05}$ * *p-value* < 0,01, ** *p-value* < 0,005 y *** *p-value* < 0,001.

Como era de esperar, el porcentaje de *contigs* compartidos entre viromas del mismo ambiente pero distinto individuo fue bajo (medianas del 14,58% en placa dental y del 11,45% en mucosa), y algo superior en saliva (mediana del 27,15%). Un análisis estadístico de estos datos mostró que el número de *contigs* compartidos entre viromas de placa dental de individuos distintos era significativamente mayor al que comparten viromas de mucosa también de individuos distintos, lo que sugiere que la diversidad beta de la comunidad de virus en placa es menor que en mucosa (**Fig. 22A**). Por otra parte, los viromas de individuos distintos (interindividuo) muestran un porcentaje mucho más bajo de *contigs* compartidos (mediana del 25,42%) que entre viromas de un mismo individuo (intraindividuo) (mediana del 70,79%), aun siendo viromas de ambientes distintos (**Fig. 22B**). Estos resultados coinciden con estudios previos que apuntan a que las comunidades virales de cada individuo son muy diferentes entre sí. A partir de

este punto, los viromas de saliva se excluyeron para los siguientes análisis debido al alto número de virus compartidos con los viromas de mucosa del mismo individuo.

Cluster	**Nombre**	**Nº *contigs***	**Lecturas**	**Longitud representante$**	**Mucosa**	**Placa**
1	*Arthrobacter phage Mudcat;Rhodococcus phage ReqiPoco6;Rhodococcus phage ReqiPepy6*	182	15.378.501	46.568	91	91
31	*Mycobacterium phage Muddy*	9	4.295.224	33.477	4	5
8	*Gordonia phage Splinter;Mycobacterium phage Nhonho*	25	2.997.167	39.128	13	12
3	*Microbacterium phage vB_MoxS-ISF9;Mycobacterium phage Jolie2*	36	2.684.773	43.776	23	14
11	*Gordonia phage Emalyn*	18	2.262.361	31.849	10	8
20	*Rhodococcus phage ReqiPoco6*	11	1.369.556	41.602	7	4
37	*Xanthomonas phage Xp15*	7	1.090.822	60.552	5	2
7	*Enterococcus phage phiFL4A*	24	1.088.896	49.737	20	6
75	*Pseudomonas phage NP1*	3	999.835	54.764	2	1
22	*Lactococcus phage P078;Lactococcus phage P118*	11	925.333	56.142	8	3
34	*Arthrobacter phage Mudcat*	8	898.081	41.823	7	1
9	*Mycobacterium phage bron*	22	892.317	35.489	10	12
2	*Streptococcus phage SM1;Streptococcus phage EJ-1*	67	878.148	42.426	50	17
19	*Lactococcus phage 1706*	12	815.937	55.783	8	4
4	*Mycobacterium phage Crossroads;Streptomyces phage phiHau3*	37	629.559	44.215	21	16

Tabla 6. Información de los 15 *clusters* de *contigs* más abundantes en función del número de lecturas alineadas. Longitud representante$: Se indica el tamaño del *contig* más largos.

Debido al bajo porcentaje de *contigs* compartidos entre viromas de individuos distintos, agrupamos los 1.557 *contigs* procedentes de todos los viromas en función de su identidad de secuencia (*Nucmer*; >80% de identidad sobre >1 kpb). De esta manera, obtuvimos 130 *clusters* (que agrupaban 1.077 *contigs*) y 480 *contigs* no agrupados. A cada *cluster* se le asignó el nombre del virus más relacionado por similitud de secuencia siguiendo un criterio similar al aplicado anteriormente para cada *contig* (la especie de virus más representada de entre todos los resultados obtenidos por *BLASTx* al comparar las pautas abiertas de lectura de todos los *contigs* de un mismo *cluster*) (**Tabla 6**). El mayor de estos *clusters* (*cluster* 1) estaba compuesto por 182 *contigs*, alguno de ellos entre los cinco más abundantes en 23 de los 27 viromas estudiados. Además, este *cluster* presentaba una distribución ubicua con 91 *contigs* de viromas de mucosa y otros 91 de viromas de placa. La mayoría de los integrantes de este *cluster* estaban relacionados (*e-value* = $8x10^{-52}$; mediana de los mejores resultados encontrados para cada *contig* del *cluster*) por similitud de secuencia con los bacteriófagos de las bacterias de suelo *Arthrobacter phage Mudcat*, *Rhodococcus phage ReqiPoco6* y *ReqiPepy6* y en menor medida con algunos bacteriófagos que infectan el género *Lactococcus*. También encontramos otros seis *clusters* relacionados con *Arthrobacter phage Mudcat* (43 *contigs*), tres *clusters* (31 *contigs*) con *Rhodococcus phage ReqiPoco6* y dos *clusters* (9 *contigs*) relacionados con *Rhodococcus phage ReqiPepy6*. La mayoría de estos *contigs* presentaban tamaños entre 30 y 50 kpb, sensiblemente inferiores a los bacteriófagos relacionados en la base de datos (53 y 78 kpb). Otro de los *clusters* más abundantes (*cluster* 2) estaba formado por 67

contigs estrechamente relacionados con *Streptococcus phage SM1* (*e-value* = 2,5x10^{-108}, mediana), un virus que infecta un poblador muy habitual de la boca como es *Streptococcus mitis*. También encontramos hasta 11 *clusters* adicionales (59 *contigs*) emparentados con bacteriófagos que infectan bacterias del género *Streptococcus* y preferentemente localizados en las muestras de mucosa. El mayor número de *clusters* asociados a un virus en particular fue a *Actinomyces virus Av1* (13 *clusters* que contenían 65 *contigs*). Los *clusters* con mayor número de componentes no se correspondían, en todos los casos, con aquellos que contenían *contigs* más abundantes. Así, por ejemplo, el *cluster* 31 era el segundo más abundante con más de 4 millones de lecturas en solo 9 *contigs* y estaba relacionado con *Mycobacterium phage Muddy*. Aunque la inmensa mayoría de los *clusters* estaban relacionados con virus que infectan bacterias, también detectamos un *cluster* relacionado con *Human betaherpesvirus 7* (*cluster* 89).

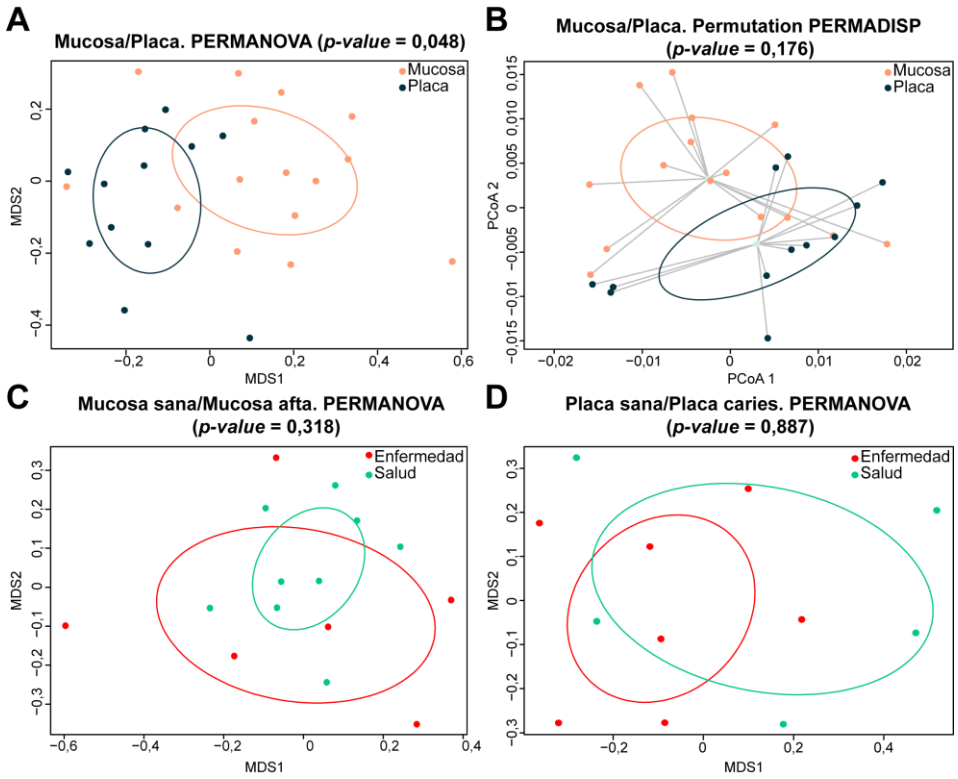

Figura 23. Sistema de ordenación bidimensional de viromas de placa dental y mucosa bucal humana. La abundancia normalizada de los *contigs* y *clusters* de *contigs* de cada viroma se usó para calcular disimilitudes Bray-Curtis, que se representaron en dos dimensiones mediante *NMDS* (A, C y D) o mediante PCoA (B). La significancia estadística de las diferencias entre grupos definidos por ambiente (A) o estado de salud (C) y (D) se evaluó mediante un test PERMANOVA y las diferencias en dispersión (B) mediante un test PERMADISP (*permutest-betadisper*) (Anderson, 2006).

Teniendo en cuenta la distribución de *clusters* y *contigs* compartidos entre viromas, calculamos las disimilitudes Bray-Curtis entre viromas y las representamos en sistemas de ordenación bidimensional *NMDS*. Los resultados obtenidos (**Fig. 23**) muestran una cierta separación entre los viromas de placa dental y de mucosa bucal, que resultó estadísticamente significativa en un estudio PERMANOVA (*p-value* = 0,048) (**Fig. 23A**). Pese a que el estudio de los *contigs* compartidos dos a dos entre viromas

sugería una diversidad beta inferior en placa dental que en mucosa bucal (**Fig. 22**), las distancias al centroide entre los dos ambientes no mostraron diferencias estadísticamente significativas en un test *permutest-betadisper* (*p-value* = 0,176) (**Fig. 23B**). Desafortunadamente, no observamos diferencias significativas entre viromas del mismo ambiente en función del estado de salud (PERMANOVA: *p-value* (mucosa sana y aftas) = 0,318, y *p-value* (placa sana y caries) = 0,887) (**Fig. 23C,D**).

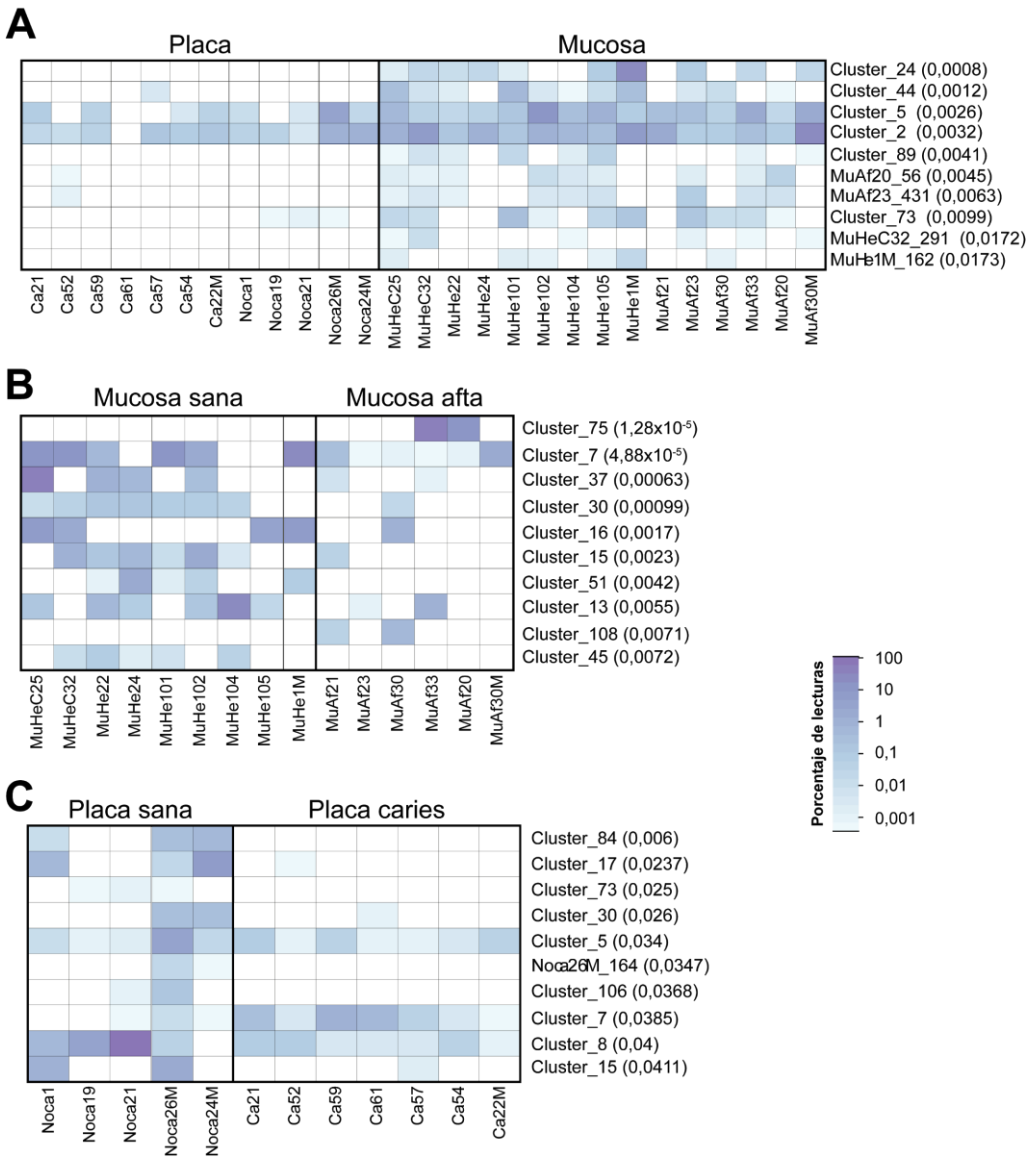

Figura 24. Mapa de calor de los 10 *clusters* o *contigs* no agrupados con distribución preferente por alguno de los dos ambientes estudiados (A) o por los estados de salud y enfermedad (B) y (C). Se muestran sólo aquellos resultados con mayor significancia estadística en un test de Mann-Whitney (A) o en *metagenomeSeq* (fitZIG) (B y C) y en orden creciente de *p-value*. El color indica la abundancia relativa de cada *cluster* o *contig* medida en porcentaje de lecturas alineadas. Entre paréntesis se indica el *p-value* de las diferencias significativas entre grupos.

Pese a que a nivel poblacional encontramos sólo una ligera separación de los viromas por ambiente, y no por condición de salud, quisimos estudiar si existían *contigs* o *clusters* con una distribución preferente para un determinado ambiente o condición de salud, que pudiera ser utilizado como biomarcador de esas

variables ambientales. Empleando un test no paramétrico de Mann-Whitney, identificamos 30 *clusters* o *contigs* asociados a un determinado ambiente, en su mayor parte a mucosa oral como significativos (**Fig. 24**). Los que presentaron *p-values* más bajos (*cluster* 24, 44, 5, 2 y 89) estaban asociados con mucosa bucal relacionados por secuencia con *Streptococcus phage YMC-2011* (*e-value* = $2x10^{-81}$), *Streptococcus phage SMP* (*e-value* = $1x10^{-53}$), *Mannheimia phage vB_MhS_1152AP2* (*e-value* = $2x10^{-119}$), *Streptococcus phage SM1* (*e-value* = $6x10^{-105}$) y *Human betaherpesvirus 7* (*e-value* = $1x10^{-167}$, 99% identidad de nucleótido), respectivamente (**Fig. 24A**). La asociación con mucosa fue también significativa para otros cuatro *contigs* de *HSV7*. En la mayoría de los casos las diferencias entre los grupos se debían a la ausencia de representantes en las muestras de placa, a excepción de los *clusters* 2 y 5 que eran prácticamente ubicuos en todos los viromas pero más abundantes en mucosa.

De la misma forma, estudiamos la existencia de especies asociadas a los estados de salud o enfermedad en ambas localizaciones. Debido a que el bajo número de viromas nos impedía hacer un test de Mann-Whitney, utilizamos un modelo para distribuciones gaussianas infladas con ceros *metagenomeSeq* (fitZIG). A pesar de que no existían diferencias estadísticas entre los estados de enfermedad y salud cuando consideramos las poblaciones completas, sí que detectamos algunos *contigs* o *clusters* con patrones de distribución diferentes entre condiciones. Así, algunos *clusters* de *contigs* relacionados con *Burkholderia phage AH2* (*cluster* 30; *e-value* = $1x10^{-69}$), *Rhodococcus phage ReqiPoco6* (*cluster* 15; *e-value* = $3x10^{-35}$) y *Actinomyces virus Av1* (*cluster* 51; *e-value* = $1x10^{-101}$) estaban asociados preferentemente con mucosa oral sana y apenas fueron detectados en viromas de aftas (**Fig. 24B**). El *cluster* 7 (relacionado con *Enterococcus phage phiFL4A*, *e-value* = 10^{-114}), pese a tener una distribución prácticamente ubicua, era también más abundante en los viromas de mucosa sana. Por el contrario, el *cluster* 75 (relacionado con *Pseudomonas phage NP1*; *e-value* = $2x10^{-90}$ (Chaudhry et al., 2017)) se asociaba con aftas. En el caso de la placa dental (**Fig. 24C**), cabe destacar la asociación de los *clusters* 84 (relacionado con *Alteromonas phage vB_AmaP_AD45-P1*; *e-value* = $4x10^{-37}$) y 17 (relacionado con *Arthrobacter phage vB_ArtM-ArV1*; *e-value* = $1{,}95x10^{-45}$) con placa dental sana, y del *cluster* 7 (*Enterococcus phage phiFL4A*; *e-value* = $1x10^{-114}$) con caries. Este último *cluster* es particularmente interesante porque en mucosa estaba asociado con el estado de salud, mientras que en placa dental estaba asociado con caries, siendo el hospedador del virus de referencia más relacionado *Enterococcus faecalis,* una bacteria implicada en el desarrollo de caries (Badet y Thebaud, 2008).

2.8. Los 444 genomas virales completos o casi completos ensamblados desde los viromas bucales se organizan en 31 *megaclusters*

Como habíamos visto anteriormente (**Fig. 20**), una buena parte de los 1.557 *contigs* de los viromas de mucosa bucal y placa dental presentaban tamaños que podrían corresponderse con genomas completos o casi completos. Esto quedó de manifiesto al comprobar que 445 *contigs* (un 29% del total) presentaban un tamaño superior al 70% de tamaño del genoma de la especie viral más relacionada en las bases de

datos y/o se comprobó su naturaleza circular mediante la detección de solapamiento entre los extremos 5' y 3' del *contig*. En el caso particular de los *contigs* que conformaban el *cluster* más abundante en la boca (*cluster* 1), muchos mostraban coberturas muy elevadas, pero tamaños notablemente inferiores a los virus más relacionados (*Rhodococcus phage ReqiPepy6* y *ReqiPoco6* con tamaños de 76.797 y 78.064 pb). Con el objetivo de que estos *contigs* estuvieran representados en estudios posteriores, escogimos el tamaño de otro virus relacionado por similitud de secuencia a un nivel muy similar pero con un tamaño menor de 59.443 pb (*Arthrobacter phage Mudcat*).

MC*	**Número de *contigs* virales**					***BLASTp* (*PHAST*)**	
	Total	**>1% lecturas&**	**Tamaño$**	**%CGs#**	**Virus**	**Identidad@**	***e-value***
A	33	17	40,6	54,5	*Enterococcus phage phiFL4A*	63,4	$5x10^{-103}$
B	12	3	37,8	63,1	*Streptomyces phage phiHau3*	64,3	$1x10^{-82}$
C	49	33	45,8	55	*Arthrobacter phage Mudcat*	73,2	$1x10^{-53}$
D	4	1	35,4	54,2	*Vibrio phage N4*	57,3	$4x10^{-173}$
E	3	3	64,5	51,3	*Corynebacterium phage P1201*	65	$1x10^{-157}$
F	1	0	6,4	66,5	*Propionibacterium phage B5*	48	$4x10^{-63}$
G	18	3	14	57,3	*Rhodococcus phage RRH1*	62,9	$5x10^{-77}$
H	18	3	39	56,8	*Corynebacterium phage BFK20*	61	$1x10^{-148}$
I	15	6	40,1	58,1	*Microbacterium phage vB_MoxS-ISF9*	62,9	$3x10^{-43}$
J	5	5	31,4	53,4	*Gordonia phage Emalyn*	68,4	$5x10^{-113}$
K	18	6	36	53,9	*Gordonia phage Splinter*	54,6	$3x10^{-106}$
L	3	0	6,7	60,5	*Ralstonia phage RSS0*	64	$8x10^{-71}$
M	5	1	64,9	41,3	*Salmonella phage FSL SP-058*	69,8	$8x10^{-166}$
N	14	6	55,4	45,5	*Xanthomonas phage Xp15*	63,1	$9x10^{-70}$
O	9	4	48,5	60	*Clavibacter phage CN1A*	60,4	$1x10^{-55}$
P	13	6	44,8	53	*Pseudomonas phage NP1*	66,9	$1x10^{-96}$
Q	13	6	43,4	46,5	*Mannheimia phage vB_MhS_1152AP2*	75	$1x10^{-79}$
R	5	0	36,9	50,6	*Vibrio phage X29*	55,8	$6x10^{-92}$
S	10	0	25	47,9	*Mannheimia phage phiMHaA1*	67,8	$5x10^{-126}$
T	9	2	36,5	50,4	*Lactobacillus prophage Lj771*	48	$8x10^{-23}$
U	12	2	17,7	36,8	*Streptococcus phage Cp-1*	91,1	$3x10^{-149}$
V	4	0	15,2	62,9	*Actinomyces virus Av1*	47,3	$3x10^{-91}$
W	59	7	16,8	52,5	*Actinomyces virus Av1*	74,6	$6x10^{-96}$
X	2	2	61,2	41,5	*Acinetobacter phage phiAC-1*	57	$1x10^{-40}$
Y	5	0	36,7	38,7	*Clostridium phage phiMMP03*	68,2	$9x10^{-132}$
Z	53	9	33,6	41,4	*Streptococcus phage SMP*	84,5	$2x10^{-94}$
AA	9	5	87,8	44,5	*Ralstonia phage RS138*	63,2	$1x10^{-138}$
AB	7	4	36,9	47,3	*Serratia phage Eta*	48,6	$1x10^{-39}$
AC	16	7	46,7	39,2	*Lactococcus phage P087*	67,3	$2x10^{-47}$
AD	11	0	4,6	46,4	*Torque teno virus 28*	71,1	$2x10^{-157}$
AE	9	1	5,7	38,7	*Parabacteroides phage YZ-2015a*	35,9	$4x10^{-31}$

Tabla 7. Características de los 31 *megaclusters*. MC*: *megacluster*. > 1% lecturas&: Contigs con más de un 1% de las lecturas de su viroma. Tamaño$: mediana de los tamaños de los *contigs* de cada *megacluster* expresados en kpb. # %CGs: media de porcentaje de CG de los *contigs* que conforman cada *megacluster*. Identidad@: % de identidad media de los mejores resultados obtenidos para cada *contig* del *megacluster*.

A continuación, generamos un árbol proteómico basado en distancias basadas en alineamientos por *BLASTx* para cada pareja de los 445 genomas completos o casi completos de virus de la boca y los 205 genomas virales de referencia más relacionados por similitud de secuencia (**Fig. 25A**). Utilizando este árbol agrupamos todos estos genomas en 31 *megaclusters* (**Tabla 7**), que representan un nivel de agregación superior al de los *clusters* empleados en secciones anteriores. Atendiendo a la agrupación de los genomas de referencia, obtuvimos resultados consistentes con estudios previos basados también en similitudes entre proteomas (Rohwer y Edwards, 2002): (i) ausencia de soporte a la actual división del orden *Caudovirales* en tres familias virales, con miembros de estas familias repartidos en *megaclusters* distintos, o (ii) agrupación de virus de las mismas subfamilias como la agrupación de los nueve genomas de referencia de la subfamilia *Autographivirinae* incluidos en el árbol proteómico dentro del *megacluster* **D**, los nueve *Peduovirinae* en el *megacluster* **S**, o los seis *Spounavirinae* y los seis *Tevenvirinae* en el *megacluster* **AA**.

A continuación, quisimos evaluar si la profundidad de muestras con la que trabajamos en esta tesis (27 viromas) era suficiente para explorar la diversidad de virus bucales. Para ello, cuantificamos el número de *megaclusters* representados cuando se escogen combinaciones al azar de los viromas (1.000 permutaciones en cada caso) (**Fig. 25C**). Los resultados muestran cómo con combinaciones de sólo 10 viromas se obtienen *contigs* completos o casi completos de 29 de los 31 *megaclusters*. También quisimos comprobar si este árbol proteómico obtenido con *contigs* casi completos era útil para representar la diversidad de los 1.557 *contigs* virales (**Fig. 26**). Para ello, construimos una red bidimensional basada en las conexiones *BLASTn* entre todos los *contigs* y observábamos que la inmensa mayoría de los *contigs* virales se encontraban dentro o muy próximos a los *megaclusters* definidos previamente en el árbol proteómico. De los 80-90 *contigs* restantes, algunos se agregaban en la red formado 12 nuevos *megaclusters* sin representantes potencialmente completos o circulares. La principal característica que presentaban muchos de estos nuevos grupos es que sus *contigs* estaban relacionados con bacteriófagos de gran tamaño (> 100 kpb). Así por ejemplo el *megacluster* **AK** contenía 33 *contigs* de los cuales 18 estaban relacionados con *Rhodococcus phage E3* (142,6 kpb), los del *megaclusters* **AL** con *Bacillus virus G* (497,5 kpb), los del *megacluster* **AM** con *Pelagibacter phage HTVC008M* (147,3 kpb) y los *contigs* de otros tres *megaclusters*: **AO, AQ** y **AP** con *Ralstonia phage RSF1* (222,9 kpb), *Ralstonia phage RSL2* (223,9 kpb) y *Bacillus phage SP-15* (221,9 kpb).

A este nivel de agrupación en *megaclusters* no detectamos ningún tipo de asociación por ambiente o por estado de salud (**Fig. 25B**), lo que sugiere que estos 31 *megaclusters* están formados por virus ampliamente distribuidos en la cavidad bucal. También cabe destacar que 20 de los 31 *megaclusters* tenían genomas de virus de referencia, el resto de *megaclusters* (**A**, **B**, **F**, **J**, **K**, **V**, **X**, **Y** y **AB**), no parecen tener representantes ni siquiera lejanos en las bases de datos.

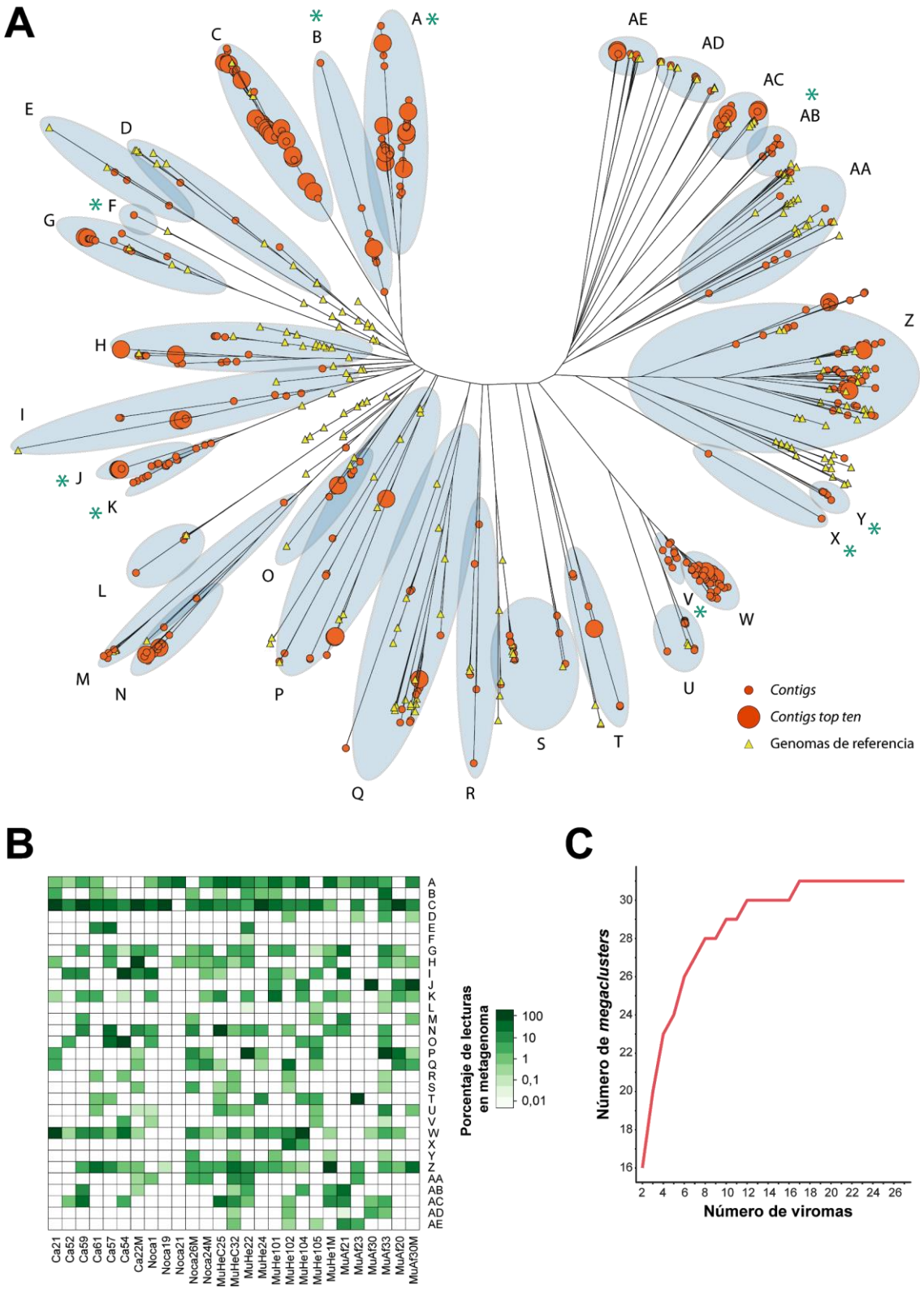

Figura 25. (A) Árbol proteómico de 444 *contigs* completos o casi completos y 205 genomas de referencia relacionados. Las distancias entre genomas se determinaron mediante una modificación del método *Dice* y se generó un árbol *Neighbor Joining* de las mismas. Los círculos de mayor tamaño indican su pertenencia a los diez *contigs* más abundantes de cada viroma. Se muestran sombreadas las 31 ramas que definen los *megaclusters* (designados con una o dos letras). Los asteriscos verdes indican qué *megaclusters* no incluyen genomas de referencia (B) Mapa de calor del porcentaje de lecturas de cada viroma que alinean con todos los *contigs* contenidos en cada *megacluster*. (C) Número de *megaclusters* con representantes cuando se elige al azar un número creciente de viromas. Se muestra la media de 1.000 permutaciones de viromas en cada punto.

