UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

CENTRO DE INVESTIGACIÓN EN CIENCIAS

“El pangenoma viral: bases de datos de referencia mejoradas para genómica viral”

TESIS PROFESIONAL PARA OBTENER EL GRADO DE:

DOCTORADO EN CIENCIAS

PRESENTA:
OSCAR ALEJANDRO USCANGA JUNCO

DIRECTORA:
DRA. BLANCA ITZELT TABOADA RAMÍREZ

CODIRECTORA:
DRA. LORENA DÍAZ GONZÁLEZ

CUERNAVACA, MORELOS

MAYO, 2026

Instituto de Investigación en Ciencias Básicas y Aplicadas

Agradecimientos

A mis padres Obdulia e Ismael por la gentileza para llevar a cabo la crianza que me hizo ser quien soy hoy.

A Lily por tu autenticidad inspiradora, por tu apoyo infinito y tu amor sincero.

A Guillermo por tu amistad y tu hambre de conocimiento.

A Javier, Kevin, Kevin y Omar por su amistad.

A la Dra. Blanca Itzelt Taboada Ramírez y a la Dra. Lorena Díaz González por sus enseñanzas constantes y su mentoría de primer nivel. Gracias por permitirme ser parte de su equipo y compartir conmigo sus toneladas de experiencia y calidez humana.

Al Dr. Rodrigo García López quien además de brindarme mentoría sobre mi proyecto de doctorado como parte de mi comité tutorial, se convirtió en un gran amigo. Gracias por las increíbles conversaciones.

A Lucas por tu amistad y felicidad contagiosa.

A mis amigas Alida y Edna por el acompañamiento emocional e intelectual a lo largo de doctorado.

Al Posgrado en Ciencias del Instituto de Investigación en Ciencias Básicas y Aplicadas (IICBA) de la Universidad Autónoma del Estado de Morelos (UAEM).

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por la beca otorgada que me permitió llevar a cabo mis estudios de Doctorado a tiempo completo.

A mi comité tutorial: Dr. José Alberto Hernández Aguilar y Dr. Armando Hernández Mendoza.

A mi comité evaluador conformador por: Dr. Armando Hernández Mendoza, Dr. Jorge Hermosillo Valadez, Dr. Ramón Antonio González García- Conde, Dra. Blanca Itzelt Taboada Ramírez, Dra. Sonia Dávila Ramos, Dr. José Alberto Hernández Aguilar y el Dr. Rodrigo García López.

A María Cristina Aranda por su impecable forma de llevar la organización del Posgrado en Ciencias.

A todos y cada uno de los integrantes del "Lab 5" del Instituto de Biotecnología de la Universidad Nacional Autónoma de México.

A la Unidad Universitaria de Secuenciación Masiva y Bioinformática (UUSMB) por darme acceso al clúster Teopanzolco y a Jérôme Verleyen por su apoyo técnico de calidad.

A la Dirección de Cómputo y de Tecnologías de Información y Comunicación (DGTIC) de la Universidad Nacional Autónoma de México por los recursos de supercómputo brindados a través del proyecto LANCADUNAM-DGTIC-350 en la supercomputadora MIZTLI.

A la Universidad Autónoma de la Ciudad de México a través del proyecto SECTEI/138/2024 “VIGILANCIA GENÓMICA DE VIRUS RESPIRATORIOS EN LA CIUDAD DE MÉXICO” otorgado a la Dra. Selene Zárate y la Dra. Blanca Taboada por el apoyo económico para poder finalizar mis semestres adicionales del Doctorado.

“*Boy oh boy… the price of freedom is steep. Embrace your dreams and whatever happens, protect your honor… as SOLDIER!*”

— Zack Fair, Crisis Core: Final Fantasy VII, 2007.

Dedicatorias

A Lily.

A mi madre Obdulia y a mi padre Ismael.

A Guillermo.

A Carlos.

A Melissa.

A Limón, Cebolla y Bubumao.