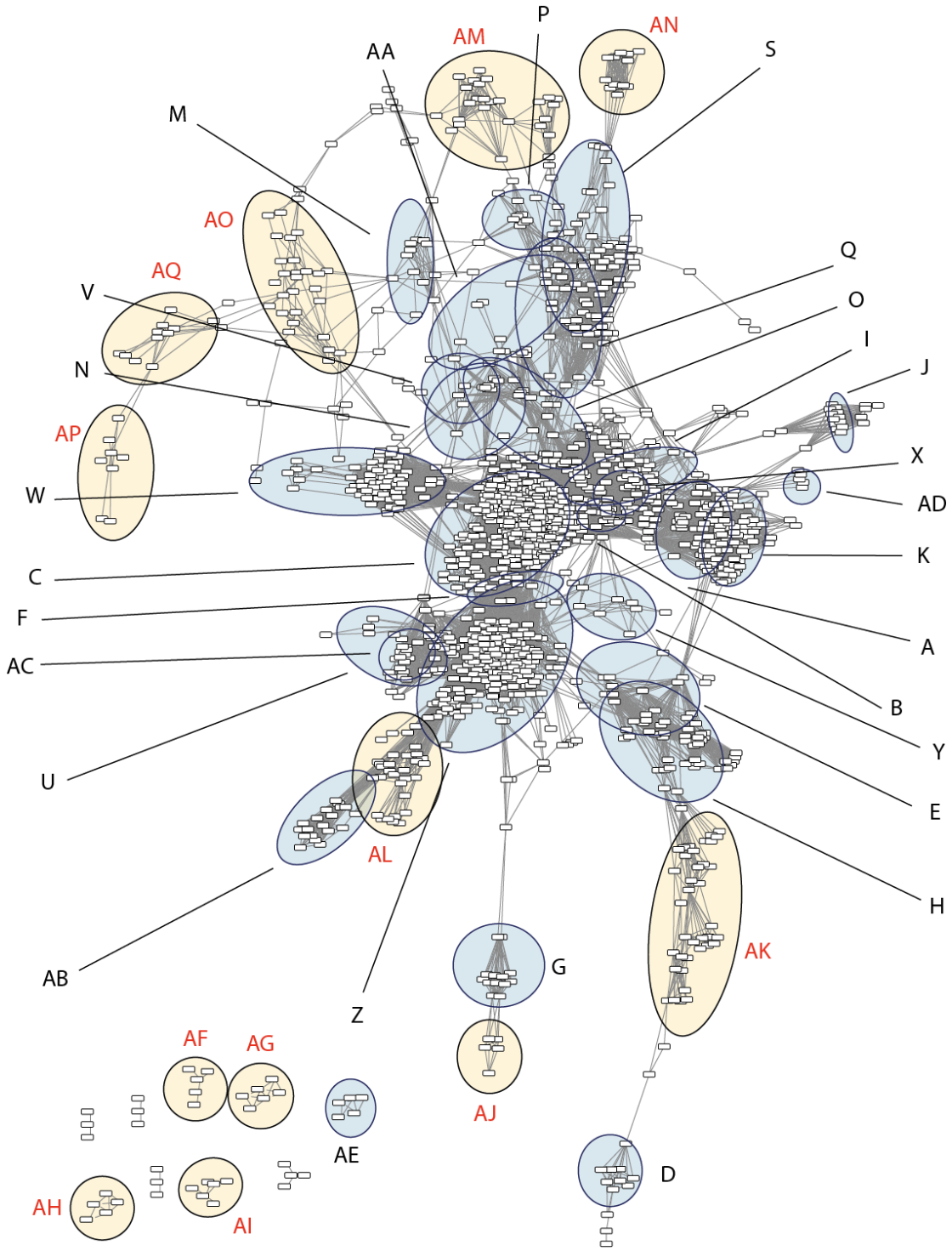

Figura 26. Red de los 1.557 *contigs* virales obtenida en función de las similitudes de *BLASTn* y agrupación por *MCL* (Enright et al., 2002)**.** Solo están representados aquellos contigs con tres o más conexiones. Se indican con un círculo amarillo 12 nuevas agrupaciones (**AF**-**AP**) formadas por al menos cinco *contigs* y sin ningún representante de los 444 genomas completos o casi completos.

2.8.1. Identificación de cuatro nuevos virus humanos de las familias *Anelloviridae* y *Papillomaviridae.*

Pese al casi absoluto dominio de bacteriófagos en las comunidades virales de la boca humana, también detectamos una población minoritaria de virus eucarióticos que se agrupaban exclusivamente dentro del *megacluster* **AD** (**Fig. 25A**). Entre ellos, pudimos ensamblar el genoma completo y circular de 13 virus de la familia *Anelloviridae* (**Fig. 27C**), cinco *Papillomaviridae* (**Fig. 27D**), y cinco virus eucarióticos de ADN de cadena sencilla circular que codifican por proteínas de replicación (*CRESS-DNA*) (Rosario et al., 2012) (**Fig. 27E**).

De los 13 genomas completos de anellovirus, 11 pudieron asociarse con especies ya secuenciadas del género *Alfatorquetenovirus*, incluyendo representantes en los tres clados monofiléticos de este género (**Fig. 27A**). Los dos anellovirus restantes (TTMV-ALH8 y TTMV-ALA22) estaban relacionados con especies del género *Betatorquetenovirus*, pero sus proteínas codificadas en los genes *ORF1* no presentaban similitud de secuencia por encima del criterio de demarcación aceptado en esta familia para designar nuevas especies (65%). Así, la proteína más parecida a la del virus TTMV-ALH8 fue la de *T-like mini virus TTMV_LY1* (con un 59,11% de identidad de aminoácido) y la de TTMV-ALA22 fue *TLMV-CLC062* (con un 60,68%). Esto nos permitió definirlos como dos especies nuevas de anellovirus humanos (Parras-Moltó et al., 2014).

De los seis papilomavirus humanos, cuatro pertenecían al género *Gammapapillomavirus* y uno a *Betapapillomavirus*. Dos de los primeros se correspondían con nuevos tipos de papilomavirus humanos, ya que el gen que codifica por la proteína *L1* de estos virus presentaba similitud de secuencia por debajo del criterio de demarcación aceptado para designar nuevos tipos humanos (90%): MuHe102-contig1113 presentaba sólo un 71,88% de identidad de nucleótido con el tipo *HPV146 y* Noca16-contig657 presentaba sólo un 74,10% con el tipo más parecido (*HPV178*). Estos dos genomas fueron clonados en los plásmidos *pSpark V* (Canvax) y *pGEM-T Easy Vector* (Promega) respectivamente, y depositados en el Centro de Referencia Internacional del Virus del Papiloma Humano (Instituto Karolinska, Suecia). En dicho instituto, se confirmaron las secuencias nucleotídicas y se les adjudicaron los nombres *HPV207* y *HPV208* respectivamente.

Los cinco virus *CREES-DNA* ensamblados presentaban dos genes dispuestos en orientaciones opuestas (*ORF1* y *ORF2*), y relacionados por secuencia con genes de los virus *Porcine stool-associated circular virus 5* (*e-value* = $3x10^{-132}$, mediana) y *Human PoSCV5-like circular virus* (*e-value* = $1x10^{-100}$, mediana). Aunque no pudimos ensamblar el genoma completo, obtuvimos también 13 *contigs* con tamaños entre 4.122 y 20.104 pb relacionados con herpesvirus humanos. Muchos de ellos presentaban una identidad de secuencia >99% con *Roseolovirus* herpesvirus humano 7 (*HSV7*) y se encontraban preferentemente en muestras de mucosa ya que encontramos alineamiento de más de cinco secuencias con el genoma de este virus en ocho de los 15 viromas de mucosa bucal, pero sólo en dos de los 12 viromas de placa

dental. También encontramos varios *contigs* del virus *Epstein Barr* (*HSV4*; >99% de identidad de secuencia) a los que se alineaban >5 secuencias de dos viromas de placa dental y uno de mucosa.

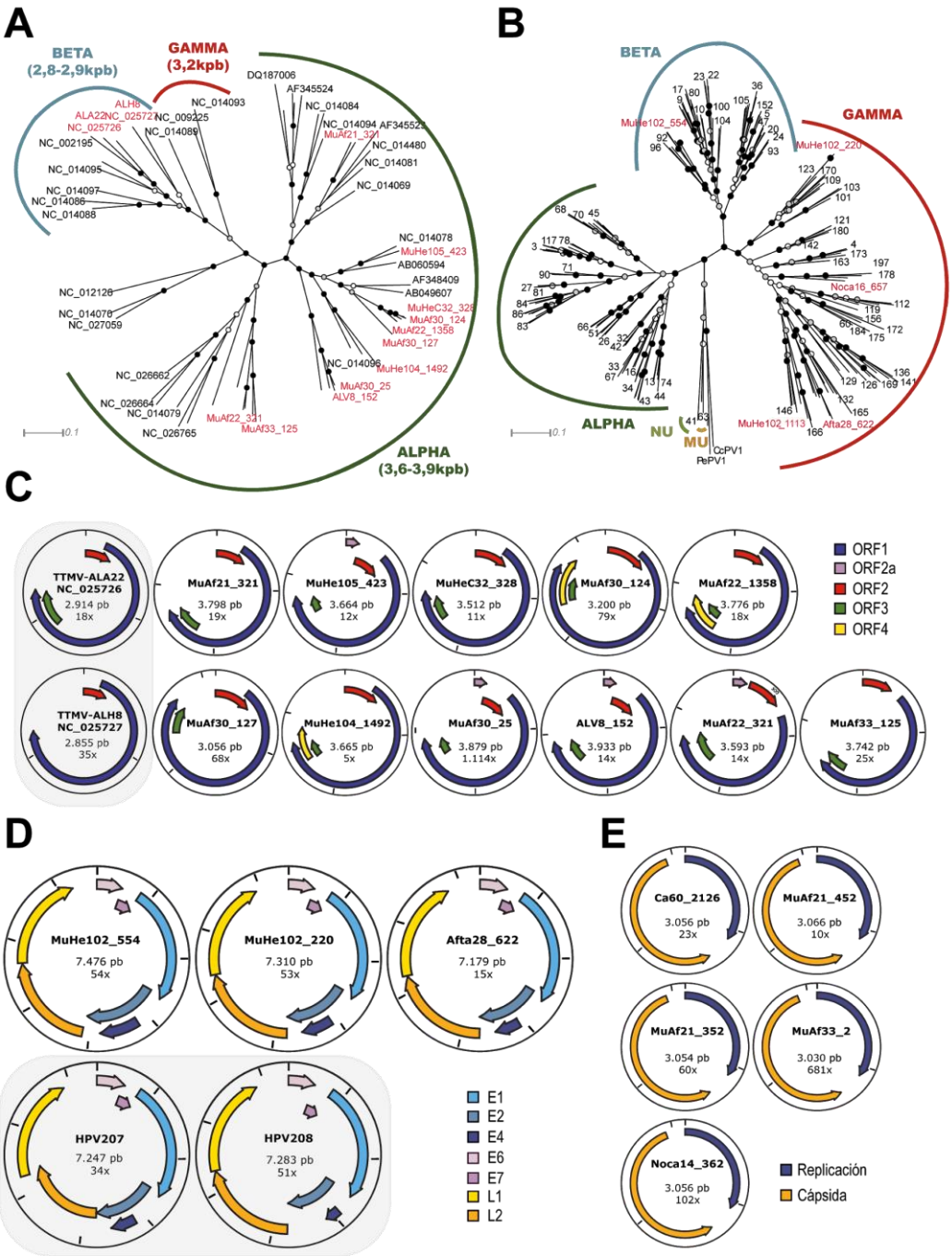

Figura 27. Análisis filogenético y estructural de algunos de los genomas de virus eucarióticos circulares ensamblados desde viromas bucales. Árboles filogenéticos basados en la proteína codificada en el *ORF1* de algunos genomas de referencia representativos de anellovirus humanos (A) y en la proteína L1 de varios papilomavirus humanos (B). Se incluyen también los genomas ensamblados en esta tesis (color rojo). Los alineamientos se hicieron mediante *Clustal Omega* y el árbol se construyó utilizando máxima verosimilitud. Los nodos de color negro indican valores de bootstrap >90% y en gris >70%. Los trazos exteriores indican los géneros de ambas familias de virus. Las figuras (C, D y E) representan los genomas completos de los anellovirus, papilomavirus y *CRESS-DNA* respectivamente, ensamblados desde los viromas bucales y dibujados con *SnapGene* (Biotech, s. f.). El color gris de fondo indica los cuatro genomas de virus humanos nuevos descritos en esta tesis.

2.8.2. La inmensa mayoría de los *megaclusters* de virus bucales están relacionados con bacteriófagos del orden *Caudovirales*

Como dijimos anteriormente, los virus de la mayoría de los 31 *megaclusters* estaban emparentados con bacteriófagos (**Tabla 7**): los de 27 *megaclusters* con bacteriófagos con cola del orden *Caudovirales*, los integrantes de los *megaclusters* **F** y **L** con bacteriófagos filamentosos de la familia *Inoviridae*, y los del *megacluster* **AE** con bacteriófagos pequeños de la familia *Microviridae*.

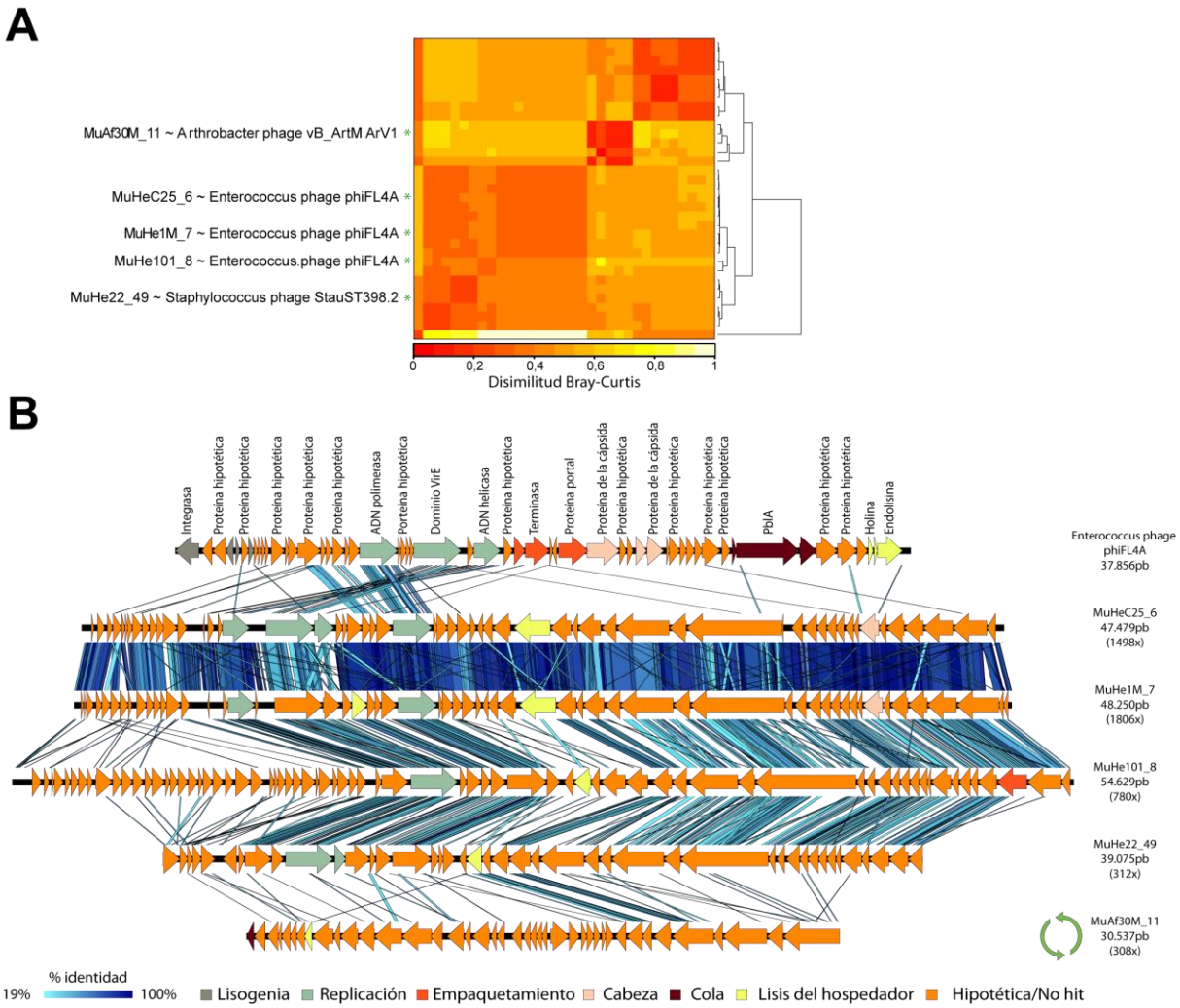

Figura 28. Estudio de la estructura genómica de los *contigs* relacionados con el genoma de referencia *Enterococcus phage phiFL4A* incluidos en el *megacluster* A. (A) Mapa de calor y dendograma de las distancias *Dice* modificadas entre alguno de los *contigs* del *megacluster* **A.** Se indican con un asterisco los *contigs* representativos seleccionados para el estudio de sintenia, indicando también el nombre del virus más relacionado en la base de datos. (B) Análisis de la sintenia entre varios *contigs* completos o casi completos representativos del *megacluster* A y el genoma de referencia *Enterococcus phage phiFL4A*. Los genes predichos con *Prodigal*, y anotados por comparación con la base de datos de proteínas *nr* del *GenBank*, se indican mediante flechas. Se muestran en azul sólo aquellos alineamientos (*tBLASTx*) con un *e-value* < $1x10^{-10}$. Las flechas verdes indican la naturaleza circular de uno de los *contigs*.

Los *megaclusters* **A** y **C** contenían un elevado número de *contigs* con gran cobertura, y muchos de ellos, situados entre los 10 *contigs* más abundantes de sus respectivos viromas. Los 17 *contigs* con mayor cobertura del *megacluster* **A** estaban formados por >1% de las lecturas de sus viromas, y muchos de ellos, emparentados con *Enterococcus phage phiFL4A*. Sin embargo, un estudio de la sintenia, mostró que incluso el alineamiento de los *contigs* más relacionados con *Enterococcus phage phiFL4A* (MuHe101_8 y MuHe1M_7) estaba restringido a una pequeña región con genes de replicación: (**Fig. 28**) la ADN polimerasa (*e-value* = $1x10^{-106}$ y $1x10^{-114}$), la ADN helicasa (*e-value* = $7x10^{-99}$ y $1x10^{-91}$), y la proteína de virulencia VirE (*e-value* = $4x10^{-90}$ y $2x10^{-08}$). Los genes restantes presentaban similitudes de secuencia con genes de una variedad amplia de virus como la subunidad mayor de la terminasa de *Mycobacterium phage vB_MapS_FF47* (*e-value* = $1x10^{-129}$), la proteína de cubierta de *Streptomyces phage Sujidade* (*e-value* = $3x10^{-49}$), la proteína principal de la cápsida de *Mycobacterium phage MosMoris* (*e-value* = $9x10^{-36}$), o Lisina A de *Rhodococcus phage ReqiDocB7* (*e-value* = $2x10^{-26}$). Además, los *contigs* de este *megacluster* presentaban genes dispuestos en las dos direcciones mientras que *Enterococcus phage phiFL4A* los tiene todos en la misma orientación. Otra diferencia sustancial es que nuestros *contigs* tienen un contenido de GCs del 54,54%, mientras que *Enterococcus phage phiFL4A*, como virus que infecta bacterias del filo *Firmicutes*, tiene un contenido de CGs del 37,8%. Estas diferencias, junto con el hecho de que ningún genoma de referencia se agrupe dentro de este *megacluster*, sugiere que los virus que conforman el *megacluster* **A** son virus nuevos no representados en las bases de datos.

Por otro lado, los 33 *contigs* más abundantes del *megacluster* **C** (formados por más de 1% de las lecturas de sus respectivos viromas) estaban emparentados con los virus *Arthrobacter phage Mudcat*, *Rhodococcus phage ReqiPoco6* y *ReqiPepy6* y *Lactococcus phage P092* (**Fig. 29**). El alineamiento de varios *contigs* representativos de este *megacluster* con *Arthrobacter phage Mudcat* mostró una clara sintenia desde el extremo 5´ hasta el extremo 3´ de los *contigs* con todos los genes orientados en el mismo sentido, apoyando la idea de que los *contigs* de este *megacluster* podrían estar completos. Aunque algunas de las principales proteínas codificadas en estos genomas presentaban similitud de secuencia con las proteínas de la cubierta, la unidad larga de la terminasa (TerL), la proteína portal o la proteína de la cabeza de *Rhodococcus phage ReqiPoco6* (medianas de *e-values* de $1x10^{-83}$, 0, $1x10^{-135}$ y $2,5x10^{-71}$, respectivamente), los *contigs* de este *megacluster* presentaban también diferencias importantes en tamaño. Así, por ejemplo, nuestros *contigs* carecían de aproximadamente 8 kpb del extremo 5´ del genoma *Rhodococcus phage ReqiPoco6* (*Arthrobacter phage Mudcat* tampoco presenta esta región) y del tercer grupo de genes de tamaño pequeño con dominios transmembrana *de Arthrobacter phage Mudcat*. Además, nuestros *contigs* también carecían por un lado del gen *LysB* y de genes de ARN transferente, presentes en *Rhodococcus phage ReqiPoco6*, y por otro, de un gen de gran tamaño que codifica por una proteína hipotética de 1.292 aminoácidos contigua al gen de la proteína de la cubierta de *Arthrobacter phage Mudcat*.

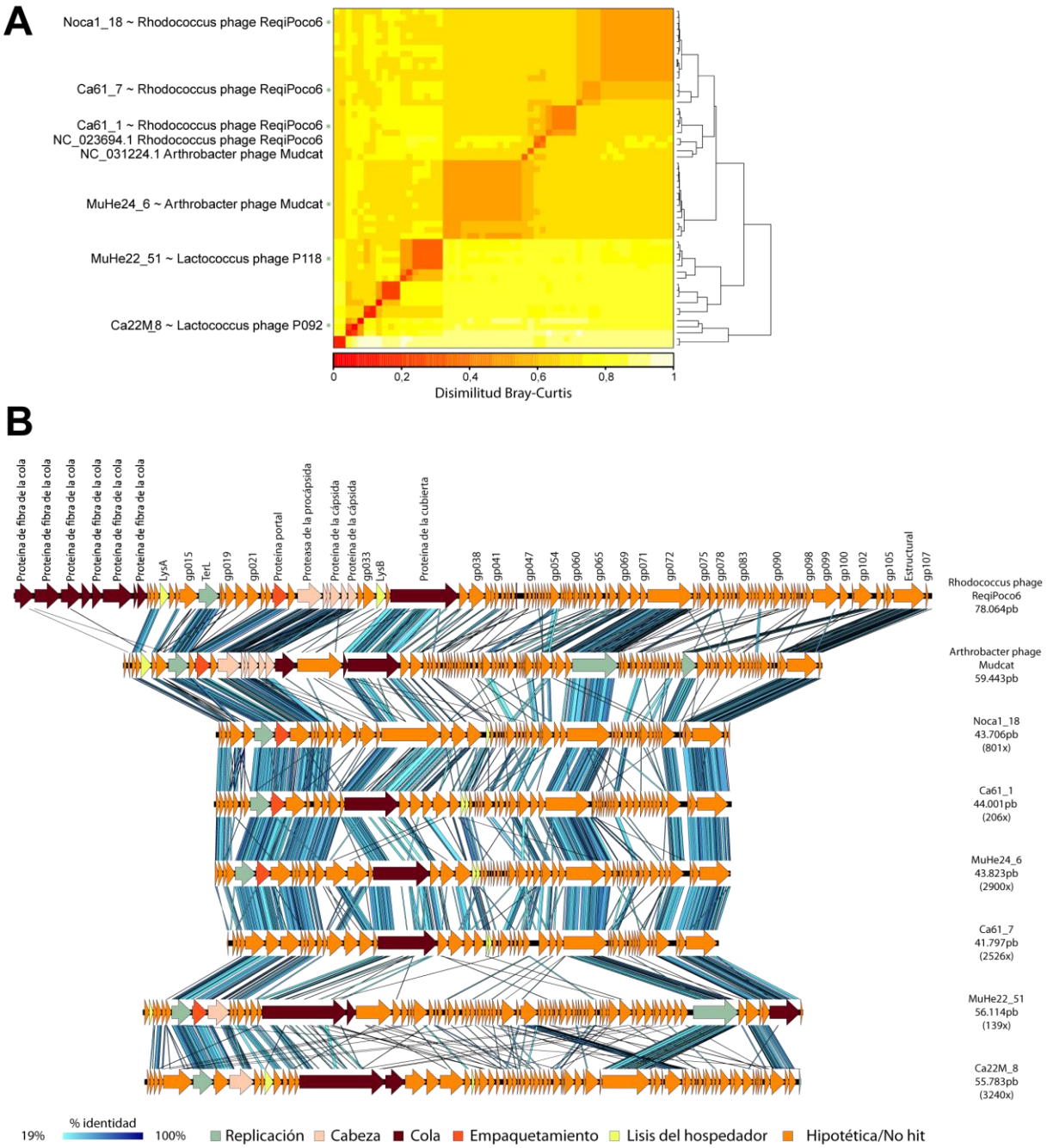

Figura 29. Estudio de la estructura genómica de los *contigs* del *megacluster* C. (A) Mapa de calor y dendograma de las distancias *Dice* modificadas entre los *contigs* del *megacluster*. Se indican con un asterisco los *contigs* representativos seleccionados de cada grupo para el posterior análisis de sintenia, indicando también el nombre del virus más relacionado en la base de datos. (B) Análisis de la sintenia de varios *contigs* completos o casi completos representativos del *megacluster* **C** y el genoma de referencia de los virus relacionados *Rhodococcus phage ReqiPoco6* y *Arthrobacter phage Mudcat*. Los genes predichos con *Prodigal* y anotados por comparación con la base de datos de proteínas *nr* del *GenBank* se indican mediante flechas. Se muestran en azul sólo aquellos alineamientos (*tBLASTx*) con un *e-value* < $1x10^{-10}$.

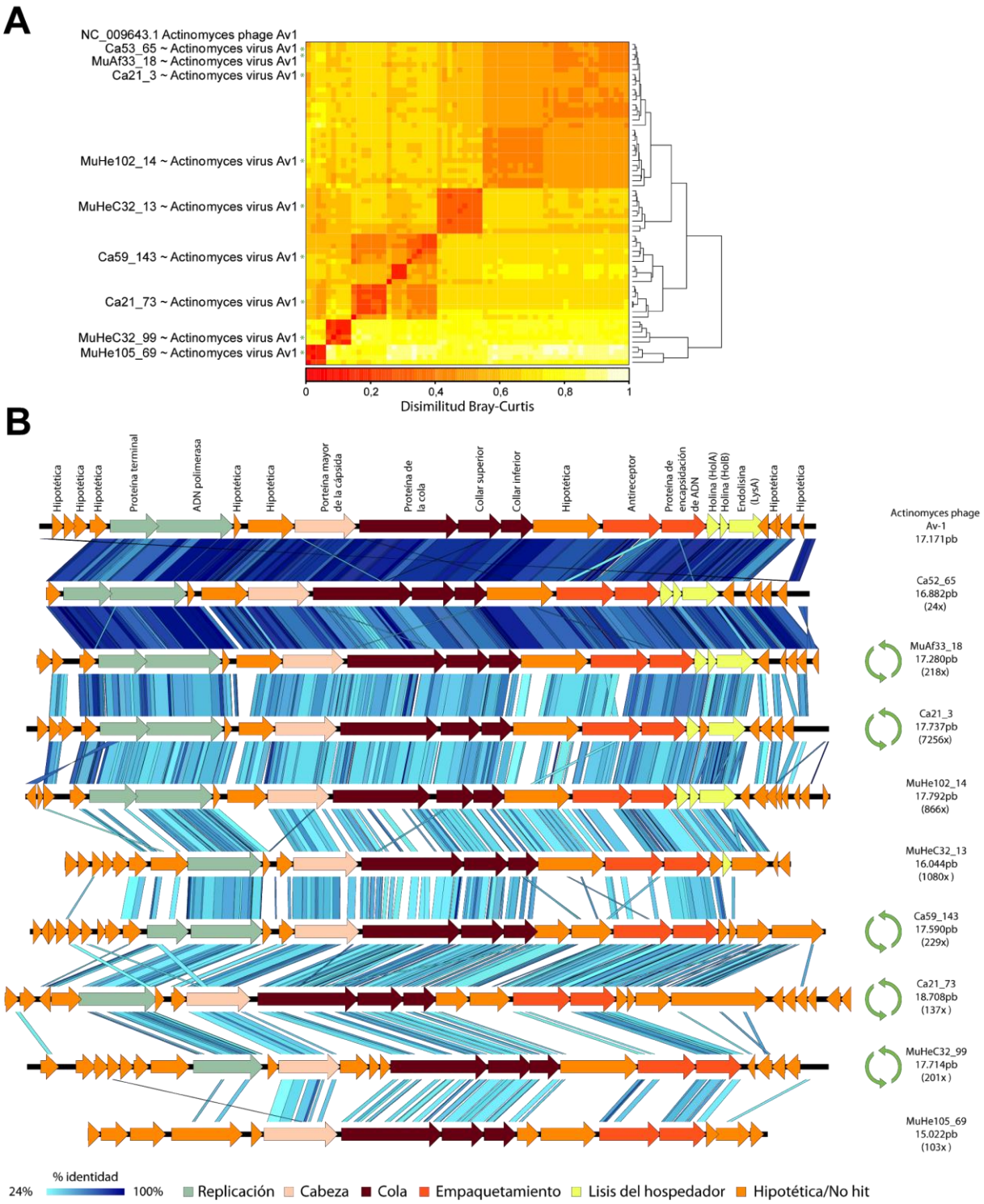

Figura 30. Estudio de la estructura genómica de los *contigs* de los *megaclusters* V y W. (A) Mapa de calor y dendograma de las distancias *Dice* modificadas entre los *contigs* de estos *megaclusters* y el genoma de referencia *Actinomyces phage Av1*. Se indica con un asterisco los *contigs* representativos seleccionados para el estudio de sintenia, indicando también el nombre del virus más relacionado en la base de datos. (B) Análisis de la sintenia de varios *contigs* completos o casi completos representativos y el genoma de referencia *Actinomyces phage Av1* Los genes predichos con *Prodigal* y anotados por comparación con la base de datos de proteínas *nr* del *GenBank* se indican mediante flechas. Se muestran en azul sólo aquellos alineamientos (*tBLASTx*) con un *e-value* < $1x10^{-10}$. Las flechas verdes indican la naturaleza circular de los de los *contigs*.

Un análisis más en profundidad de este *megacluster* nos permitió dividir sus *contigs* en dos grupos diferenciados: uno formado por 35 *contigs* con un contenido de CGs del 54,99% (± 1,88% desviación estándar) que presentaban un mayor número de genes relacionados con los bacteriófagos que infectan *Actinobacteria*: *Arthrobacter phage Mudcat* y *Rhodococcus phage ReqiPoco6/ReqiPepy6*, y otro grupo formado por 14 *contigs* con un %CGs del 39,37% (± 1,58% desviación estándar) y más relacionados con varios bacteriófagos que infectan *Firmicutes*: *Lactococcus phage P092* y *Lactococcus phage P118*. 19 de los *contigs* del primer grupo poseían endolisinas relacionadas con las de *Actinomyces phage Av1* (*e-value* = $5x10^{-34}$, mediana) mientras que los del segundo grupo tenían endolisinas relacionadas con las de algunos bacteriófagos de *Streptococcus* como *Streptococcus phage Dp-1* (*e-value* = $7x10^{-53}$, mediana).

Otros *megaclusters* con un gran número de *contigs* completos o casi completos fueron:

-El *megacluster* **W**, pese a tener una menor representación de *contigs* de alta cobertura, contenía hasta 59 *contigs* con un tamaño muy similar (16,80 kpb ± 1,32 kpb *SD*) al del virus más relacionado en las bases de datos: *Actinomyces phage Av1* (17,17 kpb) y 23 de ellos se ensamblaron en un genoma circular (**Fig. 30**). Además, este *megacluster* estaba estrechamente relacionado con el *megacluster* **V**, formado por otros cuatro *contigs* (dos de ellos circulares) también con similitud de secuencia por *Actinomyces phage Av1* y con el *megacluster* **U**, formado por 12 *contigs* completos con un tamaño de 17,68 kpb emparentados con el también podovirus (subfamilia *Picovirinae*) *Streptococcus phage Cp-1* (dos de ellos circulares). Un análisis de la ordenación genética de estos *contigs* comparada con el virus de referencia, mostró una buena sintenia en muchos de estos *contigs*, con un nivel alto de conservación en la región central donde se encuentran genes que codifican para la ADN polimerasa, la proteína de la cabeza, collar superior y collar inferior de la cápsida, o la proteína de encapsidación (*e-value* = 0, y >98% de identidad de aminoácido para el gen más relacionado en 10 de estos *contigs*) (**Fig. 30**). Sin embargo, también se encontraron, en varios de estos *contigs*, genes que no estaban presentes en *Actinomyces phage Av1* y, en muchos casos, no se parecían a ningún gen disponible en las bases de datos, aunque algunos de ellos codifican por proteínas con plegamientos tipo colágeno, que se podrían corresponder con proteínas de los filamentos laterales. Otro conjunto de genes que mostraban gran variabilidad son los relacionados con la lisis del hospedador como las holinas HolA y HolB y la endolisina. 32 de los 59 *contigs* del *megacluster* **W** presentaban endolisinas y sólo 29 de los 59 holinas, encontrando sólo 25 *contigs* con combinaciones de ambos genes. Los cuatro *contigs* del *megacluster* **V** carecía de estos genes.

-El *megacluster* **Z** estaba formado por 53 genomas completos o casi completos con un tamaño medio de 35,7 kpb que presentaban un %CGs bajo, lo que resulta coherente con el gran parecido que la mayoría de ellos presentan con bacteriófagos que infectan bacterias del género *Streptococcus* dentro del filo *Firmicutes*. También al igual que muchos de los bacteriófagos de *Streptococcus*, estos *contigs*

presentaban genes necesarios para el desarrollo de un ciclo lisogénico como integrasas/recombinasas o represores de la familia XRE.

-Los *megaclusters* **K** y **J** estaban formados por 18 y 5 genomas completos, muchos de ellos muy abundantes en sus respectivos viromas. Estos *contigs* presentaban tamaños y %CGs similares y se agrupaban muy próximos en el árbol proteómico. Además, ambos *megaclusters* mostraban similitud de secuencia por bacteriófagos del suelo aislados en bacterias del género *Gordonia*. La enorme abundancia de proteínas hipotéticas en estos *contigs*, junto con la ausencia de genomas de referencia en estos *megaclusters,* sugiere que son virus originales, muy diferentes de los disponibles en las bases de datos.

-Otros *megaclusters* interesantes fueron el *megacluster* **G**, que incluye algunos de los siphovirus con genomas más pequeños descritos, los *megaclusters* **N** y **M** con *contigs* de tamaños medios de 55,42 y 64,92 kpb respectivamente y con similitud por bacteriófagos de *Gammaproteobacteria*, el *megacluster* **AA** con algunos genomas de gran tamaño como alguno de los incluidos en la **Figura 21** y el *megacluster* **X** con un genoma de 84,45 kpb con un %CGs del 28,73% y con similitud de secuencia por algún gen conocido en solo 17 de sus 144 genes predichos.

Pese a que la mayoría de los *megaclusters* incluían genomas de referencia, es importante resaltar que las proteínas codificadas en sus *contigs* con mejor similitud de secuencia por proteínas de virus conocidos, no superan de media el 75% de identidad en la mayoría de los *megaclusters.* Este resultado subraya la originalidad de los virus de la boca humana descubiertos en esta tesis.

3. Los métodos de predicción de hospedador sugieren que los bacteriófagos de la cavidad bucal humana infectan esencialmente los filos bacterianos *Actinobacteria, Firmicutes, Proteobacteria* y *Bacteroidetes*

Con el fin de averiguar qué tipo de hospedador infectan los 1.557 *contigs* virales ensamblados desde los viromas de placa dental y mucosa bucal, empleamos cinco estrategias diferentes de predicción de hospedador.

3.1. Predicción del hospedador en función del virus de referencia más relacionado por *BLASTx*

La primera de las estrategias consistió en asumir que el hospedador de los *contigs* virales coincidía a niveles taxonómicos elevados (filo y clase) con el del virus con mayor similitud de secuencia encontrado en las bases de datos (**Fig. 31**). De esta manera, obtuvimos información taxonómica del posible hospedador para 1.449 de los 1.557 *contigs* virales. Un 46,7% se asignaron a virus que infectan el filo *Actinobacteria*, un 25,9% *Firmicutes* y un 18,1% *Proteobacteria*. Solo 32 y 6 de los 1.449 *contigs* estaban relacionados con virus que infectan *Bacteroidetes* y *Cyanobacteria*, respectivamente. Teniendo en cuenta la abundancia relativa de los *contigs*, un 70,7% de las secuencias de los viromas alineaban con *contigs* virales que podrían infectar *Actinobacteria*, un 10,5% *Firmicutes* y un 6,2% *Proteobacteria*.

Las ocho familias de hospedadores más representados fueron *Streptococcaceae*, *Nocardiaceae*, *Micrococcaceae*, *Mycobacteriaceae*, *Actinomycetaceae*, *Gordoniaceae*, *Enterobacteriaceae* y *Burkholderiaceae*, con un 16,12%, 12,59%, 10,86%, 6,3%, 5,52%, 4,95%, 4,17% y 3,85% de los *contigs*, respectivamente. Estos resultados indican que un número importante de virus de la boca infectan una amplia variedad de familias del filo *Actinobacteria*, mientras que los que infectan *Firmicutes* lo hacen mayoritariamente a miembros de la familia *Streptococcaceae*.

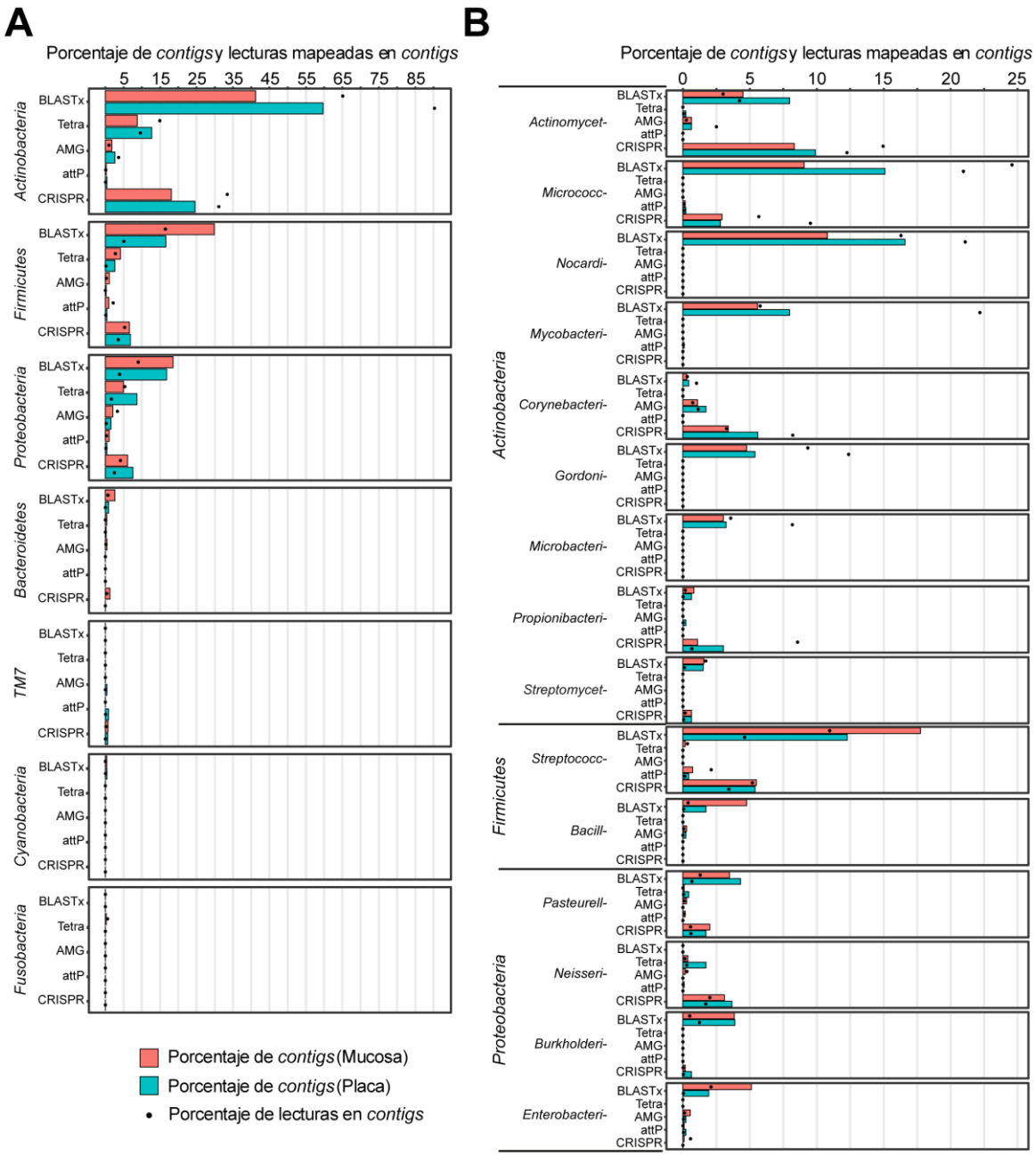

Figura 31. Predicción del hospedador de 1.557 *contigs* virales de la boca humana. Se muestran los *contigs* asociados a un determinado hospedador mediante cinco aproximaciones diferentes (BLASTx, Tetra, AMGs, attP y CRISPR). Las barras indican el porcentaje de *contigs* virales asociados a un determinado filo y familia de hospedador y los puntos hacen referencia al porcentaje de lecturas que alinean con dichos *contigs*, ambos valores relativizados al número total de contigs y lecturas en mucosa y placa, respectivamente. Se muestra los resultados de los (A) siete filos y (B) 15 familias de hospedadores bacterianos más representados.

Cabe destacar la mayor abundancia de bacteriófagos de *Actinobacteria* en placa dental con respecto a mucosa. Así, la relación entre el número de *contigs* virales que infectan bacterias de este filo y los que infectan *Firmicutes* fue de 1,38 en los viromas de mucosa y de 3,6 en los de placa dental. Estas diferencias fueron más evidentes cuando consideramos la abundancia relativa de estos *contigs* (% de lecturas alineadas) con ratios de 3,95 en mucosa y 17,8 en placa dental.

3.2. Predicción del hospedador mediante comparación de los perfiles de frecuencia de tetranucleótidos

En la segunda de las estrategias para establecer conexiones virus-hospedador, comparamos los perfiles de tetranucleótidos de los *contigs* virales y de las bacterias presentes en la boca humana. Como los virus parasitan la maquinaria de traducción de sus hospedadores, es esperable cierta coincidencia en el uso de codon, y por tanto, en la composición nucleotídica entre un virus y su hospedador. En este sentido, se ha demostrado que la composición de tetranucleótidos es una señal que puede ser utilizada para determinar el rango de hospedador de los bacteriófagos (Edwards et al., 2016; Ogilvie et al., 2013; Roux et al., 2015b). En el apartado de **Materiales y Métodos** comprobamos que, utilizando un valor del índice de correlación de Pearson de 0,84, podíamos predecir, con una fiabilidad del 90%, el hospedador de un bacteriófago a nivel de filo (**Fig. 8**). Utilizando esta estrategia para nuestros *contigs* virales y una base de datos con el genoma de 439 bacterias de la boca, detectamos conexiones para 314 *contigs* virales (un 20,23% del total). Los tres filos de hospedadores más representados mediante este método fueron *Actinobacteria*, *Proteobacteria* y *Firmicutes* (con un 9,96%, 6,04%, 3,66% de los *contigs*, y un 11,62%, 3,43% y 1,5%, de lecturas alineadas en estos *contigs*, respectivamente). A nivel de familia, y empleando un índice de correlación de Pearson de 0,94, obtuvimos conexiones para un número bajo de *contigs*, que se asociaban principalmente con bacterias de las familias *Neisseriaceae* y *Pasteurellaceae* (0,77% y 0,19% respectivamente).

3.3. Predicción del hospedador basado en la presencia de genes metabólicos auxiliares (*AMG*)

Una tercera aproximación empleada para inferir el hospedador de los *contigs* virales fue la identificación de genes metabólicos auxiliares (*AMG*). La probable adquisición de estos genes por transferencia horizontal desde sus hospedadores convierte a su asignación taxonómica en una valiosa herramienta de predicción del hospedador. Comparando los 43.590 genes predichos de los 1.557 *contigs* virales mediante la herramienta *eggNOG*, encontramos 371 genes con posibles funciones metabólicas. 120 de ellos (contenidos en 82 *contigs* virales) mostraban una similitud elevada (*e-value* < $1x10^{-50}$) con genomas bacterianos. Los principales hospedadores a nivel de filo detectados mediante este método fueron *Actinobacteria*, *Proteobacteria* y *Firmicutes*, con un 1,99% 1,86% y 0,83% de los *contigs* virales y un 2,02%, 1,77% y 0,14% de las lecturas alineadas a *contigs*, respectivamente.

3.4. Predicción del hospedador basado en secuencias de integración *attP*

La cuarta estrategia consistió en identificar secuencias de integración *attP* en los *contigs* virales y compararlas con las secuencias de ARN transferente de los genomas bacterianos. Encontramos 116 secuencias de integración *attP* en 68 *contigs* virales diferentes, de las que 45 (presentes en 37 *contigs* virales) presentaban alta similitud de secuencia (>35 pb y >90% identidad de secuencia) con regiones *attB* de genomas bacterianos. Mediante esta estrategia conseguimos adscribir solo un 1%, 0,9% y 0,77% de los *contigs* virales a los filos bacterianos *Proteobacteria*, *TM7* y *Firmicutes*, respectivamente.

3.5. Predicción del hospedador basado en las secuencias separadoras de los *CRISPRs* en microbiomas de la cavidad bucal

Por último, hicimos metagenómica mediante secuenciación de genomas completos desde el ADN total (microbiomas) contenido en los sedimentos bacterianos de las mismas muestras donde se habían analizado previamente los viromas. Empleamos las tecnologías de *HiSeq* (Illumina®) para secuenciar cinco microbiomas de mucosa y diez de placa dental (23.061.214 millones de secuencias pareadas de 2x250pb de media por microbioma), y PacBio® para secuenciar seis microbiomas de placa (132.884 secuencias de mediana por microbioma, con un tamaño medio de 4.399 pb). Los 15 microbiomas de *HiSeq* se procesaron de la misma forma que los viromas (**Fig. 32**), eliminando de media un 1,68% de secuencias de baja calidad. Además, identificamos un nivel elevado de contaminación con ADN del genoma humano en los microbiomas de mucosa (85,87% de media), por sólo un 2,98% en los microbiomas de placa dental. El ensamblaje *de novo* de estas secuencias filtradas *HiSeq* generó una mediana de 69.746 *contigs* por metagenoma (*N50* = 2.741 pb).

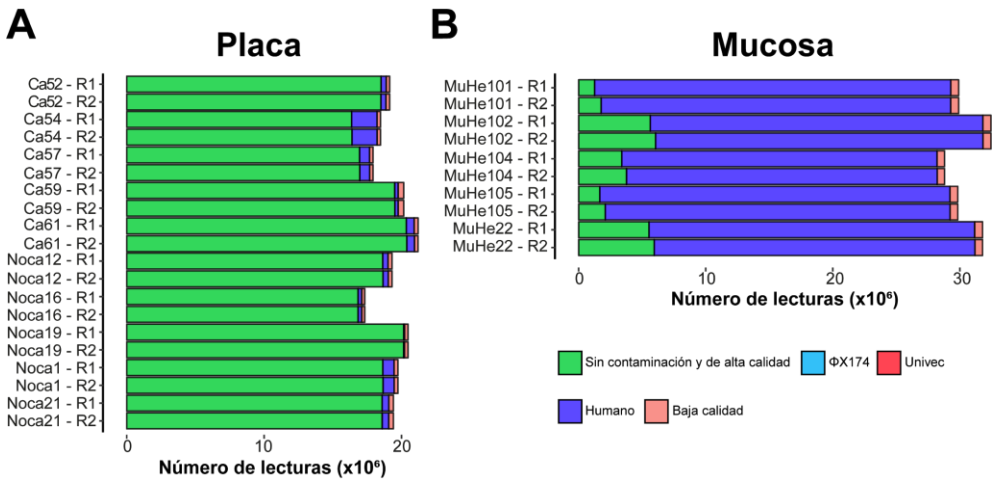

Figura 32. Representación gráfica de la cantidad de lecturas obtenidas por *HiSeq* (Illumina®) desde 15 microbiomas bucales. Se indica el número de lecturas procesadas en cada uno de los métodos de evaluación de calidad y búsqueda de secuencias contaminantes.

Usando la herramienta *CRISPRCasFinder*, encontramos que las lecturas no ensambladas de PacBio® y los *contigs* ensamblados desde las lecturas *HiSeq* contenían un total de 2.568 *CRISPRs* con 27.519 espaciadores de un tamaño medio de 36 pb. Casi todos ellos estaban en las secuencias de los microbiomas de placa dental y en los *contigs* ensamblados desde lecturas *HiSeq* (1.606 *CRISPRs*). Sólo 1.360 de estos espaciadores alineaban (95% de longitud alineada y 95% de identidad) con alguno de los *contigs* virales que habíamos encontrado previamente en los viromas de boca humana. Estos espaciadores estaban contenidos en 514 *CRISPRs* de 507 *contigs HiSeq* o lecturas de PacBio® distintos (**Tabla 8**), de los cuales, sólo 438 pudieron ser asignados a genomas bacterianos conocidos mediante *BLASTx* (*e-value* < $1x10^{-03}$). Esta estrategia nos permitió predecir hospedadores para 544 de nuestros 1.557 *contigs* virales.

	Totales	**Alinean con *contigs* virales**	**Alinean con *contigs* virales y asignados a genomas bacterianos**
Contigs* o lecturas PacBio con *CRISPRs	2.247	507	438
CRISPRs	2.568	514	445
Espaciadores	27.519	1.360	1.217

Tabla 8. Identificación de *CRISPRs* en secuencias procedentes de microbiomas de la boca y estudio de su relación con los *contigs* virales.

Siguiendo esta aproximación, los filos de bacterias infectados por estos 544 *contigs* virales fueron nuevamente *Actinobacteria*, *Firmicutes y Proteobacteria*, con un 20%, 6,68% y 6,6% del total de *contigs* virales. Las familias más representadas fueron *Actinomycetaceae* (8,8%), *Streptococcaceae* (5,46%) y *Corynebacteriaceae* (4%). Estos resultados coinciden en parte con los obtenidos previamente mediante BLASTx. Sin embargo, a diferencia de esta estrategia de predicción de hospedador, ninguno de los *CRISPRs* asociaba los *contigs* virales con miembros de las familias *Nocardiaceae, Mycobacteriaceae, Gordoniaceae* y *Microbacteriaceae* y, por el contrario, detectamos asociaciones nuevas con miembros de las familias *Corynebacteriaceae* y *Propionibacteriaceae*, del filo *Actinobacteria*, *Neisseriaceae* del filo *Proteobacteria*, y con bacterias del filo *Bacteroidetes* (0,9% de los *contigs* virales).

Teniendo en cuenta las cinco estrategias obtuvimos información acerca del filo del posible hospedador para 1.476 de los 1.557 *contigs* virales, incluyendo 423 de los 433 genomas completos o casi completos de bacteriófagos (se excluyeron 11 *contigs* del *megacluster* **AD** por ser virus eucarióticos). Las coincidencias entre los métodos con mayor número de resultados (BLASTx, Tetra y CRISPR) fueron altas a nivel de filo observando coincidencia en la predicción de hospedador para 291 de los 303 (96%) *contigs* con información obtenida simultáneamente por los métodos BLASTx y Tetra; para 159 de los 162 (98,1%) *contigs* con información por Tetra y CRISPR; y para 460 de los 529 (87%) *contigs* con información por BLASTx y CRISPR. Este nivel de coincidencia entre métodos fue mucho menor a nivel de familia, ya que solo 136 de los 509 (26,7%) *contigs* con predicción de hospedador obtenida mediante BLASTx y CRISPR coincidían a este nivel taxonómico. Finalmente, cabe señalar la coincidencia de las

distintas estrategias en asociar un gran número de *contigs* virales con bacterias de los filos *Actinobacteria*, *Proteobacteria* y *Firmicutes* principalmente. Muy pocos con *Bacteroidetes* y casi ninguno con *Fusobacteria*, pese a ser estos dos últimos filos constituyentes habituales de las comunidades de bacterias de la boca humana (Ly et al., 2014; Naidu et al., 2014; Pride et al., 2011a).

3.6. La mayoría de los 31 *megaclusters* de virus de la boca infectan por este orden *Actinobacteria, Proteobacterias* y *Firmicutes*

Las estrategias de predicción de hospedador bacteriano empleadas en el apartado anterior permitieron asociar de forma coherente muchos de los virus de los 30 *megaclusters* con unos pocos filos bacterianos. Así por ejemplo, los virus de los *megaclusters* **B**, **E**, **G**, **H**, **I**, **J**, **O**, **V** y **W** se asociaron casi exclusivamente con el filo *Actinobacteria* mediante al menos dos de los métodos empleados. Del mismo modo, los *megaclusters* **L**, **M**, **N**, **P**, **Q**, **S** y **AA** se asociaron al filo *Proteobacteria*, y los *megaclusters* **U**, **Y**, **Z** y **AC** a *Firmicutes* (**Fig. 33A**). Sin embargo, algunos de los *megaclusters* incluían *contigs* que podrían infectar bacterias de filos distintos. Un ejemplo de esto último fueron los *megaclusters* **A** y **C**, compuestos principalmente por *contigs* que infectan *Actinobacteria* y otros *contigs* que infectan *Firmicutes*, o *megaclusters* donde la asignación fue más confusa como es el caso de los *megaclusters* **D**, **R**, **X** o **AB**. A nivel de familia de hospedador obtuvimos resultados menos coherentes dentro de cada *megacluster*, debido a que la señal de los métodos es menos específica cuanto menor es el nivel taxonómico (**Fig. 33B**). Los *megaclusters* **A**, **B**, **C**, **G**, **H**, **I** y **K** presentaban *contigs* que podrían infectar al menos cuatro familias del filo *Actinobacteria*, lo que sugiere un amplio rango de hospedador para este *megacluster* de virus. Los *megaclusters* **M**, **N**, **P**, **Q**, **S** o **AA** se comportan de una manera similar pero con *Proteobacteria* y el *megacluster* **T** con *Firmicutes*. Por el contrario, existen cuatro *megaclusters* de virus que podrían infectar familias específicas, como son el caso del *megacluster* **E** (*Corynebacteriaceae*), **F** (*Propionibacteriaceae*), y los *megaclusters* **V** y **W** (*Actinomycetaceae*).

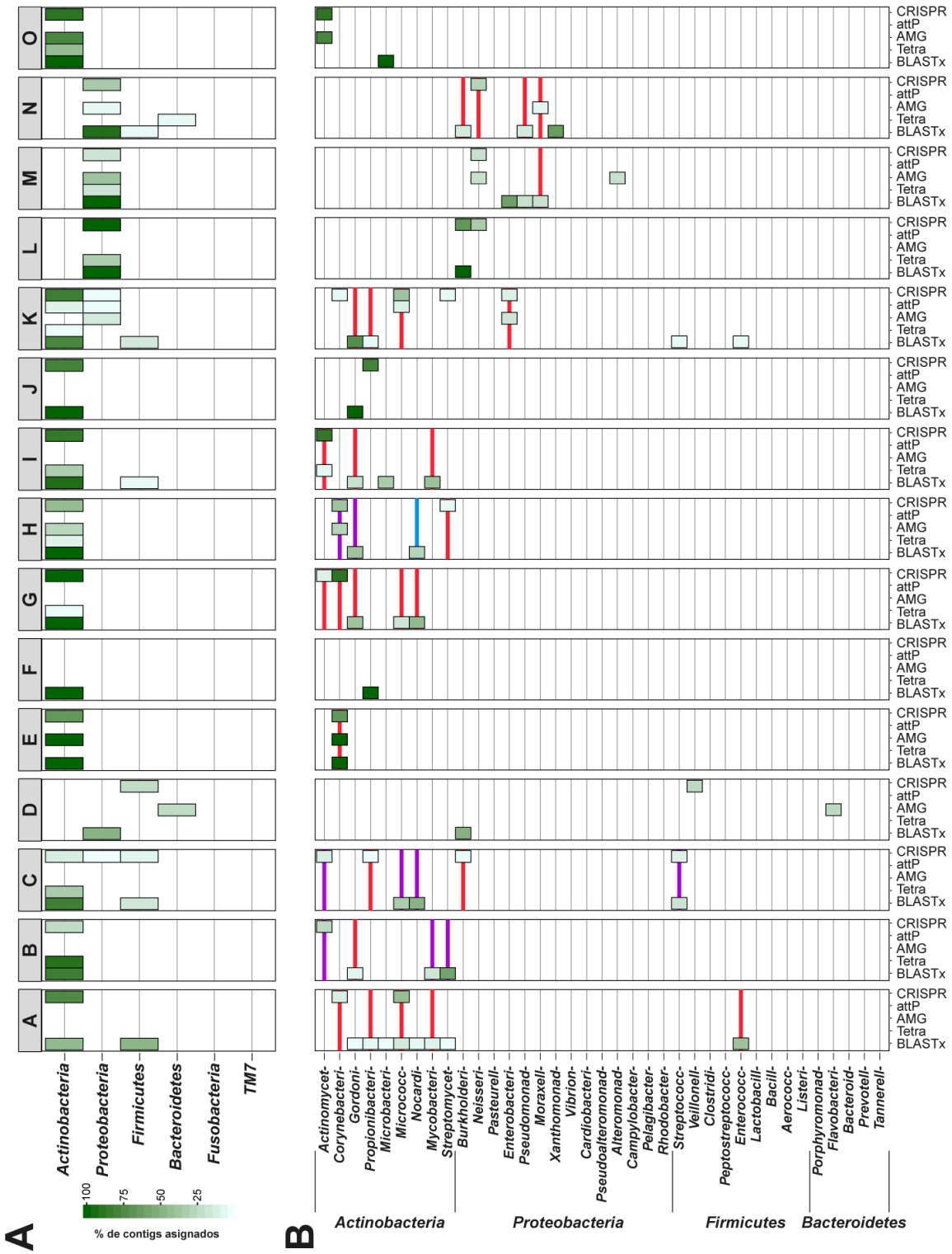

Figura 33. Predicción de hospedador para los 30 *megaclusters* de bacteriófagos de la boca. Los rectángulos que presentan un rango de color de blanco a verde indican el porcentaje de *contigs* de cada *megacluster* para el cual se ha predicho su hospedador mediante cinco métodos diferentes (BLASTx, Tetra, AMG, attP y CRISPR) a nivel de (A) filo o (B) familia. Las líneas rojas indican aquellos *megaclusters* de virus asociados con familias bacterianas en los que al menos un *contig* viral contiene genes relacionados con lisinas, las líneas azules con holinas y las líneas moradas con lisinas y holinas.