Abstract

Viruses are organized into 12 taxonomic ranks: realm, kingdom, phylum, subphylum, class, order, suborder, family, subfamily, genus, subgenus, and species. Species-level sequences belonging to the same genus contain redundant information. When a metagenomic read matches this redundant region, procedures are necessary to achieve accuracy in taxonomic annotation, such as identifying the last common ancestor (LCA) or using probabilistic methods to identify the correct species or genus by determining abundances.

This doctoral thesis presents the development of a novel methodological approach for generating reference databases with reduced redundancy based on the organization of viral sequences at the genus and species taxonomic levels. These structures are called "Viral Pangenomes" and are intended to be used for viral identification. The procedure for generating pangenomes involves reducing redundancy among conserved regions and determining those unique to viral species, as well as identifying redundant nucleotide regions shared among species belonging to the same genus. The goal of these structures is to simplify the taxonomic identification process of short reads by offering a more compact and specific alternative for use as a reference database for taxonomic annotation without compromising their phylogenetic diversity.

Viral pangenomes underwent to a 99.51% median reduction in redundancy and up to 100% average recovery of sequences associated with viral species, up to 94.1% compression ratio was observed for genus with high interspecies identity. This pangenome generation procedure can also be used to identify unique sequences at subspecies taxonomic levels, such as Influenza A’s H and N subtypes.

Resumen

Los virus se organizan en 12 rangos taxonómicos; reino mayor, reino, filo, subfilo, clase, orden, suborden, familia, subfamilia, género, subgénero y especie. Las secuencias a nivel especie que pertenecen al mismo género contienen información redundante, pues cuando un *read* tiene una coincidencia con esta zona redundante, es necesario llevar a cabo procedimientos para obtener precisión en la anotación taxonómica como la identificación de un ancestro común o LCA o un procedimiento de asignación probabilística para determinar la especie o género correctos mediante la determinación de abundancias.

En esta tesis doctoral se presenta el desarrollo de una metodología novedosa para la generación de bases de datos de referencia con redundancia reducida basada en la organización de secuencias virales en los niveles taxonómicos género y especie, a estas estructuras se les denomina "Pangenomas Virales" y tienen su uso proyectado para metagenómica viral. El procedimiento para la generación de pangenomas consiste en la detección de redundancia entre regiones de nucleótidos y la determinación de regiones únicas para cada especie, así como identificación de regiones de nucleótidos redundantes que se comparten entre especies que pertenecen al mismo género. El objetivo de estas estructuras es simplificar el procedimiento de la identificación taxonómica de lecturas metagenómicas ofreciendo una alternativa más compacta y específica para ser usada como base de datos de referencia con el objetivo de hacer anotación taxonómica sin comprometer la diversidad filogenética.

Los pangenomas virales muestran hasta un 99.51% de reducción de redundancia y un hasta 100% de recuperación media hacia secuencias asociadas a las especies virales también se encontró hasta un 94.1% de índice de compresión para géneros con alta identidad interespecie. Este procedimiento de generación de pangenomas puede ser utilizado además para identificación de secuencias únicas a niveles taxonómicos inferiores a especie, como los subtipos H y N de Influenza A.

Índice

Índice de figuras

Índice de tablas

Capítulo 1 Introducción

La metagenómica es una subdisciplina de la genética que permite estudiar el material genético de los organismos presentes en un microbioma, como bacterias, virus, hongos y otros microorganismos, sin necesidad de cultivarlos previamente. Mediante plataformas de secuenciación masiva de segunda generación, es posible obtener miles o millones de fragmentos cortos (75 a 150 pares de bases) de ADN, denominados *reads* o lecturas, que representan el material genético de los organismos presentes en la muestra (Zhang et al., 2021). Para interpretar esta información, se utilizan bases de datos de referencia, las cuales contienen secuencias homologas con una anotación taxonómica previamente establecida. De esta manera, los *reads* identificados como homólogos de las secuencias de referencia pueden ser asignados a distintos grupos taxonómicos. A partir del conjunto de estas asignaciones, es posible inferir la composición ecológica de la muestra metagenómica. Por ello, la calidad, representatividad y actualización de las bases de datos de referencia son elementos críticos para lograr una clasificación metagenómica precisa (Martí et al. 2025).