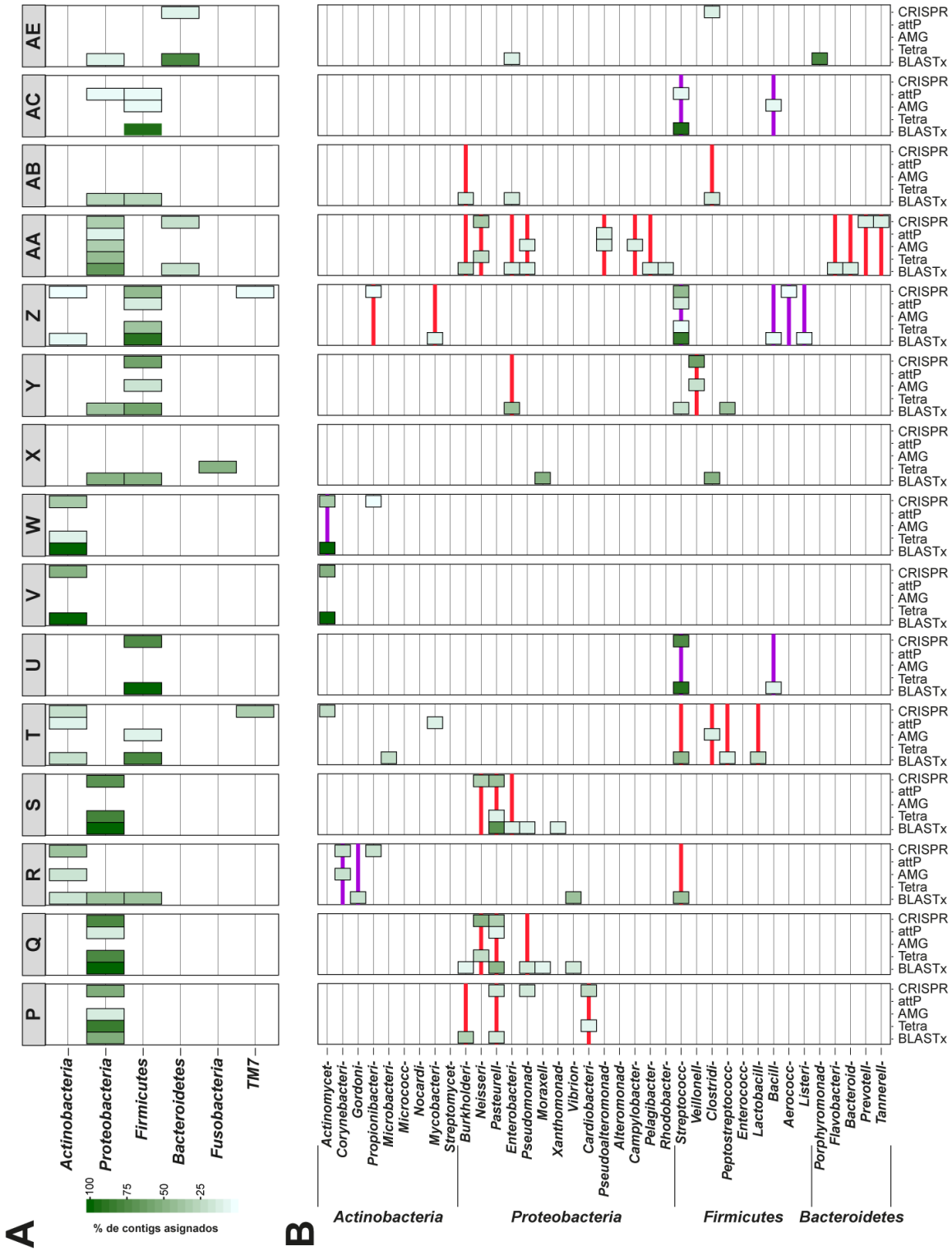

Figura 33. Continuación.

Una de las herramientas de predicción de hospedador que proporcionó más información fue la basada en el hospedador del virus más cercano encontrado por similitud de secuencia en las bases de datos. La representación gráfica del hospedador predicho con este método para los 1.557 *contigs* virales o los 444 *contigs* virales potencialmente completos (**Fig. 34**), puso de manifiesto que muchos de los *megaclusters* próximos en ambas representaciones compartían la característica de infectar hospedadores del mismo filo bacteriano, lo que sugiere que la clasificación taxonómica a nivel de filo del hospedador es un factor clave para explicar la diversidad genética de los bacteriófagos de la boca.

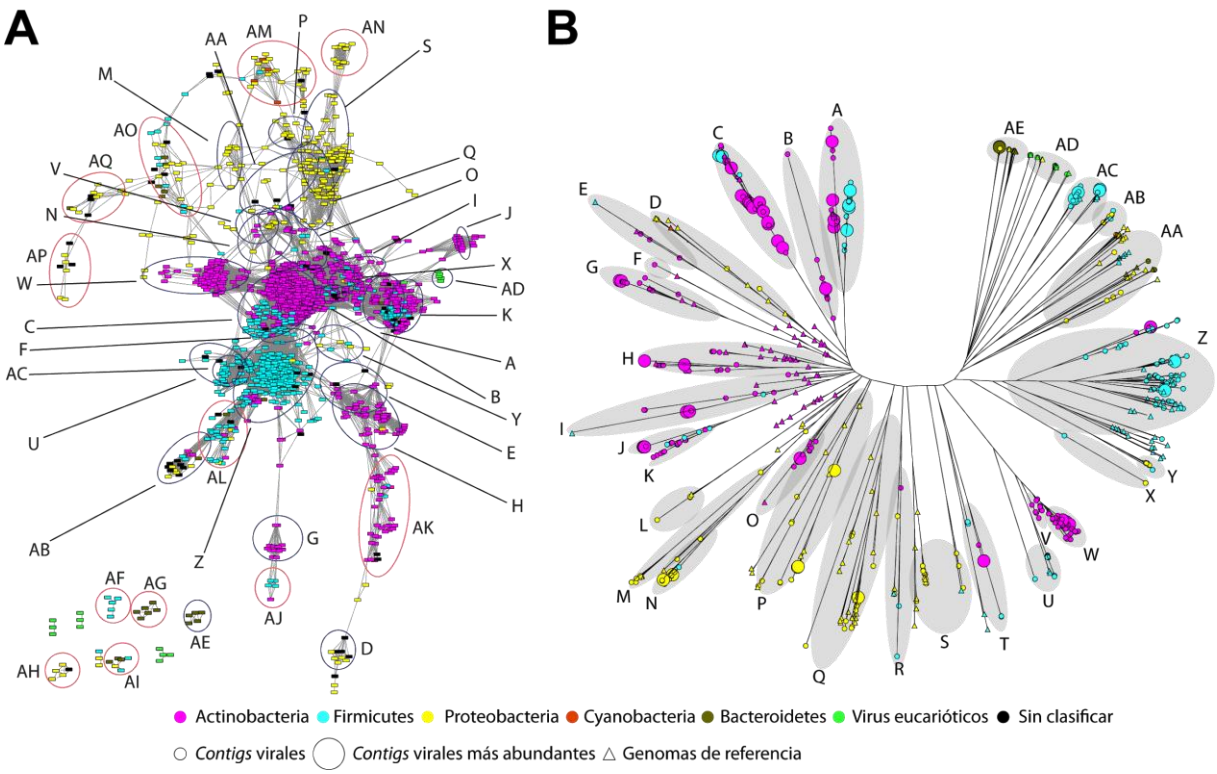

Figura 34. Predicción de hospedador para los *contigs* virales de la cavidad bucal según el hospedador del virus más cercano conocido encontrado por BLASTx. Se muestra (A) una red con los 1.557 *contigs* virales agrupados con *MCL* en función de sus distancias obtenidas mediante *BLASTn*, y (B) un árbol proteómico *Neigbour Joining* generado desde una matriz de distancias *Dice*-modificada calculadas mediante *tBLASTx* de para los 444 *contigs* virales potencialmente completos y 205 genomas de referencia relacionados. El color de los nodos indica el filo del hospedador del virus más cercano en la base de datos para cada *conti*g. Los círculos de mayor tamaño indican su pertenencia a los diez *contigs* más abundantes de cada viroma. Los 31 *megaclusters* identificados a partir del árbol proteómico se indican también en la red que incluye los 1.557 *contigs*. Adicionalmente en esta última se muestran también los 12 *megaclusters* sin representantes potencialmente completos.

4. La composición de las comunidades bacterianas es diferente entre mucosa y placa dental, y no correlaciona en términos de abundancia con los filos infectados por la comunidad de bacteriófagos

Para estudiar la composición de las comunidades bacterianas asociadas a los viromas estudiados, se amplificó la región comprendida entre las posiciones 341 y 805 del gen marcador ARNr 16S desde los sedimentos celulares de 16 muestras de mucosa y 15 de placa dental. Entre estas muestras se incluyen aquellas cuyos viromas han sido estudiados en apartados anteriores de esta tesis. Los amplicones obtenidos se secuenciaron en equipos *MiSeq* de Illumina® generando entre 108.615 y 341.102 secuencias solapantes 2x300pb por muestra. Las secuencias de alta calidad se organizaron en unidades taxonómicas operativas (*OTUs*) utilizando *Qimme*, que se clasificaron a continuación por comparación con la base de datos SILVA. Las comunidades de bacterias de estas muestras estaban dominadas por *OTUs* de los filos *Firmicutes*, *Proteobacteria, Actinobacteria*, *Bacteroidetes y Fusobacteria* en orden decreciente de abundancia, con una representación muy baja de *Spirochaetes* y *TM7* (**Fig. 35 A y C**). Encontramos diferencias en la distribución de estos filos por ambiente. Así, las proporciones de *Firmicutes* y *Proteobacterias* fueron más elevadas en mucosa (51,43% y 22% de media, respectivamente) que en placa dental (33% y 12,5%, respectivamente), y por el contrario, los filos *Actinobacteria* y *Fusobacteria* fueron superiores en placa dental (15,32% y 12%, respectivamente) que en mucosa (4,84% y 4,91%, respectivamente). Sin embargo, sólo las diferencias por ambiente del filo *Firmicutes* resultaron estadísticamente significativas en un test Mann-Whitney. Pese a que la mayoría de los *contigs* virales de la boca (incluyendo varios de los más abundantes), infectan bacterias del filo *Actinobacteria*, este filo es sólo el cuarto más abundante en *OTUs* asignadas. Por otro lado, las bacterias del filo *Firmicutes*, que son las claras dominadoras de los ecosistemas bucales, parecen ser hospedadores casi exclusivos de un grupo reducido de *megaclusters* virales (**U**, **Y**, **Z** y **AC**) y del grupo de bajo %CGs en el *megacluster* **C**. A nivel de familia, encontramos que las bacterias más representadas en la boca humana fueron en orden decreciente *Streptococcaceae* (15,88% en placa y 39,22% en mucosa), *Pasteurellaceae* (5,43% en placa y 17,86% en mucosa), *Veillonellaceae* (11,24% en placa y 4,57% en mucosa) y *Micrococcaceae* (7,62% en placa y 3,96% en mucosa). Varias familias mostraron diferencias estadísticamente significativas en su distribución entre ambientes. Así, las familias *Streptococcaceae* y *Gemellaceae* del filo *Firmicutes y* la familia *Pasteurellaceae* del filo *Proteobacteria* eran más abundantes en mucosa que en placa dental. Por el contrario, las familias *Corynebacteriaceae* y *Actinomycetaceae* del filo *Actinobacteria* y las fusobacterias de la familia *Leptotrichiaceae* y los bacteroidetes de la familia *Flavobacteriaceae* eran más abundantes en placa dental que en mucosa.

Para evaluar el impacto del tipo de ambiente bucal y el estado de salud en la composición de las comunidades de bacterias, estudiamos la distribución de los microbiomas en un sistema de ordenación *NMDS* (**Fig. 36**). Coincidiendo con las diferencias por ambiente encontradas para algunos de los filos y familias más abundantes de la cavidad bucal, el estudio de la comunidad de bacterias mostró una clara separación de los microbiomas de placa dental y de mucosa que resultó significativa estadísticamente

en un test PERMANOVA (*p-value* < 0,001). Sin embargo, y en clara correlación con los resultados obtenidos con los viromas, no detectamos diferencias significativas entre los microbiomas en los estados de salud y enfermedad.

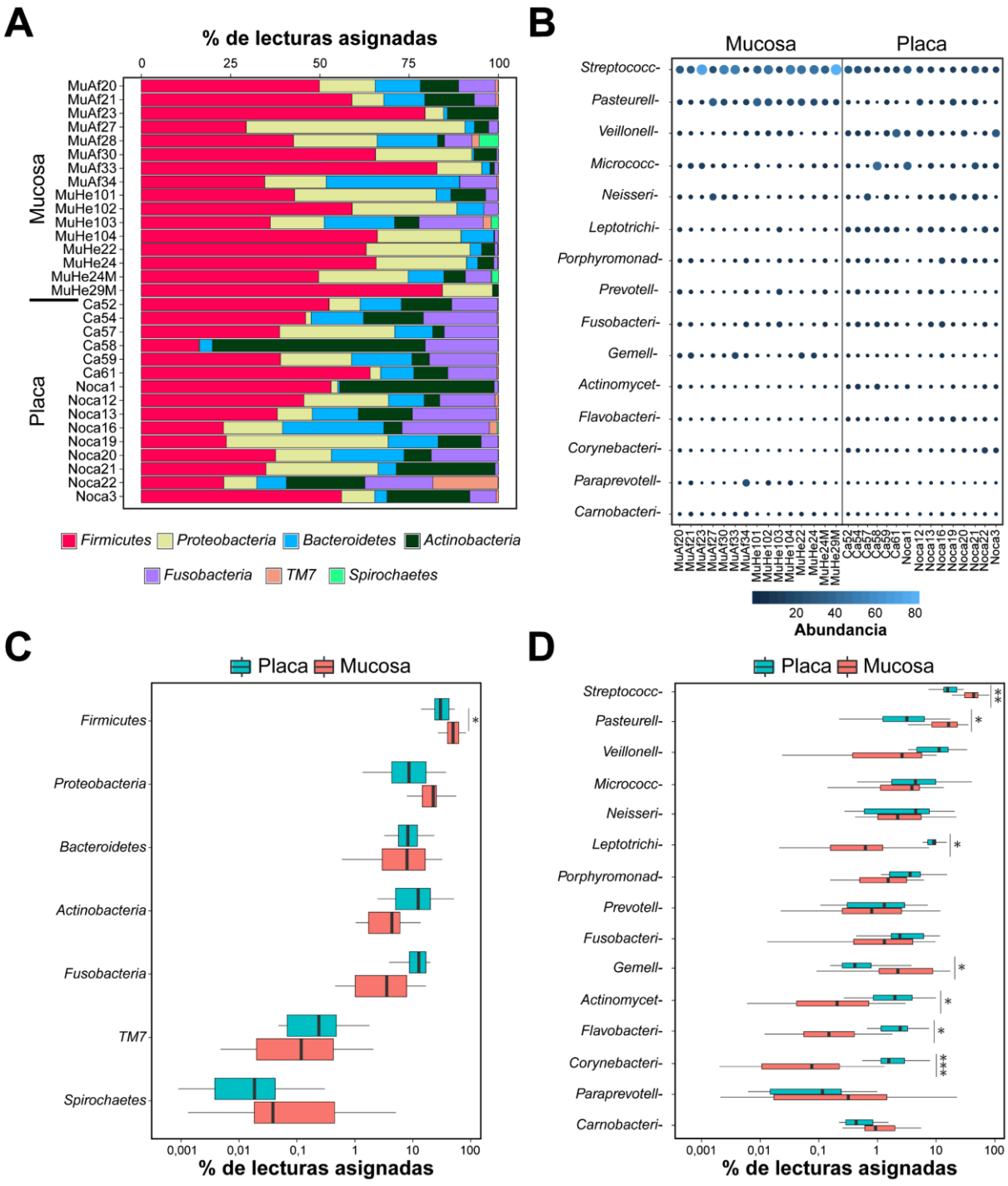

Figura 35. Estudio de la composición taxonómica de las comunidades de bacterias de la boca humana. (A y C) Proporciones relativas de las *OTUs* asociadas a filos y (B y D) a familias bacterianas. En A y B se muestran los datos individualizados a viromas y en C y D se muestran diagramas de cajas para los filos y familias más abundantes respectivamente. El tamaño e intensidad de color azul en B indica la abundancia relativa de las *OTUs* asignadas a una determinada familia. El análisis estadístico de las diferencias en abundancia de *OTUs* entre ambientes se hizo mediante un test Mann-Whitney aplicando una corrección *FDR*. * $p < 0,05$; ** $p < 0,01$; y *** $p < 0,001$.

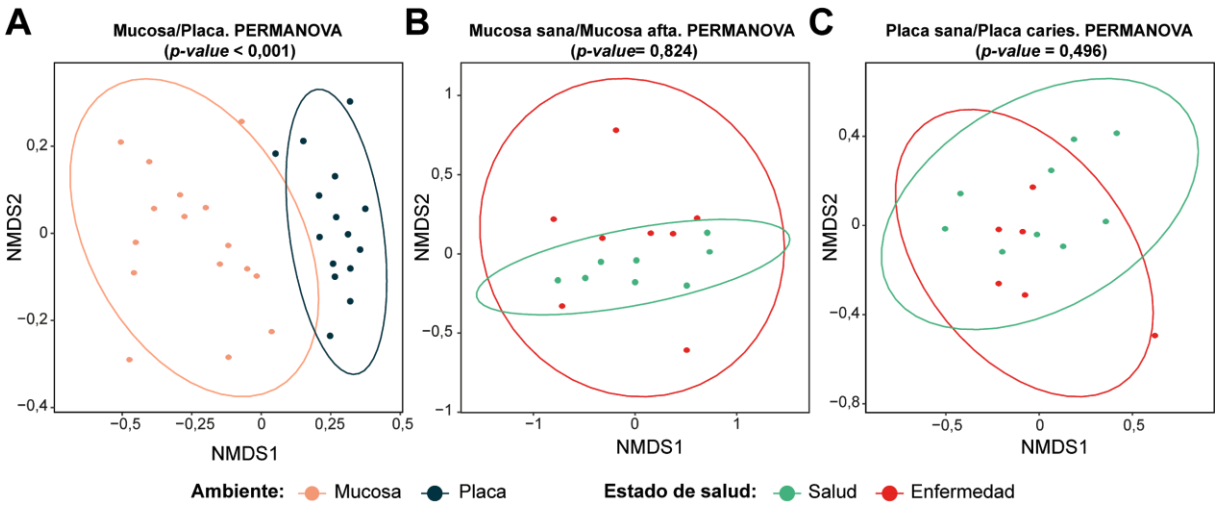

Figura 36. Sistema de ordenación bidimensional de los microbiomas de placa dental y mucosa oral humana. El perfil de abundancias de *OTUs* se utilizó para calcular las disimilitudes de Bray-Curtis que se representaron en sistemas de ordenación *NMDS*. Se indica también la significancia estadística de las diferencias entre grupos definidos por ambiente (A) o estado de salud (B y C), analizada mediante un test PERMANOVA.

5. Los virus de la cavidad bucal humana contienen un amplio arsenal de genes que codifican por lisinas y holinas

Durante la infección lítica, los bacteriófagos emplean lisinas específicas para romper la pared de peptidoglicano de sus hospedadores y holinas para hacer canales en la membrana plasmática que permitan el acceso de las lisinas a la pared de peptidoglicanos. El uso en clínica de estas proteínas representa una herramienta terapéutica prometedora por su especificidad y por el agotamiento del repertorio de antibióticos en un escenario de continua expansión de resistencias frente a los antibióticos disponibles. En este último apartado de la tesis doctoral hemos querido estudiar de forma preliminar el repertorio de genes implicados en estas funciones y contenidos en los bacteriófagos de la cavidad bucal humana. Para ello, seleccionamos aquellos genes relacionados por similitud de secuencia (*BLASTx* contra la base de proteínas *PHAST*, *e-value* < $1x10^{-03}$) con lisinas, endolisinas, amidasas, hidrolasas de la pared celular, endopeptidasas y holinas. De los 1.557 *contigs* virales, 446 distribuidos en 23 *megaclusters* contenían un total de 500 genes que codificaban por posibles lisinas y 261 *contigs* (con representantes en nueve *megaclusters*) contenían 304 holinas (**Fig. 25B**). Aunque la mayor parte de los *megaclusters* con holinas contenía también lisinas, el *megacluster* **H** contenía cuatro *contigs* con holinas y sin lisinas. Por otro lado, en los *megaclusters* **D**, **F**, **J**, **L**, **O**, **V**, **X** y **AE** no pudimos identificar ningún gen que codificara por lisinas u holinas. Gracias a la predicción de rango de hospedador que hemos abordado anteriormente para estos *contigs* viras virales, podemos proponer sobre qué filos y/o familias de bacterias podrían actuar estas lisinas y holinas. Así, las 500 lisinas y 304 holinas de nuestros *contigs* virales podrían estar implicadas en la lisis de hasta 37 familias de bacterias de cuatro filos diferentes, principalmente *Actinomycetaceae*, *Nocardiaceae*, *Mycobacteriaceae*, *Streptomycetaceae* y *Streptococcaceae*. La necesidad actual de buscar nuevas herramientas específicas para combatir

enfermedades bacterianas, o intervenir ecosistemas microbianos alterados, justifica el estudio del arsenal lítico de las comunidades de bacteriófagos. La gran diversidad de familias sobre las que estas lisinas podrían actuar abre la puerta a futuras investigaciones destinadas a probar su posible eficacia en la restauración de ecosistemas bucales saludables como complemento o alternativa a los antibióticos.

DISCUSIÓN

La cavidad bucal representa uno de los ecosistemas microbianos más diverso e importante para la salud humana. Esto es resultado de su estrecha relación con la microbiota intestinal, y de ser el principal portal de entrada de microorganismos patógenos. A diferencia de la comunidad bacteriana, la población de virus de la boca ha sido pobremente estudiada. Por ello, en este trabajo hemos querido profundizar en el conocimiento de estas comunidades, así como en las posibles interacciones que mantienen con sus hospedadores bacterianos. Para ello, hemos estudiado 48 viromas procedentes de saliva, mucosa oral y placa dental. Este trabajo representa el mayor estudio metagenómico de la comunidad de virus de la boca realizado hasta la fecha (43,4 Gpb), triplicando en profundidad de secuencia por viroma al resto de estudios publicados (1.509.608 lecturas y 906 Mpb por viroma de media en esta tesis frente a 151.200-605.754 lecturas y 61-271 Mpb por viroma en trabajos previos) (Abeles et al., 2015, 2014; Ly et al., 2016, 2014; Pérez-Brocal y Moya, 2018; Pride et al., 2011a).

1. Sesgos experimentales en el estudio de los viromas humanos

Las comunidades naturales de virus están formadas por un rango amplio de partículas virales con características morfológicas y químicas muy dispares. Esto impide el establecimiento de protocolos estándar de purificación de genomas virales que reflejen fielmente la composición de las comunidades originales de virus (Thurber et al., 2009; Willner et al., 2011). Algunos estudios han evaluado el impacto de estas fuentes de sesgo sobre comunidades sintéticas de virus de composición conocida. Sin embargo, en muchos casos, la elección de sus constituyentes no ha sido muy acertada al no reflejar el amplio rango de morfologías, de tamaños y de tipos de genomas que hay en la naturaleza, o presentan distribuciones muy desbalanceadas. Estas limitaciones pueden dificultar la identificación de algunas fuentes de sesgos o infravalorar su impacto (Castro-Mejía et al., 2015; Daly et al., 2011; Hall et al., 2014; Kleiner et al., 2015; Kohl et al., 2015; Lewandowska et al., 2017; Li et al., 2015).

1.1. Sesgos en el enriquecimiento de partículas virales

En esta tesis hemos utilizado una comunidad sintética compuesta de siete virus de ADN (diversos en cuanto al tipo de material genético y morfología) para explorar los sesgos generados por protocolos sencillos de enriquecimiento de virus y por varios métodos de amplificación al azar desde cantidades bajas de ADN molde. La cuantificación del material genético de estos virus por *PCR* cuantitativa después de un tratamiento con nucleasas, permitió evaluar sólo aquellos genomas virales protegidos por cápsidas o envueltas intactas. Este método representa una notable mejora sobre estudios previos (Conceição-Neto et al., 2015; Kleiner et al., 2015) en la preparación de comunidades virales sintéticas balanceadas. Esto es debido, por un lado, a que la *PCR* cuantitativa, a diferencia de otras técnicas más empleadas y basadas en la cuantificación con reactivos de tinción de ADN (Roux et al., 2016), permite

una correcta estimación de genomas de ADN de cadena sencilla, y por otro, a que el tratamiento con nucleasas evita la cuantificación de genomas de virus con una estructura parcialmente dañada como resultado de la purificación o el tipo de preservación. El seguimiento de las cantidades absolutas y relativas de los virus de esta comunidad sintética de composición conocida nos permitió ver que dos pasos destinados a reducir la proporción de organismos celulares, como son la centrifugación a baja velocidad y la filtración en 0,22μm, provocaban una drástica caída en el número total de genomas del virus Vaccinia WR, que es el virus más grande de la comunidad. La utilización de filtros de 0,45μm atenuó esta pérdida, lo que correlaciona con estudios previos que demuestran que la filtración en 0,45μm,, duplica la cantidad de ADN viral obtenida con filtros de 0,22μm (Hoyles et al., 2014), o mejorar la representación de virus grandes de las familias *Phycodnavirus, Mimiviridae* y *Herpesviridae* (Conceição-Neto et al., 2015; Lewandowska et al., 2017; Li et al., 2015; López-Bueno et al., 2009) o bacteriófagos Jumbo (Yuan y Gao, 2017). La reducción de la velocidad de centrifugación inicial podría valorarse en estudios futuros para reducir aún más la pérdida de virus grandes.

En nuestro protocolo, el colchón de iodixanol permite la concentración eficiente de las partículas virales sin someterlas a fuerzas físicas que pudieran afectar a la estabilidad de sus estructuras y sin afectar a la composición de la comunidad viral. Este paso, combinado con el tratamiento de nucleasas, podría resultar también muy eficaz para eliminar ADN libre de origen celular, siendo una buena alternativa a los gradientes de CsCl, los cuales separan muy bien las comunidades de virus de las de bacterias y del ADN libre, pero introducen importantes sesgos en su composición (Castro-Mejía et al., 2015; Kleiner et al., 2015; Thurber et al., 2009).

Tras estos resultados propusimos un protocolo simple de enriquecimiento de partículas virales en muestras de cavidad bucal, basado en dos pasos de centrifugación a baja velocidad seguido de una filtración en 0,45μm y el uso optativo de colchones de iodixanol. Ante la previsible presencia de cantidades bajas de contaminación bacteriana por el uso de filtros de 0,45μm, proponemos la valoración del nivel de contaminación mediante *PCR* del gen marcador de ARNr 16S.

1.2. Prevención de contaminación con ADN bacteriano y humano

En relación a la eliminación de contaminación bacteriana, otros autores han observado una eficiencia similar de eliminación entre filtros de 0,45 μm y 0,22 μm (Klieve y Swain, 1993). Aunque no hemos valorado el rendimiento de la purificación de virus frente a bacterias como se ha hecho en otros estudios (Kleiner et al., 2015), en nuestras manos, dos centrifugaciones consecutivas a baja velocidad, combinadas con la filtración en 0,45μm, redujo drásticamente el número de unidades formadoras de colonia de tres cultivos puros de bacterias, incluyendo bacterias de pequeño tamaño, y al menos 10 veces el contenido de genes que codifican por ARNr 16S en viromas de saliva. Sin embargo, la aplicación de este protocolo a 55 muestras de mucosa bucal y placa dental, sin colchones de iodixanol,

resultó en un nivel de contaminación bacteriana superior al esperado en alguna de las muestras. Esta contaminación pudo atribuirse a bacterias de pequeño tamaño como las que forman el filo *TM7*, cuyos miembros cultivados podrían teóricamente atravesar los filtros de 0,45μm. En alguno de los viromas más contaminados pudimos incluso ensamblar el genoma posiblemente completo de una de estas bacterias, que con un tamaño de 734 kpb supera al genoma de 705 kpb de una bacteria de este filo aislada de la boca humana (He et al., 2015). Las bacterias del filo *TM7*, pese a que apenas tienen representantes cultivados, son prácticamente ubicuas en una gran variedad de ambientes naturales (Hugenholtz et al., 2001) y forman parte de un grupo grande de filos que podría encerrar hasta un 15% de la diversidad total de bacterias (Brown et al., 2015; Hug et al., 2016). Su naturaleza simbionte con otras bacterias podría explicar sus limitaciones metabólicas y su pequeño tamaño, que es inferior al de algunos virus de gran tamaño como los mimivirus (Fischer et al., 2010).

Para reducir el impacto de esta contaminación bacteriana en nuestro estudio de los viromas de la cavidad bucal hemos tomado una serie de medidas como son: (i) no secuenciar 15 de las 56 muestras procesadas por presentar un alto nivel de contaminación bacteriana superior al 15% (estimando por *PCR* del gen que codifica por ARNr 16S), (ii) no analizar seis viromas que presentaban un nivel aún importante de lecturas relacionadas con genomas bacterianos de las bases de datos, (iii) limitar todos los análisis de los viromas a *contigs* de un tamaño y cobertura elevados ya que es razonable pensar que las pocas secuencias de origen bacteriano presentes en los viromas restantes se ensamblen en *contigs* de tamaño pequeño y cobertura baja, (iv) diseñar un protocolo muy exigente de asignación taxonómica de los *contigs* virales en el que se descarten aquellos que presenten un porcentaje alto de genes relacionados con genomas bacterianos de las bases de datos, y bajo de genes de origen viral.

Otra fuente de contaminación conocida en la secuenciación de metagenomas de virus procedentes de muestras humanas es el genoma humano. En nuestros estudios, esta fuente de contaminación era mayor en los viromas de mucosa bucal, donde existe un contacto directo con células humanas y nos obligó a eliminar de posteriores análisis dos viromas adicionales. El uso de centrifugación y filtración excluye la contaminación con células eucariotas intactas, por lo que el origen de esta contaminación debe ser ADN libre de células humanas rotas que no se han digerido completamente por las nucleasas, quizás por su asociación con histonas (Hauer y Gasser, 2017). Una mayor concentración de nucleasas, el uso de mezclas más complejas de estas enzimas durante el proceso de purificación o el uso de colchones de iodixanol, podría disminuir los niveles de este tipo de contaminación en los viromas humanos.

1.3. Impacto de la amplificación al azar en la composición de los viromas

Debido a las limitaciones para obtener muestras humanas y al pobre rendimiento de los protocolos de enriquecimiento de virus, la cantidad de ADN viral obtenida es a menudo insuficiente para preparar librerías de secuenciación masiva. Para solucionar este problema, se recurre a protocolos de

amplificación aleatoria del ADN que también pueden alterar las proporciones relativas de sus constituyentes, con un nivel de sesgo directamente proporcional al grado de amplificación (Binga et al., 2008; Dean et al., 2002; Direito et al., 2014; Lasken, 2009; Ning et al., 2015). En esta tesis doctoral hemos intentado avanzar en el conocimiento de estos sesgos comparando mezclas artificiales o naturales de genomas virales antes y después de la amplificación. En los viromas de saliva humana la amplificación con *MDA* desde 10 pg de ADN molde genera sesgos estocásticos, con patrones de *contigs* variables entre réplicas, y un gran número de *contigs* con una representación alterada. Además, estos *contigs* no están relacionados con los dos sesgos sistemáticos conocidos de *MDA*, como son la sobreamplificación de genomas pequeños circulares o la peor amplificación de regiones de ADN con valores extremos de %CGs.

El incremento en la cantidad de molde al rango de nanogramos en *MDA* redujo drásticamente el número de *contigs* con una representación alterada, mejoró los índices de correlación entre los viromas amplificados y el no amplificado, y dio lugar a patrones de cobertura más homogéneos. Todo ello contribuyó a que la distancia en los sistemas de ordenación bidimensional del viroma sin amplificar a los viromas amplificados desde 1 ng fuera menor que la distancia con los amplificados desde 10 pg. Otro de los parámetros analizados en esta tesis fue el tiempo de extensión en la amplificación con *MDA*. Aunque los protocolos de *MDA* sugieren reducir el tiempo de amplificación al mínimo necesario para obtener la cantidad deseada de ADN, en nuestras manos, la extensión de la reacción a 10 h desde 10 pg o 1 ng de molde, no parece tener ninguna incidencia en la composición de la comunidad.

Estudios previos habían demostrado la presencia de sesgos sistemáticos en las poblaciones de bacterias o virus amplificados desde cantidades superiores a 1 ng con *MDA* (Binga et al., 2008; Dean et al., 2002; Direito et al., 2014; Lasken, 2009; Rosseel et al., 2013; Solonenko et al., 2013; Yilmaz et al., 2010). Uno de los sesgos sistemáticos mejor estudiados en *MDA* es la sobreamplificación de genomas circulares pequeños. Este sesgo fue cuantificado en un incremento de 56x y 212x en las abundancias relativas de dos genomas circulares de menos de 2 kpb de viromas de suelo (Kim et al., 2008), y en 5,7x y 72,6x en otros genomas también circulares pero ligeramente más grandes (5,3 y 6,1 kpb) de comunidades sintéticas (Roux et al., 2016). En nuestras comunidades sintéticas, observamos una sobrerrepresentación menor de los genomas circulares del bacteriófago M13 (3,2-7,2x) y de PCV2a (3,2-14,7x). La mayor sobreamplificación de PCV2a sobre M13 podría deberse a la menor probabilidad de cortes puntuales (*nicking*) en las moléculas circulares más pequeñas. Sin embargo, en los viromas de saliva humana amplificados con *MDA* desde 1 ng de molde, observamos una enorme variabilidad en el grado de sobrerepresentación de los genomas pequeños circulares. Esta variación no estaba relacionada ni con diferencias en el contenido de CG ni con la longitud de los genomas, lo que sugiere la participación de otros factores aun por identificar. Aunque muchos de los *contigs* más sobreamplificados por *MDA* se correspondían a genomas circulares pequeños, solo dos se encontraban entre los 200 más abundantes de la comunidad, lo que indica que su influencia en el perfil global de los viromas es menor.

Apoyando este resultado, las disimilitudes de Bray-Curtis entre los viromas sin amplificar y los amplificados, y su localización relativa en los sistemas de ordenación, apenas variaron cuando se eliminaron del estudio los *contigs* circulares pequeños (datos no mostrados).

La baja influencia de la sobreamplificación de genomas pequeños circulares en las diferencias observadas entre viromas amplificados y no amplificados con *MDA* en esta tesis, junto con la ausencia de mejoras al preparar librerías desde mezclas de productos de amplificación independientes procedentes del mismo molde, observada por otros (Marine et al., 2014), nos llevó a indagar más acerca de otras posibles fuentes de sesgos sistemáticos. En esta tesis demostramos que *MDA*, y también *SISPA*, inducen sesgos sistemáticos con sobreamplificación de secuencias y *contigs* con un porcentaje de CGs en el rango de 45-60%, posiblemente por la dificultad de amplificar regiones de ADN con un contenido muy bajo y muy alto de CGs. Estos resultados concuerdan con lo anteriormente propuesto para métodos de amplificación al azar basados en *MDA* (Abulencia et al., 2006; Arriola et al., 2007; Bredel et al., 2005; Ellegaard et al., 2013; Han et al., 2012; Rhee et al., 2016; Yilmaz et al., 2010), *LASL* (Aird et al., 2011; Duhaime et al., 2012; Hoeijmakers et al., 2011; Solonenko et al., 2013) y en general, para cualquier método basado en amplificación por *PCR* (Arezi y Hogrefe, 2009; Benita et al., 2003; Mamedov et al., 2008; Pinto y Raskin, 2012). Los problemas en la accesibilidad de la polimerasa a estas regiones con contenido extremo de CGs, o la terminación prematura de la copia al principio de las estructuras secundarias ricas en CGs, se ha propuesto como la principal causa de este sesgo (Arezi et al., 2003; Bredel et al., 2005). En nuestros viromas, esta fuente de sesgo afecta a la mayor parte de las secuencias y *contigs* de los viromas amplificados por *MDA* y *SISPA* desde 1 ng, sugiriendo que ésta podría ser la principal causa de sesgo sistemático. También postulamos que las diferencias entre estos los métodos de amplificación (*MDA* y *SISPA*) podrían deberse a las distintas capacidades que tienen para amplificar regiones de alto y bajo contenido de CGs. Así, los viromas obtenidos con *SISPA* mostraron un sesgo negativo en secuencias con un porcentaje de CGs < 40%, mientras que los basados en *MDA* (mediante hexámeros aleatorios) presentaban más problemas amplificando secuencias con un porcentaje de CGs > 65%. Lo utilización de un método de amplificación *MDA* alternativo, basado en el cebado con oligonucleótidos aleatorios sintetizados por una actividad ADN primasa presente en la reacción (TruePrime™), permitió una mejor amplificación de secuencias de CGs bajo que *SISPA* y una mejor amplificación de secuencias con alto contenido de CGs que *MDA* basado en hexámeros aleatorios. Estas mejoras contribuyen al solapamiento casi perfecto de los viromas MDA_T1 y Unamp1 en los sistemas de ordenación basados en disimilitudes Bray-Curtis o índices de Sørensen. La mejor capacidad para amplificar regiones de CGs altas o bajas de este sistema de amplificación *MDA* con primasas había sido ya sugerida previamente (Direito et al., 2014), aunque un estudio reciente, si bien demuestra la reducción de sesgo con respecto a *MDA* y hexámeros random, sugiere que esta mejora no es debida a la mejor amplificación de regiones con alto contenido en CGs (Picher et al., 2016).

En esta tesis demostramos que la dificultad para amplificar regiones de contenido extremo de CGs es una fuente principal de sesgo sistemático cuando se emplean nanogramos de molde. Sin embargo, también hemos identificado en los viromas de saliva varios *contigs* grandes o no circulares y con un contenido de CGs medio que mostraban una alteración importante en su representación con respecto al viroma no amplificado. De forma similar, en las comunidades sintéticas amplificadas por *MDA*, también observamos un fuerte sesgo negativo hacia el genoma de 5 kb de cadena sencilla lineal del virus MVMp, pese a contener un porcentaje CG del 43%. Estos resultados sugieren la concurrencia de varias fuentes de sesgo simultáneamente sobre las comunidades de virus estudiadas, y que desconocemos la naturaleza de alguna de ellas.

Otra de las manifestaciones del sesgo introducido por los sistemas de amplificación aleatoria es la obtención de perfiles de cobertura irregulares a lo largo de los genomas amplificados (DePew et al., 2013; Rosseel et al., 2013). En los viromas de saliva, la comparación de tres parámetros de regularidad de la cobertura a lo largo de las secuencias (curvas de Lorenz, coeficientes de variación de cobertura y la correlación de Pearson respecto a la cobertura del genoma sin amplificar), demostró el peor comportamiento de *SISPA* en comparación con *MDA*. Este resultado se atribuyó a la presencia de picos de muy alta cobertura en muchos *contigs* de viromas amplificados por *SISPA*. Aunque algunos de estos picos se han relacionado previamente con una unión preferente de la parte constante de los oligonucleótidos pseudo-degenerados (Karlsson et al., 2013; Victoria et al., 2009), solo una pequeña proporción de los picos de alta cobertura de los *contigs* de los viromas de saliva obtenidos por *SISPA* eran específicos del oligonucleótido empleado y contenían secuencias próximas parecidas a la secuencia del oligonucleótido. Uno de los hallazgos interesantes de esta tesis es que algunos de estos picos de alta cobertura se obtenían con los tres oligonucleótidos empleados y se encontraban en regiones de baja complejidad lingüística. Las secuencias de baja complejidad se evitan durante el diseño de oligonucleótidos de *PCR* (Wang y Seed, 2003), o se filtran durante las búsquedas de *BLAST* para prevenir uniones inespecíficas (Morgulis et al., 2006). Dichas secuencias, se han asociado también con picos de *ChiP-s* que son falsos positivos debido a repeticiones colapsadas (Pickrell et al., 2011). En nuestro estudio, hemos descartado que estos picos estuvieran relacionados con (i) algún problema metodológico del alineamiento al llevar a cabo éste en condiciones estrictas, (ii) el colapso de regiones repetidas, o con (iii) una mayor abundancia de moléculas de ADN molde de estas regiones ya que estos picos no se observan los viromas no amplificados. Aunque se necesita más investigación para comprender las bases moleculares de este tipo de sesgo en zonas de baja complejidad lingüística, hipotetizamos que puede estar relacionado con una sobrerrepresentación de oligonucleótidos de baja complejidad. Estos oligonucleótidos tienden a formar más fácilmente estructuras diméricas que pueden servir como molde durante las rondas de amplificación por *PCR*, incrementando así su abundancia relativa sobre otros oligonucleótidos con una mayor complejidad lingüística. Apoyando esta hipótesis, hemos visto que más del 80% de las lecturas que alineaban con estos picos de alta cobertura en regiones

de baja complejidad lingüística presentaban dímeros de oligonucleótidos en sus extremos (datos no mostrados). Este nuevo sesgo sistemático y específico de *SISPA*, junto con las mayores dificultades de amplificación de regiones con bajo contenido de CGs, podría explicar el peor comportamiento de *SISPA* con respecto a *MDA* en los sistemas de ordenación bidimensional.

En los últimos años se está haciendo un gran esfuerzo para comprender los sesgos experimentales que se introducen durante la preparación de metagenomas de virus y que dificultan la obtención de conclusiones cuantitativas desde estos estudios (Direito et al., 2014; Duhaime et al., 2012; Ellegaard et al., 2013; Picher et al., 2016; Rhee et al., 2016; Wu et al., 2006; Zong et al., 2012). El trabajo presentado en esta tesis, junto con los estudios mencionados, ayudará a preparar viromas que reproduzcan de una forma más exacta la composición de las comunidades virales naturales, mejorando la reproductividad de los estudios y la obtención de conclusiones fiables, por ejemplo, en estudios longitudinales. Sin embargo, nuestros estudios también sugieren que el impacto de los sesgos debido a la amplificación inespecífica del ADN en estudios de comparación de viromas entre individuos distintos (diversidad beta) podría ser irrelevante. Esto es así principalmente porque los viromas de individuos distintos son únicos y comparten un porcentaje muy pequeño de sus genomas virales (Minot et al., 2011; Pride et al., 2011b; Reyes et al., 2010). Esto se observa claramente en los sistemas de ordenación bidimensional donde el viroma no amplificado y los amplificados se agrupan próximos entre sí (todos los amplificados en el caso de disimilitudes de Bray-Curtis y sólo los procedentes de la amplificación con *MDA* desde 1 ng de molde en el caso de los índices de Sørensen) y mantienen una gran distancia con otros viromas de saliva de individuos no relacionados. Cuando la diversidad beta entre individuos es tan alta, pequeños solapamientos entre muestras como los que presenta la mezcla de saliva no amplificada Unamp1 con el viroma de saliva H101 (procedente de un individuo que contribuyó a esa misma mezcla), provoca un fuerte agrupamiento en los sistemas de ordenación. Este resultado es coherente con estudios previos que demuestran la convergencia entre viromas de saliva de sujetos que cohabitan en la misma casa, aunque solo comparten un número pequeño de bacteriófagos (Ly et al., 2016; Pride et al., 2011a; Robles-Sikisaka et al., 2013).

2. Diversidad de virus de la boca

Las comunidades de virus de la cavidad bucal humana estudiadas en esta tesis presentaban índices de diversidad alfa de Shannon en mucosa de 5,46-7,16 y en placa de 5,13-6,63 y estimaciones de riqueza de especies varían entre los 174-1.969 genotipos. Estos resultados son ligeramente superiores a los obtenidos en anteriores estudios de boca, con índices de Shannon de 4 ± 0,5, 4,8 ± 0,03 y 5,0 (Abeles et al., 2015; Lim et al., 2017; Pérez-Brocal y Moya, 2018), y rangos de riqueza de especies muy similares entre 293-2.200 genotipos (Pride et al., 2011a). Estas estimaciones de diversidad resultaron consistentemente superiores a las obtenidas en el tracto intestinal con índices de Shannon para viromas de heces de 3,09 ± 0,64, 3,32 ± 0,71 3,35 y 3,4 (Kim et al., 2011; Lim et al., 2015; Reyes et al., 2010; Zuo et al., 2019).

La mayor diversidad alfa de los viromas de mucosa sobre los de placa dental resultó estadísticamente significativa. Sin embargo, no observamos diferencias en la diversidad alfa de los viromas entre los estados de salud y enfermedad, pese a que sí se han publicado cambios en la diversidad de bacterias de la cavidad bucal en estos mismos procesos de enfermedad (He et al., 2018; Hijazi et al., 2015; Jorth et al., 2014) como colitis ulcerosa (Zuo et al., 2019) o infección crónica con *Clostridium difficil*e (Lawley et al., 2012)

El bajo número de secuencias compartidas, la casi nula correlación de Pearson y la enorme distancia en los sistemas de ordenación entre los viromas de la boca de individuos distintos, indican una gran diversidad beta en las comunidades de virus de estos ecosistemas y coincide con resultados previos en saliva (Pérez-Brocal y Moya, 2018; Pride et al., 2011a), y heces (Reyes et al., 2010; Wylie et al., 2014). Esta alta diversidad beta, contrasta con el alto porcentaje de *contigs* compartidos entre los viromas de saliva y mucosa del mismo individuo, posiblemente como resultado del continuo contacto físico entre estos dos ambientes. La placa dental y la mucosa bucal también mantienen un contacto físico estrecho que podría justificar la distribución ubicua de muchos de los bacteriófagos que hemos encontrado en casi todos los viromas estudiados. Pese a ello, pudimos observar una separación de la mayoría de los viromas por ambiente que estaba en el límite de la significancia estadística (*p-value* = 0,048). Así que es probable que un estudio con un número mayor y más homogéneo de muestras (nuestros viromas proceden de individuos sanos y enfermos) lograra un soporte estadístico más robusto a las diferencias entre los viromas de mucosa bucal y placa dental. Debido a que son comunidades dominadas por virus que infectan bacterias, resultó coherente encontrar diferencias estadísticamente significativas al comparar las comunidades de bacterias de estos dos ambientes en los estudios basados en el gen marcador 16S. También observamos una mayor dispersión de los viromas de mucosa con respecto a los de placa, lo que coincide con su mayor diversidad alfa, aunque en esta ocasión, dicha diferencia no resultó significativa estadísticamente. Los principales responsables de las diferencias entre ambientes no eran los virus más abundantes de los viromas e incluían virus emparentados con *Human*

betaherpesvirus 7 y con bacteriófagos que infectan miembros del género *Streptococcus*. Estos virus se encontraban preferentemente asociados a mucosa y estaban prácticamente ausentes en los viromas de placa dental, lo que resultó consistente con la mayor abundancia de bacterias del filo *Firmicutes* (observada en esta tesis) y con la alta prevalencia de este herpesvirus en la mucosa bucal humana (Grinde, 2013; McGowin y Pyles, 2010).

Otros autores han observado diferencias a nivel poblacional entre la microbiota de la mucosa oral y de la placa dental durante el desarrollo de Estomatitis Aftosa Recurrente y Caries, respectivamente. Estas disbiosis se caracterizan por un aumento de *Prevotella* (Marchini et al., 2007) y también por una disminución de *Firmicutes* y un aumento de *Proteobacteria* (Hijazi et al., 2015) en las aftas, y por el aumento de un número variable de bacterias cariogénicas en caries (Belda-Ferre et al., 2012; He et al., 2018; Simón-Soro y Mira, 2015) que algunos han definido como un catástrofe ecológica en el ecosistema microbiano de la placa supragingival (Marsh, 2003). La gran variedad de consorcios bacterianos disbióticos en muestras con caries (Aas et al., 2008; Fejerskov, 2004; Simón-Soro y Mira, 2015) puede obedecer a que las comunidades de bacterias varían considerablemente en función de la posición de la pieza dental afectada (Simón-Soro et al., 2013), y a que comunidades taxonómicamente distintas puedan presentar capacidades metabólicas coincidentes (Simón-Soro y Mira, 2015). Estos antecedentes podrían explicar por qué no hemos encontramos diferencias a nivel poblacional entre los microbiomas y viromas sanos o afectados por caries o aftas.

De nuevo, un estudio con un mayor número de muestras, y con un perfil más acotado de enfermedad, podrían permitirnos observar diferencias a nivel poblacional entre los viromas y microbiomas, y lo que es más importante, identificar biomarcadores que aporten información para comprender mejor el papel que desempeñan los bacteriófagos en estas enfermedades comunes de la boca y que podrían, eventualmente, ser empleados en diagnóstico. En este sentido, resulta prometedor que hayamos observado una distribución preferente en la placa de personas con caries de *contigs* virales emparentados con *Enterococcus phage phiFL4A*, cuyo hospedador, *Enterococcus faecalis* se ha asociado con esta enfermedad (Kouidhi et al., 2011).

El ensamblaje de las lecturas de los viromas en *contigs* originó una media de 2.594 *contigs* por viroma, que quedaron reducidos a 58 de media por viroma (1.557 en total) al aplicar un criterio de selección de *contigs* largos y alta cobertura por un lado, y de asignación taxonómica al dominio de los virus por otro. Pese a esta drástica reducción en el número de *contigs,* hasta un 74,47% de todas las secuencias de los viromas se alineaban contra estos *contigs* largos virales, confirmando que la mayoría de información de estos metagenomas es de origen viral. El principal objetivo de esta doble filtración era la eliminación de *contigs* de origen bacteriano como mencionamos anteriormente y una evidencia contundente de que hemos logrado nuestro objetivo, es la ausencia completa en estos *contigs* de genes relacionados con ARNr 16S. El hecho de que en torno al 75% de las lecturas de los viromas se ensamblen en unos 58

contigs virales de alta cobertura (media 369,4x.), hace factible pensar que el 25% de secuencias restantes procedan de un número mucho mayor de virus poco abundantes cuyos genomas no hayan podido ensamblar en *contigs* largos, justificando las estimaciones de riqueza de especies obtenidas por *PHACCS* (174-1.969 especies).

Con el avance y desarrollo de la bioinformática se están generando herramientas que permiten la identificación de *contigs* virales desde microbiomas. *VirSorter* es una de las herramientas más utilizadas con este propósito debido a su alta fiabilidad (Roux et al., 2015a). En comparación con nuestro método, *VirSorter* es ligeramente menos sensible, logrando identificar como virales sólo un 66,2% de los *contigs* que identifica nuestro método. Otra diferencia es la inclusión por *VirSorter* de *contigs* que contienen profagos flanqueados por genes bacterianos. Estos *contigs* son descartados por nuestro método que los clasifica como bacterianos, lo que puede ser de utilidad cuando existe la certeza de contaminación bacteriana.

3. Caracterización de la composición de las comunidades de virus de la cavidad bucal humana

Una primera visión de la composición de las comunidades virales se obtuvo al comparar las lecturas de los viromas con la base de datos de proteínas *nr*. La distribución por dominios resultante fue consistente con estudios similares llevados a cabo con viromas de ecosistemas naturales y asociados a humanos (Angly et al., 2006; Minot et al., 2013), y se caracterizó por un bajo porcentaje de secuencias relacionadas con virus, siendo mayoritarias las secuencias asignadas a bacterias o sin resultados significativos. Esta distribución se debe, en parte, a la escasa información de secuencias virales disponibles en las bases de datos, a la existencia de multitud de genomas virales integrados en genomas bacterianos (profagos) y anotados como bacterias y, en menor medida, a la presencia de genes bacterianos transportados en los genomas virales durante los procesos de transferencia horizontal de genes. Además, y como mencionamos anteriormente, nuestros viromas también pueden contener niveles bajos de contaminación bacteriana real por el uso de filtros de 0,45µm. La comparación con bases de datos exclusivamente virales puede llevar también a conclusiones erróneas como la asignación incorrecta de secuencias a virus eucarióticos grandes de las familias *Mimiviridae* o *Phycodnaviridae*, como pudimos demostrar posteriormente al ensamblarlas en *contigs* de gran tamaño relacionados con genomas bacterianos del filo *TM7*. Gracias a los avances en las tecnologías de secuenciación masiva, durante los últimos años, las bases de datos están experimentando un notable aumento en el número de genomas virales depositados. Así, por ejemplo, entre Enero y Octubre de 2018, el número de genomas virales pasó de 7.000 a más de 10.000 (http://millardlab.org/bioinformatics/bacteriophage-genomes/). Esto nos permite aventurar un escenario para los próximos años en el que la comparación directa de secuencias con bases de datos menos sesgadas y más representativas de los virus que existen en la naturaleza pueda darnos una imagen más real de las comunidades virales.

Pese a estas limitaciones, esta aproximación evidenció que los viromas de la boca humana estaban dominados por bacteriófagos con cola, principalmente bacteriófagos de la familia *Siphoviridae* que siguen un estilo de vida preferentemente lisogénico. Este resultado coincide con estudios previos de viromas humanos de heces y saliva que sugieren que el tipo de regulación depredador-presa, al contrario que en ambientes acuáticos como el mar, no es el tipo de regulación principal de la microbiota humana (Edlund et al., 2015; Knowles et al., 2016; Silveira y Rohwer, 2016) y también coincide con el elevado número de profagos encontrado en los genomas de bacterias aisladas de muestras humanas (Manrique et al., 2017).

Una visión taxonómica más fiable de la composición de los viromas se puede obtenerse mediante la comparación de todos los genes que contiene cada *contig* con bases de datos de virus. Una buena parte de los 1.557 *contigs* virales presentaban tamaños distribuidos en dos picos de 17 kpb y 42 kpb, lo que refleja casi exactamente la distribución de tamaños de los *Caudovirales* disponibles en las bases de datos y sugiere que un gran porcentaje de estos *contigs* podían estar completos o casi completos. La mayoría de los contigs con tamaños cercanos a 17 kpb estaban estrechamente emparentados con *Actinomyces phage Av1*, un virus que infecta un poblador habitual de la placa dental: *Actinomyces naeslundii*, (Vielkind et al., 2015). En el pico de *contigs* con tamaños de 40-45 kpb, el grupo más abundante estaba lejanamente relacionados con virus que infectan bacterias de otros ambientes como *Arthrobacter phage Mudcat* (https://phagesdb.org/) y *Rhodococcus phage ReqiPoco6* (Summer et al., 2011). Pese a las grandes diferencias interpersonales en la composición de las comunidades de virus de la boca, estos contigs estaban presentes en prácticamente todos los viromas analizados de mucosa y placa dental. Su ubicuidad va más allá de estas nuestras, porque virus relacionados con *Actinomyces phage Av1* parecen ser también los virus más abundantes en placa subgingival (Wang et al., 2013), y los contigs relacionados con *Arthrobacter phage Mudcat* y *Rhodococcus phage* presentaban también una gran similitud de secuencia con el contig89, uno de los virus más abundantes en un estudio longitudinal de viromas de saliva de ocho individuos de Estados Unidos (Abeles et al., 2014) y en viromas de saliva de 72 individuos de Valencia (Pérez-Brocal y Moya, 2018). La mayor longitud de nuestros *contigs* sugiere que el contig89 podría ser sólo un fragmento del genoma de este virus. De hecho, los 23,55 kpb del contig89 alineaban, con una identidad de secuencia >90%, con la región 3’ de los *contigs* ensamblados en nuestros viromas, muchos de ellos con un tamaño superior a las 40 kpb. Estos dos grupos de bacteriófagos, ubicuos en la boca humana, podrían formar parte de un núcleo de virus altamente conservados en estos ecosistemas (virus *core*), al igual que el recientemente descubierto virus crAssphage en el intestino, donde es el virus más abundante y ubicuo (Dutilh et al., 2014; Guerin et al., 2018).

Para estudiar en mayor detalle estas comunidades de virus agrupamos más de 400 *contigs* virales considerados genomas completos (o casi completos) en un árbol proteómico. Este árbol es coherente con los previamente publicados que cuestionan la actual división de familias de virus del orden *Caudovirale*s, pero respaldan su actual división en subfamilias (Rohwer y Edwards, 2002). Los *contigs* de los viromas de la boca se agrupan en 31 *megaclusters* que representan bastante bien la diversidad completa de virus de la boca humana, ya que a partir de combinaciones al azar de tan solo 17 viromas se consigue tener virus representantes de los 31 *megaclusters,* y este número no aumenta al añadir nuevos viromas. Aunque la mayoría de los *megaclusters* incluían algún virus de referencia relacionado, los porcentajes de identidad media a nivel de nucleótido eran inferiores al 70% en 25 de los 31 *megaclusters*, y en muchos casos mostraban una pobre sintenia, indicando la escasa representación de los virus de la boca en las bases de datos. Pese a que la placa dental y la mucosa son ambientes muy diferentes, ninguno de los 31 *megaclusters*, presentaban una asociación clara por uno de los dos ambientes, ni por una condición de salud o enfermedad en particular, lo que sugiere su amplia distribución en la cavidad bucal. Esto puede deberse a la estrecha interacción entre ambos ecosistemas por proximidad física y por una dinámica de intercambio de virus a través de la saliva.

Tres de estos *megaclusters* presentan un elevado número de *contigs* y una distribución ubicua a lo largo de la mayoría de los viromas estudiados. El primero de ellos, el *megacluster* **W**, contenía 59 *contigs* procedentes de los viromas del 81% de las personas de nuestro estudio. Estos *contigs* mostraban una muy buena sintenia y un alto nivel de identidad de secuencia con el genoma de *Actinomyces phage Av1*. El *megacluster* **C**, con 49 *contigs* presentes en un 89% de las personas, incluía algunos *contigs* muy abundantes ensamblados con hasta el 45% de las secuencias de alguno de los viromas. Estos *contigs* presentaban cierta sintenia a lo largo de toda su secuencia y un % de CGs similar con los virus más relacionados *Arthrobacter phage Mudcat* y *Rhodococcus phage ReqiPoco6*. Sin embargo, las importantes diferencias de tamaño debidas entre otras cosas a la ausencia en nuestros *contigs* de genes que codifican por LysB, una proteína característica de este tipo de bacteriófagos, así como el hecho de que estos virus sean propios de suelo o de otros animales, apoyan la originalidad de este grupo ubicuo y dominante de los viromas de la boca. La presencia en alguno de ellos de genes que codifican endolisinas relacionadas con las de *Actinomyces phage Av1,* un bacteriófago que infecta *Actinomyces naeslundii,* sugiere que podría infectar bacterias del orden *Actinomycetales*. Curiosamente, algunos integrantes de este *megacluster* mostraban un bajo porcentaje de CGs propio de *Firmicutes* y mejor similitud de secuencia con algunos virus que infectan bacterias de este filo como *Lactococcus phage 1706*. El *megacluster* **A** es otro de los grupos más ubicuos (presente en un 78% de los individuos estudiados) y abundante (33 *contigs*) de la boca. La mayoría de los virus que lo conforman están lejanamente emparentados con *Enterococcus phage phiFL4A*, con el que presenta una pobre sintenia. Además, la circunstancia de que ningún genoma de referencia, ni siquiera *Enterococcus phage phiFL4A*, se agrupara dentro de este *megacluster*, revela la originalidad de este grupo de virus.

Aunque no tan abundante ni ubicuo como los tres *megaclusters* anteriores, el *megacluster* **AA** resultó interesante porque incluía bacteriófagos con genomas de gran tamaño (108-202 kpb) emparentados con un grupo de *Caudovirales* conocidos como bacteriófagos Jumbo (Yuan y Gao, 2017), cuyas cápsidas de gran tamaño podrían quedarse retenidas en filtros de 0,22 µm (Conceição-Neto et al., 2015; Lewandowska et al., 2017; Li et al., 2015; López-Bueno et al., 2009). La inclusión de estos virus en nuestros viromas podría justificar el uso de protocolos de filtración a baja velocidad y filtración en 0,45µm durante los pasos de enriquecimiento de virus, en lugar de los de 0,22µm más habitualmente usados.

Aunque 27 de los 31 *megaclusters* están formados por *contigs* emparentados con bacteriófagos del orden *Caudovirales*, nuestros viromas también contienen dos *megaclusters* relacionados con virus de las familias *Inoviridae* (**F** y **L**) y otro de *Microviridae* (**AE**). Pese a la sobreamplificación de sus genomas circulares de pequeño tamaño por *MDA*, estos virus no parecen ser tan abundantes en la boca como en el tracto intestinal (Minot et al., 2013; Reyes et al., 2010). Esto podría estar relacionado con la menor representación de bacterias del filo *Bacteroidetes* en la boca en relación al tracto intestinal, ya que muchos de estos bacteriófagos intestinales infectan bacterias de este filo (Krupovic y Forterre, 2011).