Los clasificadores taxonómicos son herramientas computacionales que permiten asignar los *reads* metagenómicos a distintos grupos taxonómicos. Para ello, utilizan algoritmos de comparación de secuencias que evalúan la similitud entre los *reads* y las secuencias contenidas en bases de datos de referencia. Cuando esta comparación se realiza mediante la búsqueda de la similitud máxima entre un *read* y una región anotada de una secuencia de referencia, el procedimiento se conoce como alineamiento parcial o global, y las herramientas que lo llevan a cabo se denominan alineadores. Sin embargo, la clasificación a niveles taxonómicos específicos, como género o especie, puede presentar baja sensibilidad, ya que estas herramientas no siempre logran identificar todos los taxones anotados en la base de datos de referencia (Cadenas-Castrejón et al., 2023). En este contexto, la diversidad viral presente en la muestra y la longitud de los *reads* metagenómicos son factores relevantes, debido a que pueden impactar directamente en el desempeño del clasificador taxonómico.

De acuerdo con la ICTV (*International Committee on Taxonomy of Viruses*), autoridad encargada de definir las categorías virales, los virus están organizados en 12 rangos taxonómicos: reino mayor, reino, filo, subfilo, clase, orden, suborden, familia, subfamilia, género, subgénero y especie (Simmonds et al., 2024). En los niveles taxonómicos más específicos, como género o especie, las secuencias de referencia pueden compartir regiones altamente conservadas o redundantes entre taxones filogenéticamente cercanos. Por ello, cuando un *read* se alinea con una región compartida por varias secuencias de referencia, su asignación taxonómica puede ser ambigua apuntando a varios taxones con la misma identidad entre secuencias. En estos casos es necesario llevar a cabo procedimientos de

clasificación inespecífica, como la identificación de un ancestro común o LCA por sus siglas en inglés, *Last Common Ancestor* (Wang et al., 2022) o un procedimiento de asignación probabilística para determinar la especie o género más probable (Zheng, Shaw, y Yu 2024).

Esta tesis doctoral se centra en el desarrollo de una nueva metodología para generar bases de datos de referencia virales con redundancia reducida y por lo tanto altamente específicas, denominadas "pangenomas virales", cuyo uso está orientado a la metagenómica viral. Estas bases de datos se construyen mediante un proceso de detección de redundancia que permite identificar regiones genómicas únicas para cada especie, así como regiones compartidas entre especies pertenecientes al mismo género. De esta forma, los pangenomas virales buscan simplificar el proceso de identificación taxonómica, al organizar la información genómica en una estructura que permite distinguir entre secuencias específicas de especie y secuencias conservadas a nivel de género, reduciendo así la ambigüedad durante la clasificación de *reads* metagenómicos.

1.1 Contribuciones de esta tesis

Las principales contribuciones de esta tesis son las siguientes. Primero, se desarrolló una metodología para la generación de pangenomas virales a partir de las secuencias disponibles en la base de datos NCBI Virus, con el objetivo de reducir la redundancia de las bases de referencia sin perder la anotación taxonómica asociada. Segundo, se generaron pangenomas a nivel de especie, los cuales permiten representar la información genómica de cada especie viral mediante un conjunto reducido de secuencias no redundantes. Esta reducción puede disminuir el tamaño de los índices utilizados por alineadores como Bowtie 2 y BWA, facilitando su uso en pipelines metagenómicos.