El único *megacluster* de virus eucarióticos, *megacluster* **AD**, incluía *contigs* circulares relacionados con anellovirus, papilomavirus y CREES-DNA virus. En nuestros viromas, los anellovirus y papilomavirus aparecen más frecuentemente en mucosa, un ecosistema más relacionado con la saliva, donde virus de estas familias se han detectado en estudios anteriores (Ross et al., 1999). Aunque no pudimos ensamblar el genoma completo de ningún herpesvirus en estos viromas, sí que encontramos numerosos *contigs* que mostraban una identidad de secuencia cercana al 100% con varios herpesvirus humanos (principalmente *HSV7*). Por el contrario, el resto de *contigs* de virus eucarióticos eran mucho más diferentes a los presentes las bases de datos. De hecho, dos de los genomas circulares completos de anellovirus y otros dos de papilomavirus presentaban valores de similitud de secuencia por debajo de los criterios de demarcación de especie en anellovirus y de tipo humano en papilomavirus, por lo que se propusieron como nuevos virus humanos. La identificación de estos nuevos virus humanos demuestra la enorme sensibilidad de los estudios metagenómicos, incluso en ambientes dominados por bacteriófagos. Finalmente, los cinco nuevos genomas circulares asignados al grupo *CRESS-DNA*, podrían no ser virus humanos e infectar otros microorganismos eucarióticos que habitan estos ecosistemas, como se ha sugerido para virus similares encontrados en viromas de humanos y otros mamíferos (Cheung et al., 2014; Cui et al., 2017).

4. Predicción del hospedador que infectan los bacteriófagos de la boca

Pese a que los bacteriófagos constituyen uno de los elementos de regulación más importantes de las comunidades microbianas (Fernández et al., 2018; Rohwer y Thurber, 2009; Silveira y Rohwer, 2016), queda aún un largo camino por recorrer para entender la complejidad de los tipos de interacción que establecen con sus hospedadores. En nuestra opinión, el primer paso para avanzar en este campo es caracterizar en profundidad la diversidad genética de las comunidades de bacteriófagos como hemos tratado de hacer en las primeras etapas de esta tesis. A continuación, y como segundo paso, creemos que es vital definir el rango de hospedador de los bacteriófagos más importantes de la comunidad. Con este propósito, hemos recurrido a cinco aproximaciones bioinformáticas con distinta sensibilidad. La inferencia de hospedador en función del genoma viral más próximo encontrado en la base de datos *GenBank*, nos proporcionó información acerca del 93% de los *contigs* virales. Aunque no hemos validado la fiabilidad de este método, la fuerte agregación de los *contigs* que podrían infectar bacterias del mismo filo sugiere su utilidad y revela la importancia de la coevolución entre los bacteriófagos y sus hospedadores. Este resultado también sugiere que el filo del hospedador debería ser un criterio taxonómico más importante, por ejemplo, que la morfología de las partículas virales.

Otro método de predicción de hospedador utilizado fue la comparación de la frecuencia de oligonucleótidos entre los *contigs* virales y las bacterias secuenciadas de la boca. Este método, que ya había sido usado con éxito en estudios previos (Edwards et al., 2016; Ogilvie et al., 2013; Roux et al., 2015b), fue validado en esta tesis con bacteriófagos y profagos de hospedador conocido y dio lugar a resultados congruentes con el anterior método. El origen viral de muchos de los separadores en los cassettes *CRISPRs* de las bacterias, representa una fuente fiable para averiguar la interacción entre bacteriófagos y hospedadores. La obtención de unos 20 millones de secuencias por microbioma nos ha permitido encontrar 1.360 separadores en *CRISPR*s relacionados con 544 de nuestros *contigs* virales. Pese a que el estudio se hizo principalmente con comunidades bacterianas de placa dental y no de mucosa, este método nos ha proporcionado información del posible hospedador de hasta un 30% de los *contigs* virales. La decisión de analizar sólo los separadores contenidos en *CRISPR*s con más de cuatro separadores puede haber reducido la identificación de falsos *CRISPR* (Kupczok et al., 2015), contribuyendo a mejorar el porcentaje de separadores relacionados con virus en nuestro trabajo (4,9%), con respecto al 0,82% descrito en estudios previos (Koonin et al., 2017). Por último, las predicciones de hospedador basadas en la procedencia de genes metabólicos auxiliares y en las secuencias de integración resultaron mucho menos sensibles, pero fueron también congruentes con el resto de aproximaciones. Pese al distinto nivel de sensibilidad de los cinco métodos usados, éstos mostraban altos niveles de congruencia y coincidían en que los hospedadores más frecuentes eran por orden decreciente *Actinobacteria*, *Proteobacteria* y *Firmicutes*.

Uno de los resultados más sorprendentes de esta tesis, es que la mayoría de los *contig* (82,6% en el caso del método basado en *BLAST)*, de los viromas de mucosa y placa dental, y hasta nueve *megaclusters* distintos, se asocian exclusivamente a virus que infectan bacterias del filo *Actinobacteria*, cuando este filo es mucho menos abundante que los *Firmicutes* en ambos ambientes, e incluso menos abundante que *Proteobacteria* y *Bacteroidetes* en mucosa. Esta distribución de filos bacterianos determinados en metagenomas de ARNr 16S es consistente con estudios similares previos para mucosa bucal y placa dental (Dewhirst et al., 2010; He et al., 2018; Verma et al., 2018), aunque en algún estudio de placa supragingival se ha propuesto una abundancia mayor de *Actinobacteria* (Ziouani et al., 2015). En este sentido, estudios recientes basados en mejoras de la técnica *FISH*, han demostrado que la estructura del *biofilm* de la placa dental está dominada por actinobacterias del género *Corynebacterium* que forman una estructura central de la cual penden filamentos que se extienden radialmente (Ferrer y Mira, 2016). La importancia numérica de *Corynebacterium* y otro género de actinobacterias: *Actinomycetes* se había propuesto anteriormente en estudios metatranscriptómicos donde son los dos géneros más activos transcripcionalmente de la placa dental (Benítez-Páez et al., 2014). Así, una posible explicación a esta falta de correlación entre las bacterias de la microbiota bucal y las bacterias predichas que infectan las comunidades de bacteriófagos, podría ser que las bacterias más activas metabólica y transcripcionalmente contribuyan a la comunidad de virus libres con una mayor progenie viral. Otra posible explicación es la peor amplificación por *PCR* del gen que codifica por ARNr 16S de las bacterias de este filo debido a su elevado % de CGs (Cabrera-Rubio et al., 2012; Simón-Soro et al., 2014).

Otro aspecto interesante de este estudio es que, pese a que algunos de los *megaclusters* están formados por virus que infectan una familia definida de hospedadores, en otros muchos casos están formados por virus que podrían infectar varias familias distintas de bacterias. Esto último podría ser resultado simplemente de una pérdida de precisión de las técnicas de predicción de hospedador, pero también podría tener que ver con un rango de hospedador amplio de los bacteriófagos de estos *megaclusters*. En este sentido, estudios recientes sugieren que el rango de hospedador de los bacteriófagos podría no ser tan estrecho como se pensaba (Flores et al., 2013, 2011; Weitz et al., 2013).

Otra discrepancia importante entre la comunidad de bacterias de las muestras analizadas y los hospedadores predichos para su bacteriófagos, es la casi ausencia de interacciones con *Fusobacteria*, *Bacteroidetes* y *Spirochaeta*, pese a ser pobladores habituales de la cavidad bucal. Este resultado puede estar relacionado con la menor cantidad de genomas de bacterias de estos filos (Land et al., 2015), así como de los bacteriófagos que las infectan, en las bases de datos (https://www.genome.jp/virushostdb). Aunque también puede obedecer a una menor presión selectiva de los bacteriófagos sobre estos filos.

5. Repertorio de enzimas líticas codificadas en el genoma de los bacteriófagos de la boca

La Organización Mundial de la Salud lleva años alertando acerca de la crisis sanitaria que producirá el imparable aumento de bacterias patogénicas multirresistentes a antibióticos (https://www.who.int/antimicrobial-resistance/publications/surveillancereport/en/). Las previsiones del impacto económico de este problema (O'Neill, 2014) deberían estimular la investigación de nuevas formas de control bacteriano (Simpkin et al., 2017). Una de las estrategias que se están explorando es el uso de enzimas líticas de la pared de peptidoglicanos codificadas en el genoma de los bacteriófagos: las lisinas. La alta especificidad de acción de estas enzimas rompiendo la pared bacteriana de sus hospedadores podría representar una ventaja sobre los antibióticos, cuyo amplio espectro de acción altera drásticamente la microbiota intestinal (Panda et al., 2014; Schubert et al., 2015).

En este trabajo describimos la presencia de 500 nuevas lisinas presentes en 23 de los 30 *megacluster* de bacteriófagos de la boca. El esfuerzo que hemos realizado para predecir el hospedador más probable de estos bacteriófagos nos lleva a hipotetizar sobre qué familias de bacterias podrían ser eficaces estas lisinas. Entre este amplio repertorio de lisinas destacan las codificadas en el genoma de los bacteriófagos relacionados con *Actinomyces phage Av1*. El hospedador de este virus, *Actinomyces naeslundii*, produce enfermedades que pueden ser graves como la actinomicosis (Valour et al., 2014). También hemos identificado lisinas potencialmente específicas para bacterias de las familias *Lactobacillaceae* y *Corynebacteriaceae*, relacionadas con el desarrollo de caries; *Porphyromonadaceae*, *Pasteurellaceae*, y *Peptostreptococcaceae*, implicadas en el desarrollo de periodontitis; *Flavobacteriaceae*, *Bacteroidaceae* o *Peptostreptococcaceae*, que podría estar relacionadas con cáncer bucal; o multitud de lisinas para *Streptococcaceae* que podrían tener interés clínico en el tratamiento de éstas y otras afecciones de la cavidad oral.

Desconocemos la eficacia real que el uso de este repertorio de enzimas líticas podría tener en la práctica clínica. Sin embargo, su previsible eficacia biológica como mecanismo lítico de salida de la célula infectada para los bacteriófagos que las contienen y el actual escenario de emergencia de bacterias multirresistentes, hace que, en nuestra opinión, merezca la pena explorar esta vía de investigación.

CONCLUSIONES

1. Hemos puesto a punto un protocolo sencillo de enriquecimiento de partículas virales que preserva la composición de la comunidad viral, permitiendo la recuperación parcial de virus de gran tamaño.
2. La amplificación al azar mediante *MDA* de los viromas de la boca introduce sesgos estocásticos desde picogramos de ADN molde, y sistemáticos desde nanogramos. Los sesgos sistemáticos se deben principalmente a la dificultad para amplificar regiones con %CGs extremos, un problema compartido con otros tipos de amplificación por *PCR* como *SISPA*. Además, *SISPA* produce picos de alta cobertura en regiones de baja complejidad lingüística.
3. Los sesgos debidos a la amplificación al azar no tienen apenas impacto en los estudios de comparación de viromas humanos debido a su gran variabilidad interpersonal.
4. Hemos puesto a punto un protocolo de análisis bioinformático para la identificación de *contigs* procedentes de virus libres. Este protocolo descarta eficazmente *contigs* de bacterias contaminantes de pequeño tamaño que atraviesan los filtros de 0,45µm.
5. Esta tesis representa el mayor estudio metagenómico de virus de la cavidad bucal realizado hasta la fecha. Los viromas generados están dominados por bacteriófagos del orden *Caudovirales* y contienen un pequeño porcentaje de virus eucarióticos, incluyendo cuatro virus humanos nuevos de las familias *Anelloviridae* y *Papillomaviridae*.
6. Existe una clara diferencia en la composición del microbioma de placa dental y mucosa oral, que coincide con una sutil separación también de sus viromas. No hemos encontrado diferencias ni en el microbioma ni en el viroma, entre individuos sanos y afectados por caries o estomatitis aftosa recurrente.
7. Hemos ensamblado *de novo* más de 400 genomas completos, o casi completos, de virus que se agrupan en 31 *megaclusters*. La mayoría de estos virus están pobremente relacionados con los virus disponibles en las bases de datos y algunos de los más abundantes y ubicuos lo son también en otros viromas bucales publicados.
8. La predicción del hospedador de los bacteriófagos de la boca mediante cinco estrategias bioinformáticas diferentes dio lugar a resultados coincidentes. Las bacterias del filo *Actinobacteria*, pese a no ser los componentes mayoritarios de la microbiota bucal, constituyen el hospedador más frecuente de los bacteriófagos de este ecosistema.
9. La agrupación de los bacteriófagos de la boca en función del filo de su posible hospedador es una clara evidencia de la importancia de la coevolución entre los bacteriófagos y las bacterias que infectan. Este resultado también sugiere que el filo del hospedador debería ser un criterio taxonómico más importante que la morfología de las partículas virales.
10. Los bacteriófagos de la boca humana tienen un repertorio de cientos de genes que codifican enzimas que rompen la pared de peptidoglicano. La predicción del hospedador de los bacteriófagos que las contienen constituye un punto de partida para evaluar su utilidad como alternativa a los antibióticos.

BIBLIOGRAFÍA

Aagaard, K., Ma, J., Antony, K.M., Ganu, R., Petrosino, J., Versalovic, J., 2014. The placenta harbors a unique microbiome. Sci. Transl. Med. 6, 237ra65.

Aas, J.A., Griffen, A.L., Dardis, S.R., Lee, A.M., Olsen, I., Dewhirst, F.E., Leys, E.J., Paster, B.J., 2008. Bacteria of dental caries in primary and permanent teeth in children and young adults. J. Clin. Microbiol. 46, 1407-1417.

Abby, S.S., Néron, B., Ménager, H., Touchon, M., Rocha, E.P.C., 2014. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One 9, e110726.

Abeles, S.R., Ly, M., Santiago-Rodriguez, T.M., Pride, D.T., 2015. Effects of long term antibiotic therapy on human oral and fecal viromes. PLoS One 10, e0134941.

Abeles, S.R., Robles-Sikisaka, R., Ly, M., Lum, A.G., Salzman, J., Boehm, T.K., Pride, D.T., 2014. Human oral viruses are personal, persistent and gender-consistent. ISME J. 8, 1753-1767.

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., Chang, S.H., Chang, H.W., Watson, D., Brodie, E.L., Hazen, T.C., Keller, M., 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. Appl. Environ. Microbiol. 72, 3291-3301.

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, 537-544.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18.

Anderson, D., 2016. Wrote Letter 39 of 1683-09-17 (AB 76) to Francis Aston. Lens on Leeuwenhoek 44, 76.

Anderson, M.J., 2006. Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62, 245-253.

Anderson, M.J., Walsh, D.C.I., 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol. Monogr. 83, 557-574.

Ando, H., Lemire, S., Pires, D.P., Lu, T.K., 2015. Engineering modular viral scaffolds for targeted bacterial population editing. Cell Syst 1, 187-196.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A., Rohwer, F., 2006. The marine viromes of four oceanic regions. PLoS Biol. 4, e368.

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D.A., Barott, K., Cottrell, M.T., Desnues, C., Dinsdale, E.A., Furlan, M., Haynes, M., Henn, M.R., Hu, Y., Kirchman, D.L., McDole, T., McPherson, J.D., Meyer, F., Miller, R.M., Mundt, E., Naviaux, R.K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., Rohwer, F., 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput. Biol. 5, e1000593.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., Rohwer, F., 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6, 41.

Arezi, B., Hogrefe, H., 2009. Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer. Nucleic Acids Res. 37, 473-481.

Arezi, B., Xing, W., Sorge, J.A., Hogrefe, H.H., 2003. Amplification efficiency of thermostable DNA polymerases. Anal. Biochem. 321, 226-235.

Aronesty, E., 2013. Comparison of sequencing utility programs. Open Bioinforma. J. 7, 1-8.

Arriola, E., Lambros, M.B.K., Jones, C., Dexter, T., Mackay, A., Tan, D.S.P., Tamber, N., Fenwick, K., Ashworth, A., Dowsett, M., Reis-Filho, J.S., 2007. Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. Lab. Invest. 87, 75-83.

Babalova, E.G., Katsitadze, K.T., Sakvarelidze, L.A., Imnaishvili, N.S., Sharashidze, T.G., Badashvili, V.A., Kiknadze, G.P., Meĭpariani, A.N., Gendzekhadze, N.D., Machavariani, E.V., Gogoberidze, K.L., Gozalov, E.I., Dekanosidze, N.G., 1968. Preventive value of dried dysentery bacteriophage. Zh. Mikrobiol. Epidemiol. Immunobiol. 45, 143-145.

Badet, C., Thebaud, N.B., 2008. Ecology of lactobacilli in the oral cavity: a review of literature. Open Microbiol. J. 2, 38-48.

Bagramian, R.A., Garcia-Godoy, F., Volpe, A.R., 2009. The global increase in dental caries. A pending public health crisis. Am. J. Dent. 22, 3-8.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455-477.

Barr, J.J., Auro, R., Furlan, M., Whiteson, K.L., Erb, M.L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A.S., Doran, K.S., Salamon, P., Youle, M., Rohwer, F., 2013. Bacteriophage adhering to mucus provide a non–host-derived immunity. Proc. Natl. Acad. Sci. U. S. A. 110, 10771-10776.

Belda-Ferre, P., Alcaraz, L.D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M., Mira, A., 2012. The oral metagenome in health and disease. ISME J. 6, 46-56.

Benita, Y., Oosting, R.S., Lok, M.C., Wise, M.J., Humphery-Smith, I., 2003. Regionalized GC content of template DNA as a predictor of PCR success. Nucleic Acids Res. 31, e99-e106.

Benítez-Páez, A., Belda-Ferre, P., Simón-Soro, A., Mira, A., 2014. Microbiota diversity and gene expression dynamics in human oral biofilms. BMC Genomics 15, 311.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., Sayers, E.W., 2018. GenBank. Nucleic Acids Res. 46, 41-47.

Bik, E.M., Long, C.D., Armitage, G.C., Loomer, P., Emerson, J., Mongodin, E.F., Nelson, K.E., Gill, S.R., Fraser-Liggett, C.M., Relman, D.A., 2010. Bacterial diversity in the oral cavity of 10 healthy individuals. ISME J. 4, 962-974.

Binga, E.K., Lasken, R.S., Neufeld, J.D., 2008. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J. 2, 233-241.

Biotech, G.S.L., s. f. SnapGene® software.

Blanco, L., Bernad, A., Lázaro, J.M., Martín, G., Garmendia, C., Salas, M., 1989. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. J. Biol. Chem. 264, 8935-8940.

Bogovazova, G.G., Voroshilova, N.N., Bondarenko, V.M., Gorbatkova, G.A., Afanas'eva, E.V., Kazakova, T.B., Smirnov, V.D., Mamleeva, A.G., Glukharev, I.A., Erastova, E.I., 1992. Immunobiological properties and therapeutic effectiveness of preparations from Klebsiella bacteriophages. Zh. Mikrobiol. Epidemiol. Immunobiol. 30-33.

Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., Madden, T.L., 2012. Domain enhanced lookup time accelerated BLAST. Biol. Direct 7, 12.

Börnigen, D., Ren, B., Pickard, R., Li, J., Ozer, E., Hartmann, E.M., Xiao, W., Tickle, T., Rider, J., Gevers, D., Franzosa, E.A., Davey, M.E., Gillison, M.L., Huttenhower, C., 2017. Alterations in oral bacterial communities are associated with risk factors for oral and oropharyngeal cancer. Sci. Rep. 7, 17686.

Bray, J.R., Curtis, J.T., 1957. An Ordination of the upland forest community of southern Wisconsin. Ecology Monographs 27, 325–349.

Bredel, M., Bredel, C., Juric, D., Kim, Y., Vogel, H., Harsh, G.R., Recht, L.D., Pollack, J.R., Sikic, B.I., 2005. Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. J. Mol. Diagn. 7, 171-182.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., Rohwer, F., 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. 185, 6220-6223.

Breitbart, M., Rohwer, F., 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques 39, 729-736.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer, F., 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. 99, 14250-14255.

Breitbart, M., Thompson, L., Suttle, C., Sullivan, M., 2007. Exploring the vast diversity of marine viruses. Oceanography 20, 135-139.

Briers, Y., Walmagh, M., Van Puyenbroeck, V., Cornelissen, A., Cenens, W., Aertsen, A., Oliveira, H., Azeredo, J., Verween, G., Pirnay, J.-P., Miller, S., Volckaert, G., Lavigne, R., 2014. Engineered endolysin-based «Artilysins» to combat multidrug-resistant gram-negative pathogens. MBio 5, e01379-14.

Briese, T., Paweska, J.T., McMullan, L.K., Hutchison, S.K., Street, C., Palacios, G., Khristova, M.L., Weyer, J., Swanepoel, R., Egholm, M., Nichol, S.T., Lipkin, W.I., 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog. 5, e1000455.

Broecker, F., Russo, G., Klumpp, J., Moelling, K., 2017. Stable core virome despite variable microbiome after fecal transfer. Gut Microbes 8, 214-220.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., Banfield, J.F., 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208-211.

Bruynoghe, R., Maisin, J., Others, 1921. Essais de thérapeutique au moyen du bacteriophage. CR Soc. Biol 85, 1120-1121.

Cabrera-Rubio, R., Collado, M.C., Laitinen, K., Salminen, S., Isolauri, E., Mira, A., 2012. The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. Am. J. Clin. Nutr. 96, 544-551.

Camelo-Castillo, A., Benítez-Páez, A., Belda-Ferre, P., Cabrera-Rubio, R., Mira, A., 2014. Streptococcus dentisani sp. nov., a novel member of the mitis group. Int. J. Syst. Evol. Microbiol. 64, 60-65.

Camelo-Castillo, A.J., Mira, A., Pico, A., Nibali, L., Henderson, B., Donos, N., Tomás, I., 2015. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front. Microbiol. 6, 119.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335-336.

Carlton, R.M., 1999. Phage therapy: past history and future prospects. Arch. Immunol. Ther. Exp. 47, 267-274.

Casjens, S., 2003. Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. 49, 277-300.

Castro-Mejía, J.L., Muhammed, M.K., Kot, W., Neve, H., Franz, C.M.A.P., Hansen, L.H., Vogensen, F.K.,

Nielsen, D.S., 2015. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. Microbiome 3, 64.

Chadha, P., Katare, O.P., Chhibber, S., 2017. Liposome loaded phage cocktail: enhanced therapeutic potential in resolving Klebsiella pneumoniae mediated burn wound infections. Burns 43, 1532-1543.

Cha, K., Oh, H.K., Jang, J.Y., Jo, Y., Kim, W.K., Ha, G.U., Ko, K.S., Myung, H., 2018. Characterization of two novel bacteriophages infecting multidrug-resistant (MDR) Acinetobacter baumannii and evaluation of their therapeutic efficacy in vivo. Front. Microbiol. 9, 696.

Chan, B.K., Turner, P.E., Kim, S., Mojibian, H.R., Elefteriades, J.A., Narayan, D., 2018. Phage treatment of an aortic graft infected with Pseudomonas aeruginosa. Evol Med Public Health 2018, 60-66.

Chaudhry, W.N., Concepción-Acevedo, J., Park, T., Andleeb, S., Bull, J.J., Levin, B.R., 2017. Synergy and Order Effects of Antibiotics and Phages in Killing Pseudomonas aeruginosa Biofilms. PLoS One 12, e0168615.

Chen, C., Huang, H., Wu, C.H., 2017. Protein bioinformatics databases and resources, en: Wu, C.H., Arighi, C.N., Ross, K.E. (Eds.), Protein bioinformatics: from protein modifications and networks to proteomics. Springer New York, New York, NY, pp. 3-39.

Chen, K., Pachter, L., 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput. Biol. 1, 106-112.

Cheung, A.K., Ng, T.F.F., Lager, K.M., Alt, D.P., Delwart, E.L., Pogranichniy, R.M., 2014. Identification of a novel single-stranded circular DNA virus in pig feces. Genome Announc. 2, e00347-14.

Clemente, J.C., Ursell, L.K., Parfrey, L.W., Knight, R., 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148, 1258-1270.

Coffey, B., Mills, S., Coffey, A., McAuliffe, O., Ross, R.P., 2010. Phage and their lysins as biocontrol agents for food safety applications. Annu. Rev. Food Sci. Technol. 1, 449-468.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G., Yoosuf, N., Million, M., La Scola, B., Raoult, D., 2013. Evidence of the megavirome in humans. J. Clin. Virol. 57, 191-200.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C.K., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E., Matthijnssens, J., 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci. Rep. 5, 16532.

Corby, P.M., Lyons-Weiler, J., Bretz, W.A., Hart, T.C., Aas, J.A., Boumenna, T., Goss, J., Corby, A.L., Junior, H.M., Weyant, R.J., Paster, B.J., 2005. Microbial risk indicators of early childhood caries. J. Clin. Microbiol. 43, 5753-5759.

Corsini, B., Díez-Martínez, R., Aguinagalde, L., González-Camacho, F., García-Fernández, E., Letrado, P., García, P., Yuste, J., 2018. Chemotherapy with phage lysins reduces pneumococcal colonization of the respiratory tract. Antimicrob. Agents Chemother. 62, e02212-17.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., Pourcel, C., 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 46, 246-251.

Cui, L., Morris, A., Huang, L., Beck, J.M., Twigg, H.L., III, Von Mutius, E., Ghedin, E., 2014. The microbiome and the lung. Ann. Am. Thorac. Soc. 11, 227-232.

Cui, L., Wu, B., Zhu, X., Guo, X., Ge, Y., Zhao, K., Qi, X., Shi, Z., Zhu, F., Sun, L., Zhou, M., 2017. Identification and genetic characterization of a novel circular single-stranded DNA virus in a human upper respiratory tract sample. Arch. Virol. 162, 3305-3312.

Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V.A., Foster, S., Gilmore, B.F., Hancock, R.E.W., Harper, D., Henderson, I.R., Hilpert, K., Jones, B.V., Kadioglu, A., Knowles, D., Ólafsdóttir, S., Payne, D., Projan, S., Shaunak, S., Silverman, J., Thomas, C.M., Trevor J Trust, Warn, P., Rex, J.H., 2016.

Alternatives to antibiotics—a pipeline portfolio review. Lancet Infect. Dis. 16, 239-251.

Daly, G.M., Bexfield, N., Heaney, J., Stubbs, S., Mayer, A.P., Palser, A., Kellam, P., Drou, N., Caccamo, M., Tiley, L., Alexander, G.J.M., Bernal, W., Heeney, J.L., 2011. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. PLoS One 6, e28879.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S.F., Egholm, M., Lasken, R.S., 2002. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. U. S. A. 99, 5261-5266.

DePew, J., Zhou, B., McCorrison, J.M., Wentworth, D.E., Purushe, J., Koroleva, G., Fouts, D.E., 2013. Sequencing viral genomes from a single isolated plaque. Virology Journal 10, 181.

De Vlaminck, I., Khush, K.K., Strehl, C., Kohli, B., Luikart, H., Neff, N.F., Okamoto, J., Snyder, T.M., Cornfield, D.N., Nicolls, M.R., Weill, D., Bernstein, D., Valantine, H.A., Quake, S.R., 2013. Temporal response of the human virome to immunosuppression and antiviral therapy. Cell 155, 1178-1187.

Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C.R., Yu, W.-H., Lakshmanan, A., Wade, W.G., 2010. The human oral microbiome. J. Bacteriol. 192, 5002-5017.

d'Herelle, F., 1917. Sur un microbe invisible antagoniste des bacilles dysentériques. CR Acad. Sci. Paris 165, 373-375.

Direito, S.O.L., Zaura, E., Little, M., Ehrenfreund, P., Röling, W.F.M., 2014. Systematic evaluation of bias in microbial community profiles induced by whole genome amplification. Environ. Microbiol. 16, 643-657.

Domenech, M., García, E., Moscoso, M., 2011. In vitro destruction of Streptococcus pneumoniae biofilms with bacterial and phage peptidoglycan hydrolases. Antimicrob. Agents Chemother. 55, 4144-4148.

Dridi, B., Raoult, D., Drancourt, M., 2011. Archaea as emerging organisms in complex human microbiomes. Anaerobe 17, 56-63.

Duhaime, M.B., Deng, L., Poulos, B.T., Sullivan, M.B., 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environ. Microbiol. 14, 2526-2537.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., Felts, B., Dinsdale, E.A., Mokili, J.L., Edwards, R.A., 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nature Communications 5, 4498.

Dye, B., Thornton-Evans, G., Li, X., Iafolla, T., 2015. Dental caries and tooth loss in adults in the United States, 2011-2012. NCHS Data Brief 197.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

Edlund, A., Santiago-Rodriguez, T.M., Boehm, T.K., Pride, D.T., 2015. Bacteriophage and their potential roles in the human oral cavity. J. Oral Microbiol. 7, 27423.

Edwards, R.A., McNair, K., Faust, K., Raes, J., Dutilh, B.E., 2016. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol. Rev. 40, 258-272.

Eke, P.I., Dye, B.A., Wei, L., Slade, G.D., Thornton-Evans, G.O., Borgnakke, W.S., Taylor, G.W., Page, R.C., Beck, J.D., Genco, R.J., 2015. Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. J. Periodontol. 86, 611-622.

Ellegaard, K.M., Klasson, L., Andersson, S.G., 2013. Testing the reproducibility of multiple displacement amplification on genomes of clonal endosymbiont populations. PLoS One 8, e82319.

Endersen, L., O'Mahony, J., Hill, C., Ross, R.P., McAuliffe, O., Coffey, A., 2014. Phage therapy in the food industry. Annu. Rev. Food Sci. Technol. 5, 327-349.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575-1584.

Farrell, J.J., Zhang, L., Zhou, H., Chia, D., Elashoff, D., Akin, D., Paster, B.J., Joshipura, K., Wong, D.T.W., 2012. Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. Gut 61, 582-588.

Fejerskov, O., 2004. Changing paradigms in concepts on dental caries: consequences for oral health care. Caries Res. 38, 182-191.

Fernández, L., Rodríguez, A., García, P., 2018. Phage or foe: an insight into the impact of viral predation on microbial communities. ISME J. 12, 1171-1179.

Ferrer, M.D., Mira, A., 2016. Oral Biofilm Architecture at the Microbial Scale. Trends Microbiol. 24, 246-248.

Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J.A., Schoenfeld, D., Nomicos, E., Park, M., NIH Intramural Sequencing Center Comparative Sequencing Program, Kong, H.H., Segre, J.A., 2013. Topographic diversity of fungal and bacterial communities in human skin. Nature 498, 367-370.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, 279-285.

Fischer, M.G., Allen, M.J., Wilson, W.H., Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc. Natl. Acad. Sci. U. S. A. 107, 19508-19513.

Fitzgerald, R.J., Keyes, P.H., 1960. Demonstration of the etiologic role of streptococci in experimental caries in the hamster. J. Am. Dent. Assoc. 61, 9-19.

Flores, C.O., Meyer, J.R., Valverde, S., Farr, L., Weitz, J.S., 2011. Statistical structure of host–phage interactions. Proc. Natl. Acad. Sci. U. S. A. 108, 288-297.

Flores, C.O., Valverde, S., Weitz, J.S., 2013. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. ISME J. 7, 520-532.

Forti, F., Roach, D.R., Cafora, M., Pasini, M.E., Horner, D.S., Fiscarelli, E.V., Rossitto, M., Cariani, L., Briani, F., Debarbieux, L., Ghisotti, D., 2018. Design of a broad-range bacteriophage cocktail that reduces Pseudomonas aeruginosa biofilms and treats acute infections in two animal models. Antimicrob. Agents Chemother. 62, e02573-17.

Frank, J.A., Reich, C.I., Sharma, S., Weisbaum, J.S., Wilson, B.A., Olsen, G.J., 2008. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Appl. Environ. Microbiol. 74, 2461-2470.

Froussard, P., 1992. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. Nucleic Acids Res. 20, 2900.

Fuhrman, J.A., 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399, 541-548.

Furfaro, L.L., Payne, M.S., Chang, B.J., 2018. Bacteriophage therapy: clinical trials and regulatory hurdles. Front. Cell. Infect. Microbiol. 8, 376.

Furuta, R.A., Sakamoto, H., Kuroishi, A., Yasiui, K., Matsukura, H., Hirayama, F., 2015. Metagenomic profiling of the viromes of plasma collected from blood donors with elevated serum alanine aminotransferase levels. Transfusion 55, 1889-1899.

García, J.L., García, E., Arrarás, A., García, P., Ronda, C., López, R., 1987. Cloning, purification, and biochemical characterization of the pneumococcal bacteriophage Cp-1 lysin. J. Virol. 61, 2573-2580.

Ghannoum, M.A., Jurevic, R.J., Mukherjee, P.K., Cui, F., Sikaroodi, M., Naqvi, A., Gillevet, P.M., 2010. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog. 6, e1000713.

Glöckner, F.O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G., Yarza, P., Peplies, J., Westram, R., Ludwig, W., 2017. 25 years of serving the community with ribosomal RNA gene reference

databases and tools. J. Biotechnol. 261, 169-176.

Goltsman, D.S.A., Sun, C.L., Proctor, D.M., DiGiulio, D.B., Robaczewska, A., Thomas, B.C., Shaw, G.M., Stevenson, D.K., Holmes, S.P., Banfield, J.F., Relman, D.A., 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res. 28, 1467-1480.

Gomar-Vercher, S., Simón-Soro, A., Montiel-Company, J.M., Almerich-Silla, J.M., Mira, A., 2018. Stimulated and unstimulated saliva samples have significantly different bacterial profiles. PLoS One 13, e0198021.

Grandgirard, D., Loeffler, J.M., Fischetti, V.A., Leib, S.L., 2008. Phage lytic enzyme Cpl-1 for antibacterial therapy in experimental pneumococcal meningitis. J. Infect. Dis. 197, 1519-1522.

Grard, G., Fair, J.N., Lee, D., Slikas, E., Steffen, I., Muyembe, J.-J., Sittler, T., Veeraraghavan, N., Ruby, J.G., Wang, C., Makuwa, M., Mulembakani, P., Tesh, R.B., Mazet, J., Rimoin, A.W., Taylor, T., Schneider, B.S., Simmons, G., Delwart, E., Wolfe, N.D., Chiu, C.Y., Leroy, E.M., 2012. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog. 8, e1002924.

Grinde, B., 2013. Herpesviruses: latency and reactivation - viral strategies and host response. J. Oral Microbiol. 5, 22766.

Grissa, I., Vergnaud, G., Pourcel, C., 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 35, 52-57.

Gross, E.L., Beall, C.J., Kutsch, S.R., Firestone, N.D., Leys, E.J., Griffen, A.L., 2012. Beyond Streptococcus mutans: dental caries onset linked to multiple species by 16S rRNA community analysis. PLoS One 7, e47722.

Grunenwald, C.M., Bennett, M.R., Skaar, E.P., 2018. Nonconventional therapeutics against Staphylococcus aureus. Microbiol Spectr 6.

Guerin, E., Shkoporov, A., Stockdale, S., Clooney, A.G., 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. bioRxiv 24, 653-664.

Gu, Q., Li, P., 2016. Biosynthesis of vitamins by probiotic bacteria, en: Rao, V., Rao, L.G. (Eds.), Probiotics and prebiotics in human nutrition and health. InTech, pp. 135-148.

Gutiérrez, D., Fernández, L., Rodríguez, A., García, P., 2018. Are phage lytic proteins the secret weapon to kill Staphylococcus aureus? mBio 9, e01923-17.

Hajishengallis, G., 2015. Periodontitis: from microbial immune subversion to systemic inflammation. Nat. Rev. Immunol. 15, 30-44.

Hajishengallis, G., Lamont, R.J., 2012. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol. Oral Microbiol. 27, 409-419.

Hall, R.J., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S., Strydom, H., Moore, N.E., Ren, X., Huang, Q.S., Carter, P.E., Peacey, M., 2014. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. J. Virol. Methods 195, 194-204.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, 245-249.

Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodkinson, B.P., SanMiguel, A.J., Minot, S., Bushman, F.D., Grice, E.A., 2015. The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. MBio 6, e01578-15.

Hansen, M.A., s. f. biopieces. Github.

Han, T., Chang, C.-W., Kwekel, J.C., Chen, Y., Ge, Y., Martinez-Murillo, F., Roscoe, D., Težak, Z., Philip, R., Bijwaard, K., Fuscoe, J.C., 2012. Characterization of whole genome amplified (WGA) DNA for use in genotyping assay development. BMC Genomics 13, 217.

Hauer, M.H., Gasser, S.M., 2017. Chromatin and nucleosome dynamics in DNA damage and repair. Genes Dev.

31, 2204-2221.

He, J., Tu, Q., Ge, Y., Qin, Y., Cui, B., Hu, X., Wang, Y., Deng, Y., Wang, K., Van Nostrand, J.D., Li, J., Zhou, J., Li, Y., Zhou, X., 2018. Taxonomic and functional analyses of the supragingival microbiome from caries-affected and caries-free hosts. Microb. Ecol. 75, 543-554.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.-Y., Dorrestein, P.C., Esquenazi, E., Hunter, R.C., Cheng, G., Nelson, K.E., Lux, R., Shi, W., 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc. Natl. Acad. Sci. U. S. A. 112, 244-249.

Hijazi, K., Lowe, T., Meharg, C., Berry, S.H., Foley, J., Hold, G.L., 2015. Mucosal microbiome in patients with recurrent aphthous stomatitis. J. Dent. Res. 94, 87-94.

Hoeijmakers, W.A.M., Bártfai, R., Françoijs, K.-J., Stunnenberg, H.G., 2011. Linear amplification for deep sequencing. Nat. Protoc. 6, 1026-1036.

Holmfeldt, K., Odić, D., Sullivan, M.B., Middelboe, M., Riemann, L., 2012. Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. Appl. Environ. Microbiol. 78, 892-894.

Holtz, L.R., Cao, S., Zhao, G., Bauer, I.K., Denno, D.M., Klein, E.J., Antonio, M., Stine, O.C., Snelling, T.L., Kirkwood, C.D., Wang, D., 2014. Geographic variation in the eukaryotic virome of human diarrhea. Virology 468-470, 556-564.

Hoyles, L., McCartney, A.L., Neve, H., Gibson, G.R., Sanderson, J.D., Heller, K.J., van Sinderen, D., 2014. Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res. Microbiol. 165, 803-812.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol. Biol. Evol. 34, 2115-2122.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., Von Mering, C., Bork, P., 2016. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, 286-293.

Hugenholtz, P., Tyson, G.W., Webb, R.I., Wagner, A.M., Blackall, L.L., 2001. Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. Appl. Environ. Microbiol. 67, 411-419.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F., 2016. A new view of the tree of life. Nat Microbiol 1, 16048.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., Le Mercier, P., 2011. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 39, 576-582.

Human Microbiome Project Consortium, 2012. A framework for human microbiome research. Nature 486, 215-221.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R., 2016. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput. Biol. 12, e1004957.

Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061-1067.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.

Indiani, C., Sauve, K., Raz, A., Abdelhady, W., Xiong, Y.Q., Cassino, C., Bayer, A.S., Schuch, R., 2019. The anti-

staphylococcal lysin, CF-301, activates key host factors in human blood to potentiate MRSA bacteriolysis. Antimicrobial Agents and Chemotherapy 63, e02291-18.

Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M., Piazza, P., Bowden, R.J., Paten, B., Mwaigwisya, S., Batty, E.M., Simpson, J.T., Snutch, T.P., Birney, E., Buck, D., Goodwin, S., Jansen, H.J., O'Grady, J., Olsen, H.E., MinION Analysis and Reference Consortium, 2015. MinION analysis and reference consortium: Phase 1 data release and analysis. F1000Res. 4, 1075.

Jado, I., López, R., García, E., Fenoll, A., Casal, J., García, P., Spanish Pneumococcal Infection Study Network, 2003. Phage lytic enzymes as therapy for antibiotic-resistant Streptococcus pneumoniae infection in a murine sepsis model. J. Antimicrob. Chemother. 52, 967-973.

Jakovljevic, A., Andric, M., Knezevic, A., Soldatovic, I., Nikolic, N., Karalic, D., Milasin, J., 2015. Human Cytomegalovirus and Epstein-Barr Virus Genotypes in Apical Periodontitis Lesions. J. Endod. 41, 1847-1851.

Jean, J., Goldberg, S., Khare, R., Bailey, L.C., Forrest, C.B., Hajishengallis, E., Koo, H., 2018. Retrospective analysis of Candida-related conditions in infancy and early childhood caries. Pediatr. Dent. 40, 131-135.

Jenkinson, H.F., Lamont, R.J., 2005. Oral microbial communities in sickness and in health. Trends Microbiol. 13, 589-595.

Jiang, S.C., Paul, J.H., 1998. Significance of lysogeny in the marine environment: studies with isolates and a model of lysogenic phage production. Microb. Ecol. 35, 235-243.

Jorth, P., Turner, K.H., Gumus, P., Nizam, N., Buduneli, N., Whiteley, M., 2014. Metatranscriptomics of the human oral microbiome during health and disease. MBio 5, e01012-14.

Kanamaru, S., Ishiwata, Y., Suzuki, T., Rossmann, M.G., Arisaka, F., 2005. Control of bacteriophage T4 tail lysozyme activity during the infection process. J. Mol. Biol. 346, 1013-1020.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, 353-361.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, 457-462.

Karlsson, O.E., Belák, S., Granberg, F., 2013. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. Biosecur. Bioterror. 11, 227-234.

Kato, I., Vasquez, A., Moyerbrailean, G., Land, S., Djuric, Z., Sun, J., Lin, H.-S., Ram, J.L., 2017. Nutritional correlates of human oral microbiome. J. Am. Coll. Nutr. 36, 88-98.

Kernbauer, E., Ding, Y., Cadwell, K., 2014. An enteric virus can replace the beneficial function of commensal bacteria. Nature 516, 94-98.

Kholy, K.E., Genco, R.J., Van Dyke, T.E., 2015. Oral infections and cardiovascular disease. Trends Endocrinol. Metab. 26, 315-321.

Kim, K.-H., Bae, J.-W., 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl. Environ. Microbiol. 77, 7663-7668.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S.W., Kim, M.-S., Sung, Y., Jeon, C.O., Oh, H.-M., Bae, J.-W., 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl. Environ. Microbiol. 74, 5975-5985.

Kim, M.-S., Park, E.-J., Roh, S.W., Bae, J.-W., 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062-8070.

Kinane, D.F., Stathopoulou, P.G., Papapanou, P.N., 2017. Periodontal diseases. Nature Reviews Disease Primers

17038.

Kleiner, M., Hooper, L.V., Duerkop, B.A., 2015. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics 16, 7.

Klieve, A.V., Swain, R.A., 1993. Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. Appl. Environ. Microbiol. 59, 2299-2303.

Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobián-Güemes, A.G., Coutinho, F.H., Dinsdale, E.A., Felts, B., Furby, K.A., George, E.E., Green, K.T., Gregoracci, G.B., Haas, A.F., Haggerty, J.M., Hester, E.R., Hisakawa, N., Kelly, L.W., Lim, Y.W., Little, M., Luque, A., McDole-Somera, T., McNair, K., de Oliveira, L.S., Quistad, S.D., Robinett, N.L., Sala, E., Salamon, P., Sanchez, S.E., Sandin, S., Silva, G.G.Z., Smith, J., Sullivan, C., Thompson, C., Vermeij, M.J.A., Youle, M., Young, C., Zgliczynski, B., Brainard, R., Edwards, R.A., Nulton, J., Thompson, F., Rohwer, F., 2016. Lytic to temperate switching of viral communities. Nature 531, 466-470.

Kohl, C., Brinkmann, A., Dabrowski, P.W., Radonić, A., Nitsche, A., Kurth, A., 2015. Protocol for metagenomic virus detection in clinical specimens. Emerg. Infect. Dis. 21, 48-57.

Koonin, E.V., Makarova, K.S., Zhang, F., 2017. Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. 37, 67-78.

Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M., Clancy, T.E., Chung, D.C., Lochhead, P., Hold, G.L., El-Omar, E.M., Brenner, D., Fuchs, C.S., Meyerson, M., Garrett, W.S., 2013. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe 14, 207-215.

Kouidhi, B., Zmantar, T., Mahdouani, K., Hentati, H., Bakhrouf, A., 2011. Antibiotic resistance and adhesion properties of oral Enterococci associated to dental caries. BMC Microbiol. 11, 155.

Kowalski, K., Mulak, A., 2019. Brain-gut-microbiota axis in Alzheimer's disease. J. Neurogastroenterol. Motil. 25, 48-60.

Krajmalnik-Brown, R., Ilhan, Z.-E., Kang, D.-W., DiBaise, J.K., 2012. Effects of gut microbes on nutrient absorption and energy regulation. Nutr. Clin. Pract. 27, 201-214.

Kreimer, A.R., Clifford, G.M., Boyle, P., Franceschi, S., 2005. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review. Cancer Epidemiol. Biomarkers Prev. 14, 467-475.

Krupovic, M., Forterre, P., 2011. Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. PLoS One 6, e19893.

Kupczok, A., Landan, G., Dagan, T., 2015. The contribution of genetic recombination to CRISPR array evolution. Genome Biol. Evol. 7, 1925-1939.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5, R12.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., Ussery, D.W., 2015. Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics 15, 141-161.

Langmead, B., Salzberg, S.L., 2013. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359.

Lasken, R.S., 2009. Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem. Soc. Trans. 37, 450-453.

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., Carissimi, C., 2018. Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. OMICS 22, 248-254.

Lawley, T.D., Clare, S., Walker, A.W., Stares, M.D., Connor, T.R., Raisen, C., Goulding, D., Rad, R., Schreiber, F., Brandt, C., Deakin, L.J., Pickard, D.J., Duncan, S.H., Flint, H.J., Clark, T.G., Parkhill, J., Dougan, G., 2012. Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS Pathog. 8, e1002995.

Lederberg, J., McCray, A.T., 2001. Ome SweetOmics--A genealogical treasury of words. Scientist 15, 8-8.

Lewandowska, D.W., Zagordi, O., Geissberger, F.-D., Kufner, V., Schmutz, S., Böni, J., Metzner, K.J., Trkola, A., Huber, M., 2017. Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. Microbiome 5, 94.

Ley, R.E., Turnbaugh, P.J., Klein, S., Gordon, J.I., 2006. Microbial ecology: human gut microbes associated with obesity. Nature 444, 1022-1023.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, J., Helmerhorst, E.J., Leone, C.W., Troxler, R.F., Yaskell, T., Haffajee, A.D., Socransky, S.S., Oppenheim, F.G., 2004. Identification of early microbial colonizers in human dental biofilm. J. Appl. Microbiol. 97, 1311-1318.

Li, L., Deng, X., Mee, E.T., Collot-Teixeira, S., Anderson, R., Schepelmann, S., Minor, P.D., Delwart, E., 2015. Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. J. Virol. Methods 213, 139-146.

Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M., Warner, B.B., Tarr, P.I., Wang, D., Holtz, L.R., 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat. Med. 21, 1228-1234.

Lim, Y., Totsika, M., Morrison, M., Punyadeera, C., 2017. The saliva microbiome profiles are minimally affected by collection method or DNA extraction protocols. Sci. Rep. 7, 8523.

Lloyd-Price, J., Abu-Ali, G., Huttenhower, C., 2016. The healthy human microbiome. Genome Med. 8, 51.

Loeffler, J.M., 2001. Rapid killing of Streptococcus pneumoniae with a bacteriophage cell wall hydrolase. Science 294, 2170-2172.

Loesche, W.J., Rowan, J., Straffon, L.H., Loos, P.J., 1975. Association of Streptococcus mutans with human dental decay. Infect. Immun. 11, 1252-1260.

Lood, R., Winer, B.Y., Pelzek, A.J., Diez-Martinez, R., Thandar, M., Euler, C.W., Schuch, R., Fischetti, V.A., 2015. Novel phage lysin capable of killing the multidrug-resistant Gram-negative bacterium Acinetobacter baumannii in a mouse bacteremia model. Antimicrobial Agents and Chemotherapy 59, 1983-1991.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A., Alcamí, A., 2009. High diversity of the viral community from an Antarctic lake. Science 326, 858-861.

López-López, A., Camelo-Castillo, A., Ferrer, M.D., Simon-Soro, Á., Mira, A., 2017. Health-associated niche inhabitants as oral probiotics: the case of Streptococcus dentisani. Front. Microbiol. 8, 379.

López-Pérez, M., Haro-Moreno, J.M., Gonzalez-Serrano, R., Parras-Moltó, M., Rodriguez-Valera, F., 2017. Genome diversity of marine phages recovered from mediterranean metagenomes: Size matters. PLoS Genet. 13, e1007018.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964.

Lv, F., Chen, S., Wang, L., Jiang, R., Tian, H., Li, J., Yao, Y., Zhuo, C., 2017. The role of microbiota in the pathogenesis of schizophrenia and major depressive disorder and the possibility of targeting microbiota as a treatment option. Oncotarget 8, 100899-100907.

Ly, M., Abeles, S.R., Boehm, T.K., Robles-Sikisaka, R., Naidu, M., Santiago-Rodriguez, T., Pride, D.T., 2014.

Altered oral viral ecology in association with periodontal disease. MBio 5, e01133-14.

Ly, M., Jones, M.B., Abeles, S.R., Santiago-Rodriguez, T.M., Gao, J., Chan, I.C., Ghose, C., Pride, D.T., 2016. Transmission of viruses via our microbiomes. Microbiome 4, 64.

Mager, D.L., Haffajee, A.D., Devlin, P.M., Norris, C.M., Posner, M.R., Goodson, J.M., 2005. The salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. J. Transl. Med. 3, 27.

Mamedov, T.G., Pienaar, E., Whitney, S.E., TerMaat, J.R., Carvill, G., Goliath, R., Subramanian, A., Viljoen, H.J., 2008. A fundamental study of the PCR amplification of GC-rich DNA templates. Comput. Biol. Chem. 32, 452-457.

Manrique, P., Dills, M., Young, M.J., 2017. The human gut phage community and its implications for health and disease. Viruses 9.

Marcenes, W., Kassebaum, N.J., Bernabé, E., Flaxman, A., Naghavi, M., Lopez, A., Murray, C.J.L., 2013. Global burden of oral conditions in 1990-2010: a systematic analysis. J. Dent. Res. 92, 592-597.

Marchini, L., Campos, M.S., Silva, A.M., Paulino, L.C., Nobrega, F.G., 2007. Bacterial diversity in aphthous ulcers. Oral Microbiol. Immunol. 22, 225-231.

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S.W., Wommack, K.E., 2014. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2, 3.

Marsh, P.D., 2010. Microbiology of dental plaque biofilms and their role in oral health and caries. Dent. Clin. North Am. 54, 441-454.

Marsh, P.D., 2004. Dental plaque as a microbial biofilm. Caries Res. 38, 204-211.

Marsh, P.D., 2003. Are dental diseases examples of ecological catastrophes? Microbiology 149, 279-294.

Matarazzo, F., Ribeiro, A.C., Feres, M., Faveri, M., Mayer, M.P.A., 2011. Diversity and quantitative analysis of Archaea in aggressive periodontitis and periodontally healthy subjects. J. Clin. Periodontol. 38, 621-627.

McGowin, C.L., Pyles, R.B., 2010. Mucosal treatments for herpes simplex virus: insights on targeted immunoprophylaxis and therapy. Future Microbiol. 5, 15-22.

McMurdie, P.J., Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Międzybrodzki, R., Borysowski, J., Weber-Dąbrowska, B., Fortuna, W., Letkiewicz, S., Szufnarowski, K., Pawełczyk, Z., Rogóż, P., Kłak, M., Wojtasik, E., Górski, A., 2012. Clinical aspects of phage therapy. Adv. Virus Res. 83, 73-121.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., Bushman, F.D., 2013. Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. U. S. A. 110, 12450-12455.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., Bushman, F.D., 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 21, 1616-1625.

Mira, A., Simon-Soro, A., Curtis, M.A., 2017. Role of microbial communities in the pathogenesis of periodontal diseases and caries. J. Clin. Periodontol. 44, 23-38.

Mitra, A., Skrzypczak, M., Ginalski, K., Rowicka, M., 2015. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. PLoS One 10, e0120520.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., Ghai, R., 2013. Expanding the marine virosphere using metagenomics. PLoS Genet. 9, e1003987.

Monk, A.B., Rees, C.D., Barrow, P., Hagens, S., Harper, D.R., 2010. Bacteriophage applications: where are we now? Lett. Appl. Microbiol. 51, 363-369.

Morgulis, A., Gertz, E.M., Schäffer, A.A., Agarwala, R., 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. 13, 1028-1040.

Morhart, R.E., Fitzgerald, R.J., 1976. Nutritional determinants of the ecology of the oral flora. Dent. Clin. North Am. 20, 473-489.

Motley, S.T., Picuri, J.M., Crowder, C.D., Minich, J.J., Hofstadler, S.A., Eshoo, M.W., 2014. Improved multiple displacement amplification (iMDA) and ultraclean reagents. BMC Genomics 15, 443.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K.E., Venter, J.C., Telenti, A., 2017. The blood DNA virome in 8,000 humans. PLoS Pathog. 13, e1006292.

Naidu, M., Robles-Sikisaka, R., Abeles, S.R., Boehm, T.K., Pride, D.T., 2014. Characterization of bacteriophage communities and CRISPR profiles from dental plaque. BMC Microbiol. 14, 175.

Nannapaneni, R., Soni, K.A., 2015. Use of bacteriophages to remove biofilms of listeria monocytogenes and other foodborne bacterial pathogens in the food environment, en: Pometto, A.L., III, Demirci, A. (Eds.), Biofilms in the Food Environment. John Wiley & Sons, Ltd, Chichester, UK, pp. 131-144.

Nash, A.K., Auchtung, T.A., Wong, M.C., Smith, D.P., Gesell, J.R., Ross, M.C., Stewart, C.J., Metcalf, G.A., Muzny, D.M., Gibbs, R.A., Ajami, N.J., Petrosino, J.F., 2017. The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome 5, 153.

Nasidze, I., Li, J., Quinque, D., Tang, K., 2009. Global diversity in the human salivary microbiome. Genome Res 19, 636-643.

Natah, S.S., Konttinen, Y.T., Enattah, N.S., Ashammakhi, N., Sharkey, K.A., Häyrinen-Immonen, R., 2004. Recurrent aphthous ulcers today: a review of the growing knowledge. Int. J. Oral Maxillofac. Surg. 33, 221-234.

Nazir, M.A., 2017. Prevalence of periodontal disease, its association with systemic diseases and prevention. Int. J. Health Sci. 11, 72-80.

Ndegwa, N., Ploner, A., Liu, Z., Roosaar, A., Axéll, T., Ye, W., 2018. Association between poor oral health and gastric cancer: A prospective cohort study. Int. J. Cancer 143, 2281-2288.

Nelson, D., Loomis, L., Fischetti, V.A., 2001. Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. Proc. Natl. Acad. Sci. U. S. A. 98, 4107-4112.

Nelson-Filho, P., Borba, I.G., Mesquita, K.S.F. de, Silva, R.A.B., Queiroz, A.M. de, Silva, L.A.B., 2013. Dynamics of microbial colonization of the oral cavity in newborns. Braz. Dent. J. 24, 415-419.

Ning, L., Li, Z., Wang, G., Hu, W., Hou, Q., Tong, Y., Zhang, M., Chen, Y., Qin, L., Chen, X., Man, H.-Y., Liu, P., He, J., 2015. Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. Sci. Rep. 5, 11415.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., Goto, S., 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33, 2379-2380.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell 160, 447-460.

Nyvad, B., Kilian, M., 1987. Microbiology of the early colonization of human enamel and root surfaces in vivo. Scand. J. Dent. Res. 95, 369-380.

Obeso, J.M., Martínez, B., Rodríguez, A., García, P., 2008. Lytic activity of the recombinant staphylococcal bacteriophage ΦH5 endolysin active against Staphylococcus aureus in milk. Int. J. Food Microbiol. 128, 212-218.

Ogilvie, L.A., Bowler, L.D., Caplin, J., Dedi, C., Diston, D., Cheek, E., Taylor, H., Ebdon, J.E., Jones, B.V., 2013. Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. Nat.

Commun. 4, 2420.

Oh, J., Byrd, A.L., Park, M., NISC Comparative Sequencing Program, Kong, H.H., Segre, J.A., 2016. Temporal stability of the human skin microbiome. Cell 165, 854-866.

Oksanen, J., Blanchet, F.G., Kindt, R., 2011. vegan: Community ecology package. R package.

O'Neill, J., 2014. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Review on antimicrobial resistance 2-16.

Package, S., View, M., Bairoch, A., Project, H.G., 2000. Bioinformatics. Nat. Biotechnol. 18, IT31.

Panda, S., El khader, I., Casellas, F., López Vivancos, J., García Cors, M., Santiago, A., Cuenca, S., Guarner, F., Manichanh, C., 2014. Short-term effect of antibiotics on human gut microbiota. PLoS One 9, e95476.

Parfrey, L.W., Walters, W.A., Lauber, C.L., Clemente, J.C., Berg-Lyons, D., Teiling, C., Kodira, C., Mohiuddin, M., Brunelle, J., Driscoll, M., Fierer, N., Gilbert, J.A., Knight, R., 2014. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Front. Microbiol. 5, 298.

Parras-Moltó, M., López-Bueno, A., 2018. Methods for enrichment and sequencing of oral viral assemblages: saliva, oral mucosa, and dental plaque viromes. Methods Mol. Biol. 1838, 143-161.

Parras-Moltó, M., Suárez-Rodríguez, P., Eguia, A., Aguirre-Urizar, J.M., López-Bueno, A., 2014. Genome sequence of two novel species of torque teno minivirus from the human oral cavity. Genome Announc. 2, 5-6.

Paul, J.H., 2008. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME J. 2, 579-589.

Payne, A., Holmes, N., Rakyan, V., Loose, M., 2018. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. BioRxiv 1-28.

Pedersen, A., Hornsleth, A., 1993. Recurrent aphthous ulceration: a possible clinical manifestation of reactivation of varicella zoster or cytomegalovirus infection. J. Oral Pathol. Med. 22, 64-68.

Perepanova, T.S., Darbeeva, O.S., Kotliarova, G.A., Kondrat'eva, E.M., Maĭskaia, L.M., Malysheva, V.F., Baĭguzina, F.A., Grishkova, N.V., 1995. The efficacy of bacteriophage preparations in treating inflammatory urologic diseases. Urol. Nefrol. 14-17.

Pérez-Brocal, V., Moya, A., 2018. The analysis of the oral DNA virome reveals which viruses are widespread and rare among healthy young adults in Valencia (Spain). PLoS One 13, e0191867.

Petersen, P.E., 2003. The World Oral Health report 2003: continuous improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme. Community Dent. Oral Epidemiol. 31, 3-24.

Picher, Á.J., Budeus, B., Wafzig, O., Krüger, C., García-Gómez, S., Martínez-Jiménez, M.I., Díaz-Talavera, A., Weber, D., Blanco, L., Schneider, A., 2016. TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. Nat. Commun. 7, 13296.

Pickrell, J.K., Gaffney, D.J., Gilad, Y., Pritchard, J.K., 2011. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics 27, 2144-2146.

Pinto, A.J., Raskin, L., 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS One 7, e43093.

Pirnay, J.-P., Verbeken, G., Ceyssens, P.-J., Huys, I., De Vos, D., Ameloot, C., Fauconnier, A., 2018. The magistral phage. Viruses 10, 64.

Plotree, D., Plotgram, D., 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics 5, 6.

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Sonthirod, C., Naktang, C., Thongtham, N.,

Tangphatsornruang, S., 2017. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. Sci. Rep. 7, 2774.

Popgeorgiev, N., Colson, P., Thuret, I., Chiarioni, J., Gallian, P., Raoult, D., Desnues, C., 2013. Marseillevirus prevalence in multitransfused patients suggests blood transmission. J. Clin. Virol. 58, 722-725.

Porter, S.R., Scully, C., Pedersen, A., 1998. Recurrent aphthous stomatitis. Crit. Rev. Oral Biol. Med. 9, 306-321.

Pride, D.T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R.A., Loomer, P., Armitage, G.C., Relman, D.A., 2011a. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. ISME J. 6, 915-926.

Pride, D.T., Sun, C.L., Salzman, J., Rao, N., Loomer, P., Armitage, G.C., Banfield, J.F., Relman, D.A., 2011b. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. Genome Res. 21, 126-136.

Proctor, D.M., Relman, D.A., 2017. The landscape ecology and microbiota of the human nose, mouth, and throat. Cell Host Microbe 21, 421-432.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188-7196.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33, 501-504.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S.D., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J., 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55-60.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J.M., Kennedy, S.P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S.D., Zheng, S., Li, L., 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature 513, 59-64.

Quick, J., Quinlan, A.R., Loman, N.J., 2014. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. Gigascience 3, 22.