Tercero, se generaron pangenomas a nivel de género que distinguen entre regiones específicas de especie y regiones conservadas entre especies del mismo género. Esta organización permite identificar explícitamente cuándo un *read* corresponde a una región compartida dentro de un género, lo que puede reducir la ambigüedad taxonómica y disminuir la necesidad de procedimientos posteriores como análisis de ancestro común o estimaciones de abundancia para resolver asignaciones ambiguas. Finalmente, se demostró que la metodología puede adaptarse a niveles taxonómicos o biológicos por debajo de especie, como subtipos virales. En particular, se aplicó al caso de Influenza A para identificar secuencias únicas asociadas a subtipos, mostrando que este enfoque puede extenderse a otros virus con clasificaciones intraespecíficas, como serotipos, subtipos o linajes y subespecies.

1.2 Bases de datos de referencia de virus

Actualmente, el subconjunto viral de la base de datos de nucleótidos de NCBI, denominada "NT" contiene 14,565,917 secuencias de nucleótidos y ocupa aproximadamente 275 GB. Para utilizar esta base de datos con herramientas de alineamiento como Bowtie2 (Langmead & Salzberg, 2012) o BWA (Li & Durbin, 2009), es necesario construir un índice de rápida navegación, también conocido como índice FM. Sin embargo, este índice puede ser considerablemente más grande que la base de datos original, lo que vuelve computacionalmente demandante su procesamiento.

Otra base de datos ampliamente utilizada es RefSeq, la cual corresponde a un subconjunto curado de las secuencias genómicas de referencia disponibles en NCBI. RefSeq es producto de un esfuerzo de curaduría por generar una base de datos de referencia de organismos, ya que generalmente contiene una secuencia genómica representativa por organismo, sepa, tipo, genoma o variante de importancia. En el caso de los virus, esta base de datos se utiliza principalmente como un estándar de referencia para la comparación de especies. Precisamente, al incluir un número limitado de secuencias representativas por especie viral, RefSeq no captura toda la variabilidad genética existente dentro de cada especie. De acuerdo con Goldfarb et al. (2025), RefSeq contiene una cantidad limitada de referencias para más de 14,761virus (junio 2026). Esta cifra contrasta con las 66,660 especies virales incluidas en los nodos taxonómicos correspondientes a las secuencias de la INSD *(International Nucleotide Sequence Database Collaboration)*, lo que evidencia que no existe una secuencia de referencia disponible para todas las especies virales anotadas. Por lo tanto, el uso exclusivo de RefSeq como base de datos de referencia en metagenómica viral puede ser limitado cuando se busca considerar tanto la diversidad genética intraespecífica como la mayor cantidad posible de especies virales anotadas.

Una de las dificultades asociadas con el uso de bases de datos extensas es la presencia de secuencias altamente similares o redundantes. En (Kim et al., 2016), observaron que las especies bacterianas más estudiadas suelen tener numerosos genomas muy similares entre sí. Para reducir esta redundancia, los autores desarrollaron un procedimiento que elimina información repetida entre genomas de una misma especie, con el objetivo de disminuir el tamaño del índice FM. Este procedimiento inicia con las dos secuencias que comparten el mayor número de *k*-meros dentro de un grupo. Posteriormente, ambas secuencias, denominadas G1 y G2, son comparadas con Nucmer (Kurtz et al., 2004), herramienta que identifica regiones altamente similares entre los genomas. Para generar una versión reducida, los dos genomas se combinan, pero se descartan las regiones de G2 que presentan más de 99% de similitud con G1. Este proceso se repite de manera iterativa, comparando cada nuevo genoma contra la secuencia combinada previamente reducida. Mediante esta estrategia

basada en la identificación iterativa de regiones altamente similares, la versión reducida de los la base de datos de *Salmonella enterica* pasó de 655 Mbp a 107 Mbp.