Quigley, E.M.M., 2013. Gut bacteria in health and disease. Gastroenterol. Hepatol. 9, 560-569.

Quirós, P., Colomer-Lluch, M., Martínez-Castillo, A., Miró, E., Argente, M., Jofre, J., Navarro, F., Muniesa, M., 2014. Antibiotic resistance genes in the bacteriophage DNA fraction of human fecal samples. Antimicrob. Agents Chemother. 58, 606-609.

Rani, A., Ranjan, R., McGee, H.S., Metwally, A., Hajjiri, Z., Brennan, D.C., Finn, P.W., Perkins, D.L., 2016. A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. Sci Rep 6: 33327 6, 13.

Rashel, M., Uchiyama, J., Ujihara, T., 2007. Efficient elimination of multidrug-resistant Staphylococcus aureus by cloned lysin derived from bacteriophage ϕMR11. J Infect Dis 196, 1237-1247.

R Core Team (2017), 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reyes, A., Blanton, L.V., Cao, S., Zhao, G., Manary, M., Trehan, I., Smith, M.I., Wang, D., Virgin, H.W., Rohwer, F., Gordon, J.I., 2015. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc. Natl. Acad. Sci. U. S. A. 112, 11941-11946.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., Gordon, J.I., 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334-338.

Rhee, M., Light, Y.K., Meagher, R.J., Singh, A.K., 2016. Digital droplet multiple displacement amplification (ddMDA) for whole genome sequencing of limited DNA samples. PLoS One 11, e0153699.

Robles-Sikisaka, R., Ly, M., Boehm, T., Naidu, M., Salzman, J., Pride, D.T., 2013. Association between living environment and human oral viral ecology. ISME J. 7, 1710-1724.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., Felts, B., Haynes, M., Liu, H., Lipson, D., Mahaffy, J., Martin-Cuadrado, A.B., Mira, A., Nulton, J., Pasić, L., Rayhawk, S., Rodriguez-Mueller, J., Rodriguez-Valera, F., Salamon, P., Srinagesh, S., Thingstad, T.F., Tran, T., Thurber, R.V., Willner, D., Youle, M., Rohwer, F., 2010. Viral and microbial community dynamics in four aquatic environments. ISME J. 4, 739-751.

Rodriguez-Cerrato, V., Garcia, P., Huelves, L., Garcia, E., del Prado, G., Gracia, M., Ponte, C., Lopez, R., Soriano, F., 2007. Pneumococcal LytA Autolysin, a Potent Therapeutic Agent in Experimental Peritonitis-Sepsis Caused by Highly -Lactam-Resistant Streptococcus pneumoniae. Antimicrobial Agents and Chemotherapy 51, 3371-3373.

Rohwer, F., Edwards, R., 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol. 184, 4529-4535.

Rohwer, F., Thurber, R.V., 2009. Viruses manipulate the marine environment. Nature 459, 207-212.

Rosario, K., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. Arch. Virol. 157, 1851-1871.

Rose, T., Verbeken, G., Vos, D.D., Merabishvili, M., Vaneechoutte, M., Lavigne, R., Jennes, S., Zizi, M., Pirnay, J.-P., 2014. Experimental phage therapy of burn wound infection: difficult first steps. Int. J. Burns Trauma 4, 66-73.

Rosseel, T., Van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M., Höper, D., 2013. The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. PLoS One 8, e76144.

Ross, R.S., Viazov, S., Runde, V., Schaefer, U.W., Roggendorf, M., 1999. Detection of TT virus DNA in specimens other than blood. J. Clin. Virol. 13, 181-184.

Roux, S., Enault, F., Hurwitz, B.L., Sullivan, M.B., 2015a. VirSorter: mining viral signal from microbial genomic data. PeerJ 3, e985.

Roux, S., Hallam, S.J., Woyke, T., Sullivan, M.B., 2015b. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife 4, 1-20.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B., Coleman, M.L., Breitbart, M., Sullivan, M.B., 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ 4, e2777.

Rubinstein, M.R., Wang, X., Liu, W., Hao, Y., Cai, G., Han, Y.W., 2013. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe 14, 195-206.

Salmond, G.P.C., Fineran, P.C., 2015. A century of the phage: past, present and future. Nat. Rev. Microbiol. 13, 777-786.

Santiago-Rodriguez, T.M., Ly, M., Bonilla, N., Pride, D.T., 2015. The human urine virome in association with urinary tract infections. Front. Microbiol. 6, 14.

Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592.

Schmelcher, M., Donovan, D.M., Loessner, M.J., 2012. Bacteriophage endolysins as novel antimicrobials. Future

Microbiol. 7, 1147-1171.

Schmidt, T.M., DeLong, E.F., Pace, N.R., 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. 173, 4371-4378.

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-864.

Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., Segall, A.M., Taplitz, R., Smith, D.M., Kerr, K., Kumaraswamy, M., Nizet, V., Lin, L., McCauley, M.D., Strathdee, S.A., Benson, C.A., Pope, R.K., Leroux, B.M., Picel, A.C., Mateczun, A.J., Cilwa, K.E., Regeimbal, J.M., Estrella, L.A., Wolfe, D.M., Henry, M.S., Quinones, J., Salka, S., Bishop-Lilly, K.A., Young, R., Hamilton, T., 2017. Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant Acinetobacter baumannii infection. Antimicrob. Agents Chemother. 61.

Schubert, A.M., Sinani, H., Schloss, P.D., 2015. Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against Clostridium difficile. MBio 6, e00974.

Schuch, R., Nelson, D., Fischetti, V.A., 2002. A bacteriolytic agent that detects and kills Bacillus anthracis. Nature 418, 884-889.

Sender, R., Fuchs, S., Milo, R., 2016. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol. 14, e1002533.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-2504.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R.Y., Partensky, F., Koonin, E.V., Wolf, Y.I., Nelson, N., Béjà, O., 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461, 258-262.

Sheehan, M.M., García, J.L., López, R., García, P., 1997. The lytic enzyme of the pneumococcal phage Dp-1: a chimeric lysin of intergeneric origin. Mol. Microbiol. 25, 717-725.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135-1145.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., Koonin, E.V., 2017. The CRISPR spacer space Is dominated by sequences from species-specific mobilomes. mBio 8, e01397-17.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.

Silveira, C.B., Rohwer, F.L., 2016. Piggyback-the-Winner in host-associated microbial communities. NPJ Biofilms Microbiomes 2, 16010.

Simón-Soro, A., Guillen-Navarro, M., Mira, A., 2014. Metatranscriptomics reveals overall active bacterial composition in caries lesions. J. Oral Microbiol. 6, 25443.

Simón-Soro, A., Mira, A., 2015. Solving the etiology of dental caries. Trends Microbiol. 23, 76-82.

Simón-Soro, A., Tomás, I., Cabrera-Rubio, R., Catalan, M.D., Nyvad, B., Mira, A., 2013. Microbial geography of the oral cavity. J. Dent. Res. 92, 616-621.

Simpkin, V.L., Renwick, M.J., Kelly, R., Mossialos, E., 2017. Incentivising innovation in antibiotic drug discovery and development: progress, challenges and next steps. J. Antibiot. 70, 1087-1096.

Slopek, S., Durlakowa, I., Weber-Dabrowska, B., Dabrowski, M., Kucharewicz-Krukowska, A., 1984. Results of bacteriophage treatment of suppurative bacterial infections. III. Detailed evaluation of the results obtained in further 150 cases. Arch. Immunol. Ther. Exp. 32, 317-335.

Slopek, S., Durlakowa, I., Weber-Dabrowska, B., Kucharewicz-Krukowska, A., Dabrowski, M., Bisikiewicz, R., 1983. Results of bacteriophage treatment of suppurative bacterial infections. I. General evaluation of the results. Arch. Immunol. Ther. Exp. 31, 267-291.

Smits, S.L., Schapendonk, C.M.E., van Beek, J., Vennema, H., Schürch, A.C., Schipper, D., Bodewes, R., Haagmans, B.L., Osterhaus, A.D.M.E., Koopmans, M.P., 2014. New viruses in idiopathic human diarrhea cases, the Netherlands. Emerg. Infect. Dis. 20, 1218-1222.

Socransky, S.S., Haffajee, A.D., 2005. Periodontal microbial ecology. Periodontology 2000 38, 135-187.

Socransky, S.S., Haffajee, A.D., Cugini, M.A., Smith, C., Kent, R.L., Jr, 1998. Microbial complexes in subgingival plaque. J. Clin. Periodontol. 25, 134-144.

Solonenko, S.A., Ignacio-Espinoza, J.C., Alberti, A., Cruaud, C., Hallam, S., Konstantinidis, K., Tyson, G., Wincker, P., Sullivan, M.B., 2013. Sequencing platform and library preparation choices impact viral metagenomes. BMC Genomics 14, 320.

Sørensen, T.J., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. I kommission hos E. Munksgaard.

Stahl, D.A., Lane, D.J., Olsen, G.J., Pace, N.R., 1984. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. Science 224, 409-411.

Staley, J.T., Konopka, A., 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39, 321-346.

Sullivan, M.J., Petty, N.K., Beatson, S.A., 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27, 1009-1010.

Summer, E.J., Liu, M., Gill, J.J., Grant, M., Chan-Cortes, T.N., Ferguson, L., Janes, C., Lange, K., Bertoli, M., Moore, C., Orchard, R.C., Cohen, N.D., Young, R., 2011. Genomic and functional analyses of Rhodococcus equi phages ReqiPepy6, ReqiPoco6, ReqiPine5, and ReqiDocB7. Appl. Environ. Microbiol. 77, 669-683.

Summers, W.C., 1999. Felix dHerelle and the origins of molecular biology. Yale University Press.

Suttle, C.A., 2007. Marine viruses—major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801.

Szafrański, S.P., Winkel, A., Stiesch, M., 2017. The use of bacteriophages to biocontrol oral biofilms. J. Biotechnol. 250, 29-44.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., Nishijima, M., 2014. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. PLoS One 9, e105592.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731-2739.

Tarakji, B., Gazal, G., Al-Maweri, S.A., Azzeghaiby, S.N., Alaizari, N., 2015. Guideline for the diagnosis and treatment of recurrent aphthous stomatitis for dental practitioners. J Int Oral Health 7, 74-80.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33-36.

Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N., 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nature Medicine 25, 667-678.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., Rohwer, F., 2009. Laboratory procedures to generate viral metagenomes. Nat. Protoc. 4, 470-483.

Towner, J.S., Sealy, T.K., Khristova, M.L., Albariño, C.G., Conlan, S., Reeder, S.A., Quan, P.-L., Lipkin, W.I., Downing, R., Tappero, J.W., Okware, S., Lutwama, J., Bakamutumaho, B., Kayiwa, J., Comer, J.A., Rollin, P.E., Ksiazek, T.G., Nichol, S.T., 2008. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 4, e1000212.

Treangen, T.J., Sommer, D.D., Angly, F.E., Koren, S., Pop, M., 2011. Next generation sequence assembly with AMOS, en: Curr Protoc Bioinformatics. p. Unit 11.8.

Trifonov E.N., 1990. Making sense of the human genome, en: R. H. Sarma and M. H. Sarma (Ed.), Structure and methods, human genome initiative and DNA recombination. Adenine Press, New York, pp. 69-77.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., Egholm, M., Henrissat, B., Heath, A.C., Knight, R., Gordon, J.I., 2009. A core gut microbiome in obese and lean twins. Nature 457, 480-484.

Twort, F.W., 1915. An investigation on the nature of ultra-microscopic viruses. Lancet 186, 1241-1243.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37-43.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115.

Valour, F., Sénéchal, A., Dupieux, C., Karsenty, J., Lustig, S., Breton, P., Gleizal, A., Boussel, L., Laurent, F., Braun, E., Chidiac, C., Ader, F., Ferry, T., 2014. Actinomycosis: etiology, clinical features, diagnosis, treatment, and management. Infect. Drug Resist. 7, 183-197.

van Houte, J., 1994. Role of micro-organisms in caries etiology. J. Dent. Res. 73, 672-681.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Verma, D., Garg, P.K., Dubey, A.K., 2018. Insights into the human oral microbiome. Arch. Microbiol. 200, 525-540.

Vianna, M.E., Conrads, G., Gomes, B.P.F.A., Horz, H.P., 2006. Identification and quantification of archaea involved in primary endodontic infections. J. Clin. Microbiol. 44, 1274-1282.

Victoria, J.G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., Delwart, E., 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. 83, 4642-4651.

Vielkind, P., Jentsch, H., Eschrich, K., Rodloff, A.C., Stingu, C.-S., 2015. Prevalence of Actinomyces spp. in patients with chronic periodontitis. Int. J. Med. Microbiol. 305, 682-688.

Wade, W.G., 2013. The oral microbiome in health and disease. Pharmacol. Res. 69, 137-143.

Waldor, M.K., Mekalanos, J.J., 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272, 1910-1914.

Wang, F., Kaplan, J.L., Gold, B.D., Bhasin, M.K., Ward, N.L., Kellermayer, R., Kirschner, B.S., Heyman, M.B., Dowd, S.E., Cox, S.B., Dogan, H., Steven, B., Ferry, G.D., Cohen, S.A., Baldassano, R.N., Moran, C.J., Garnett, E.A., Drake, L., Otu, H.H., Mirny, L.A., Libermann, T.A., Winter, H.S., Korolev, K.S., 2016. Detecting microbial dysbiosis associated with pediatric Crohn disease despite the high variability of the gut microbiota. Cell Rep. 14, 945-955.

Wang, J., Qi, J., Zhao, H., He, S., Zhang, Y., Wei, S., Zhao, F., 2013. Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. Sci. Rep. 3, 1843.

Wang, Q., Euler, C.W., Delaune, A., Fischetti, V.A., 2015. Using a novel lysin to help control Clostridium difficile infections. Antimicrob. Agents Chemother. 59, 7447-7457.

Wang, X., Seed, B., 2003. A PCR primer bank for quantitative gene expression analysis. Nucleic Acids Res. 31, e154.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191.

Weber-Dąbrowska, B., Mulczyk, M., Górski, A., 2001. Bacteriophage therapy of bacterial infections: an update of our institute's experience, en: Górski, A., Krotkiewski, H., Zimecki, M. (Eds.), Inflammation. Springer Netherlands, Dordrecht, pp. 201-209.

Weitz, J.S., Poisot, T., Meyer, J.R., Flores, C.O., Valverde, S., Sullivan, M.B., Hochberg, M.E., 2013. Phage-bacteria infection networks. Trends Microbiol. 21, 82-91.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., Rohwer, F., 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4, e7370.

Willner, D., Furlan, M., Schmieder, R., Grasis, J.A., Pride, D.T., Relman, D.A., Angly, F.E., McDole, T., Mariella, R.P., Jr, Rohwer, F., Haynes, M., 2011. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc. Natl. Acad. Sci. U. S. A. 108 Suppl 1, 4547-4553.

Winter, C., Bouvier, T., Weinbauer, M.G., Thingstad, T.F., 2010. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the «killing the winner» hypothesis revisited. Microbiol. Mol. Biol. Rev. 74, 42-57.

Winter, G., Hart, R.A., Charlesworth, R.P.G., Sharpley, C.F., 2018. Gut microbiome and depression: what we know and what we need to know. Rev. Neurosci. 29, 629-643.

Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221-271.

Wright, A., Hawkins, C.H., Änggård, E.E., Harper, D.R., 2009. A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant Pseudomonas aeruginosa; a preliminary report of efficacy. Clin. Otolaryngol. 34, 349-357.

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., Zhou, K., 2018. Metagenomics biomarkers selected for prediction of three different diseases in chinese population. Biomed Res. Int. 2018, 2936257.

Wu, L., Liu, X., Schadt, C.W., Zhou, J., 2006. Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. Appl. Environ. Microbiol. 72, 4931-4941.

Wu, L., Wen, C., Qin, Y., Yin, H., Tu, Q., Van Nostrand, J.D., Yuan, T., Yuan, M., Deng, Y., Zhou, J., 2015. Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. BMC Microbiol. 15, 125.

Wylie, K.M., Mihindukulasuriya, K.A., Zhou, Y., Sodergren, E., Storch, G.A., Weinstock, G.M., 2014. Metagenomic analysis of double-stranded DNA viruses in healthy adults. BMC Biol. 12, 71.

Ximenez, C., Torres, J., 2017. Development of microbiota in infants and its role in maturation of gut mucosa and immune system. Arch. Med. Res. 48, 666-680.

Yan, X., Yang, M., Liu, J., Gao, R., Hu, J., Li, J., Zhang, L., Shi, Y., Guo, H., Cheng, J., Razi, M., Pang, S., Yu, X., Hu, S., 2015. Discovery and validation of potential bacterial biomarkers for lung cancer. Am. J. Cancer Res. 5, 3111-3122.

Yilmaz, S., Allgaier, M., Hugenholtz, P., 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat. Methods 7, 943-944.

Young, R., 1992. Bacteriophage lysis: mechanism and regulation. Microbiol. Rev. 56, 430-481.

Yuan, Y., Gao, M., 2017. Jumbo bacteriophages: an overview. Front. Microbiol. 8, 403.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.-Q., Wei, C.L., Soh, S.W.L., Hibberd, M.L., Liu, E.T., Rohwer, F., Ruan, Y., 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 4, e3.

Zhao, H., Chu, M., Huang, Z., Yang, X., Ran, S., Hu, B., Zhang, C., Liang, J., 2017. Variations in oral microbiota associated with oral cancer. Sci. Rep. 7, 11773.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., Wishart, D.S., 2011. PHAST: a fast phage search tool. Nucleic Acids Res. 39, 347-352.

Zhu, C., Li, F., Wong, M.C.M., Feng, X.-P., Lu, H.-X., Xu, W., 2015. Association between Herpesviruses and Chronic Periodontitis: A Meta-Analysis Based on Case-Control Studies. PLoS One 10, e0144319.

Ziouani, S., Khelil, N.K., Benyelles, I., 2015. Oral microflora of supragingival and subgingival biofilms in Algerian healthy adults. African Journal of Microbiology Research 9, 1548-1557.

Zong, C., Lu, S., Chapman, A.R., Xie, X.S., 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science 338, 1622-1626.

Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F., Tang, W., Ching, J.Y.L., Zhao, R., Chan, P.K.S., Sung, J.J.Y., Yu, J., Chan, F.K.L., Cao, Q., Sheng, J.-Q., Ng, S.C., 2019. Gut mucosal virome alterations in ulcerative colitis. Gut gutjnl-2018-318131.

Zuo, T., Wong, S.H., Lam, K., Lui, R., Cheung, K., Tang, W., Ching, J.Y.L., Chan, P.K.S., Chan, M.C.W., Wu, J.C.Y., Chan, F.K.L., Yu, J., Sung, J.J.Y., Ng, S.C., 2018. Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome. Gut 67, 634-643.

ANEXOS

1. Oligonucleótidos

Oligonucleótido	Aplicación	Secuencia
M13-F HindIII	Estándar	AAGCTTCACTCCGCTGAAACTGTTGA
M13-R EcoRV	Estándar	GATATCGTCAGACGATTGGCCTTGAT
WR-F	*PCR* cuantitativa	AATAGACGGCGACAGGTTTC
WR-R	*PCR* cuantitativa	ACGGTGCACAAATACCTACG
Lambda-F	*PCR* cuantitativa	AGTACCCAATGATCCCATGC
Lambda-R	*PCR* cuantitativa	TCAGCCAAACGTCTCTTCAG
AdenoV-F	*PCR* cuantitativa	CGGATGGAACCATTATACCG
AdenoV-R	*PCR* cuantitativa	CTGGGCGAAGATATTTCTGG
Φ29-F	*PCR* cuantitativa	TGCGAACCCTAGAAGAAAGC
Φ29-R	*PCR* cuantitativa	ATCAGTTCATCTGCCGCATC
M13-F	*PCR* cuantitativa	TGAGGGTTGTCTGTGGAATG
M13-R	*PCR* cuantitativa	TAGCAAGCCCAATAGGAACC
MVMp-F	*PCR* cuantitativa	AGGGTTTAAGGGATGGTTGG
MVMp-R	*PCR* cuantitativa	TTGGTTGGTTCTCCTTGGTC
PCV2a-F	*PCR* cuantitativa	AATGAGGAAGGACGAACACC
PCV2a-R	*PCR* cuantitativa	CAGTTCCTTTGGCTTTCTCG
FR20RV	*SISPA*	GCCGGAGCTCTGCAGATATC
FR26RV-12N	*SISPA*	GCCGGAGCTCTGCAGATATCNNNNNNNNNNNN
1N-FR20RV	*SISPA*	NGCCGGAGCTCTGCAGATATC
2N-FR20RV	*SISPA*	NNGCCGGAGCTCTGCAGATATC
3N-FR20RV	*SISPA*	NNNGCCGGAGCTCTGCAGATATC
4N-FR20RV	*SISPA*	NNNNGCCGGAGCTCTGCAGATATC
Primer_K	*SISPA*	GACCATCTAGCGACCTCCAC
Primer_K-8N	*SISPA*	GACCATCTAGCGACCTCCACNNNNNNNN
K-12N	*SISPA*	GACCATCTAGCGACCTCCACNNNNNNNNNNNN
1N-K	*SISPA*	NGACCATCTAGCGACCTCCAC
2N-K	*SISPA*	NNGACCATCTAGCGACCTCCAC
3N-K	*SISPA*	NNNGACCATCTAGCGACCTCCAC
4N-K	*SISPA*	NNNNGACCATCTAGCGACCTCCAC
454-A	*SISPA*	ATCGTCGTCGTAGGCTGCT
454-A-12N	*SISPA*	ATCGTCGTCGTAGGCTGCTNNNNNNNNNNNN
1N-454-A	*SISPA*	NATCGTCGTCGTAGGCTGCT
2N-454-A	*SISPA*	NNATCGTCGTCGTAGGCTGCT
3N-454-A	*SISPA*	NNNATCGTCGTCGTAGGCTGCT
4N-454-A	*SISPA*	NNNNATCGTCGTCGTAGGCTGCT
27-F	Contaminación 16S	AGAGTTTGATCMTGGCTCAG
1492-R	Contaminación 16S	TACGGYTACCTTGTTACGACTT
CS1FsL-Pro341f	Amplificación 16S	ACACTGACGACATGGTTCTACA##CCTACGGGNBGCASCAG
CS2FsL-Pro805R	Amplificación 16S	TACGGTAGCAGAGACTTGGTCT##GACTACNVGGGTATCTAATCC
P5Cs1	Amplificación 16S	AATGATACGGCGACCACCGAGATCTACACTGACGACATGGTTCTACA
P7-bc1-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTTACCGACGTACGGTAGCAGAGACTTGGTCT
P7-bc2-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATATTGGACACTACGGTAGCAGAGACTTGGTCT

P7-bc3-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTCGCATGGATACGGTAGCAGAGACTTGGTCT
P7-bc5-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATAGCTTCGACTACGGTAGCAGAGACTTGGTCT
P7-bc6-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGTCAGCCGTTACGGTAGCAGAGACTTGGTCT
P7-bc7-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTCCAGATAGTACGGTAGCAGAGACTTGGTCT
P7-bc9-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGCTCACAATTACGGTAGCAGAGACTTGGTCT
P7-bc11-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCTTAGAACGTACGGTAGCAGAGACTTGGTCT
P7-bc12-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCGGTTCACATACGGTAGCAGAGACTTGGTCT
P7-bc13-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCGATAGGCCTACGGTAGCAGAGACTTGGTCT
P7-bc14-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGCTATATCCTACGGTAGCAGAGACTTGGTCT
P7-bc15-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGTCTTCAGCTACGGTAGCAGAGACTTGGTCT
P7-bc16-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTAGACACCGTACGGTAGCAGAGACTTGGTCT
P7-bc17-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTCAGCTGACTACGGTAGCAGAGACTTGGTCT
P7-bc18-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTAAGTCGGCTACGGTAGCAGAGACTTGGTCT
P7-bc19-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGCTCCTTAGTACGGTAGCAGAGACTTGGTCT
P7-bc20-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATATGGCCTGATACGGTAGCAGAGACTTGGTCT
P7-bc21-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTTGCAAGTATACGGTAGCAGAGACTTGGTCT
P7-bc22-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCCTAGTAAGTACGGTAGCAGAGACTTGGTCT
P7-bc24-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTATGAACGTTACGGTAGCAGAGACTTGGTCT
P7-bc26-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCACGATGGTTACGGTAGCAGAGACTTGGTCT
P7-bc27-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATACGTGCCTTTACGGTAGCAGAGACTTGGTCT
P7-bc28-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTGAACTAGCTACGGTAGCAGAGACTTGGTCT
P7-bc30-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATTAATCGGTGTACGGTAGCAGAGACTTGGTCT
P7-bc31-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGCGTCCATGTACGGTAGCAGAGACTTGGTCT
P7-bc32-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCGTAAGATGTACGGTAGCAGAGACTTGGTCT
P7-bc33-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCTGTTACAGTACGGTAGCAGAGACTTGGTCT
P7-bc35-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGTAACGGCTTACGGTAGCAGAGACTTGGTCT
P7-bc36-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCCATGCTTATACGGTAGCAGAGACTTGGTCT
P7-bc37-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATGTACGCACATACGGTAGCAGAGACTTGGTCT
P7-bc56-Cs2	Amplificación 16S	CAAGCAGAAGACGGCATACGAGATCTCTGAGGTTACGGTAGCAGAGACTTGGTCT
ALH7-F	Papilomavirus	TTGACTGGAGAACTAACTATTCT
ALH7-R	Papilomavirus	TAACTGCTGCAAAGTCCGAACG
ALV10-F	Papilomavirus	GTTAAGCATTTGCAGCAACTATT
ALV10-R	Papilomavirus	GCCTATTTCAGTAGCAACTAAGC
T7	pGEM-t Easy	TAATACGACTCACTATAGGG
SP6	pGEM-t Easy	ATTTAGGTGACACTATAG
M13-F	pSpark V	CCCAGTCACGACGTTGTAAAACG
M13-R	pSpark V	AGCGGATAACAATTTCACACAGG

2. Scripts

Script I - Trifonov_Complex.pl

Este script evalúa la complejidad de las secuencias a través del modelo de complejidad lingüística de Trifonov

Usage

1- Modificar la ruta del archivo (*path*) y el nombre de archivo de entrada y salida (*file_contigs* y *file_Salida* respectivamente)
2- Indicar por argumentos el tamaño de ventana de nucleótidos a analizar (primer parámetro) y el tamaño del paso a dar (segundo parámetro).

Ejemplo: "perl Trifonov_Complex.pl 50 20"

Enlace: **https://git.io/fj3Bc**

Script II - Busqueda_Primers_Mapeo_SISPA.pl

Este script busca regiones en cada *contig* con similitud de secuencia con los primers *SISPA* utilizados (En nuestro caso FRV20, K y 454).

Usage

1- Modificar la ruta del archivo (*path*) y el nombre de archivo de entrada y salida (*file_contigs* y *file_Salida* respectivamente)
2- En caso de querer probar otros primers en lugar de los que vienen por defecto, modificar la secuencia dentro del script
3- Para ejecutar el script, utilizar el argumento "V"

Ejemplo: "perl Busqueda_Primers_Mapeo_SISPA.pl V

Enlace: **https://git.io/fj3B8**

Script III - ParseadorBlast.pl

La principal funcionalidad de este script consiste en analizar los resultados de *BLAST* de las pautas abiertas de lectura de todos los *contigs* lanzados por *Blastx* contras las bases de datos *nr* y *PHAST*. Categoriza a los *contigs* en función de los cinco primeros hits de cada pauta abierta de lectura y lo asigna a *Virus*, *Bacteria*, *No hit* y *Sin clasificar*.

Usage

* Modificar el /*path*/ y poner la ruta de los archivos .*blast* en formato de salida tipo 0.
* Son necesarios 3 archivos para la taxonomía de virus y su host. Estos archivos deben estar en la misma ruta que el *script*:

    -Resultados_GeneBank20.txt

    -Taxa_ProtVir_2016.txt

    -ICTV-Master-Species-List-2014.csv

* Los archivos con las pautas abiertas de lectura extraídos mediante *Prodigal* se colocarán en una carpeta a un nivel superior al de los archivos .*blast* (../), conservando el mismo código de nombre para el metagenoma que en el archivo .*blast*.
* Los archivos con los *contigs* de CLC se colocarán en una carpeta a dos niveles superiores de los archivos blast (../../), conservando el mismo código de nombre para el metagenoma que en el archivo blast.

Ejemplo:

    Blast: ALH6_CLC_assembly-nr.blast

    FileORF: ../ALH6_CLC_assembly.fa

    FileContigs: ../../ALH6_CLC_assembly.fa

1- Si no se quiere realizar una criba por longitud y cobertura de los *contigs*, indicarlo con el parámetro "N".

    Ejemplo: "ParseadorBlast.pl N"

2- Si se quiere realizar una criba por longitud y cobertura de los *contigs* con un doble criterio, indicar los cuatro parámetros del modo: longitud 1, cobertura 1, longitud 2, cobertura 2.

Ejemplo: "ParseadorBlast.pl 3000 15 10000 4"

3- Los archivos de salida se generarán dentro de la carpeta llamada *Clasificacion_contigs*.

Enlace: **https://git.io/fj3BB**

Script IV - Famio_Breadth.pl

Este script calcula la distancia *Dice* entre dos genomas, teniendo en cuenta el *breadth coverage* del más corto en el numerador, y el *breadth coverage* del más corto contra sí mismo en cada comparación. Identidad mínima de 35% y longitud mínima de 45 aminoácidos por defecto.

Usage

1- Modificar la ruta del archivo (path) para indicar la carpeta donde se alojan las secuencias fasta que se van a analizar
2- Para ejecutar el script, utilizar el argumento "V"
3- Los resultados se generan dentro del archivo de salida *Resultados_score.txt*

Ejemplo: "perl Famio_Breadth.pl"

Enlace: **https://git.io/fj3B0**

Script V - Calculo_frecuencias_nucleotidicas.R

Mediante este script realizamos el cálculo de frecuencia nucleotídica (de tetranucleótidos por defecto) de genomas o *contigs* desde archivos en formato fasta. Posteriormente realiza un estudio de correlación de Pearson entre todas las frecuencias calculadas.

Usage

1- Ejecutar y seleccionar aquella carpeta donde se encuentren los archivos fasta almacenados.
2- Los resultados se almacenan dentro del objeto *resultados_cor*.

Enlace: **https://git.io/fj3Bz**