Centrifugue (Kim *et al*., 2016) es una herramienta de clasificación taxonómica basada en índices FM, diseñada para reducir el tamaño de las bases de datos mediante la eliminación de regiones redundantes entre genomas altamente similares. Posteriormente, Centrifugue (Song & Langmead, 2024), retomó el uso de índices FM e incorporó procedimientos adicionales de compresión para reducir aún más su tamaño. Entre ellos se incluye la compresión Run-Block, que permite representar de forma más compacta regiones de baja complejidad en los genomas indexados. Además, Centrifugue realiza por defecto un procedimiento de asignación mediante ancestro común más bajo, o LCA, de manera similar a Kraken 2 (Wood, Lu y Langmead 2019). Sin embargo, tanto Centrifugue como Kraken 2 están diseñados principalmente para reducir la complejidad espacial asociada con la construcción y el almacenamiento de índices FM. Por ello, aunque estos enfoques disminuyen la redundancia desde una perspectiva computacional, la reducción no necesariamente está orientada a preservar una estructura taxonómica explícita, sino que depende de procedimientos posteriores, como LCA, para resolver asignaciones ambiguas.

La primera versión de Centrifugue permanece como la más utilizada; sin embargo, su base de datos de NT no se ha actualizado desde 2018. En el trabajo de Martí et al., (2025), se crearon nuevas bases de datos compatibles con Centrifugue en forma de capturas temporales para brindar estabilidad entre las versiones de la base de datos y el árbol taxonómico de NCBI. Este trabajo es relevante porque resalta la necesidad de considerar las bases de datos de referencia como entidades dinámicas, sujetas a procesos continuos de control de calidad, actualización y versionado. En este sentido, la curación de una base de datos de referencia debe entenderse como un procedimiento sistemático y reproducible, capaz de incorporar actualizaciones constantes sin comprometer la trazabilidad de los análisis metagenómicos.

Kraken (Wood y Salzberg 2014) es otra de las herramientas más utilizadas para la anotación taxonómica de datos metagenómicos. A diferencia de los métodos basados en alineamiento completo, Kraken utiliza coincidencias exactas de *k*-meros. Su base de datos de referencia contiene *k*-meros asociados con anotaciones taxonómicas obtenidas a partir del ancestro común más bajo de los organismos cuyos genomas contienen dichos *k*-meros. Para construir esta base de datos, se calculan todos los *k*-meros de una longitud definida, comúnmente 31 nucleótidos, aunque este parámetro puede modificarse. La asignación taxonómica de cada *k*-mer se realiza con base en los nodos taxonómicos de NCBI. Este enfoque permite una clasificación rápida y eficiente; sin embargo, puede verse afectado cuando los reads provienen de regiones conservadas compartidas entre taxones cercanamente relacionados, ya

que en estos casos la asignación puede resolverse únicamente a un nivel taxonómico superior mediante LCA.

De manera similar, CLARK (Ounit et al., 2015) también utiliza *k*-meros para la clasificación taxonómica, aunque se diferencia de Kraken en que considera únicamente *k*-meros discriminatorios para determinar la clasificación de los *reads*. Esta estrategia puede aumentar la especificidad de la asignación taxonómica, pero también puede reducir la sensibilidad cuando los *reads* no contienen *k*-meros suficientemente discriminatorios.

Kraken 2 (Wood, Lu, y Langmead 2019) retoma el enfoque basado en *k*-meros, pero incorpora el uso de minimizadores para reducir el tamaño de la base de datos. Un minimizador es una subsecuencia de longitud *l* contenida dentro de un *k*-mer de longitud *k*, donde *l* es menor que *k*. Para cada *k*-mer, el minimizador corresponde al *l*-mer lexicográficamente menor, considerando también su representación en la cadena complementaria inversa. Debido a que muchos *k*-meros almacenados en la base de datos de Kraken son altamente similares, almacenar una sola vez los minimizadores compartidos permite reducir de manera considerable el tamaño de la base de datos. En Kraken 2, esta estrategia disminuye el tamaño de almacenamiento aproximadamente en 85%, con una reducción limitada en la precisión reportada, cercana a 1% (Wood y Salzberg 2014). No obstante, el uso de minimizadores implica una representación más compacta de la información contenida en los *k*-meros, por lo que algunas señales taxonómicas específicas pueden perderse o resolverse con menor precisión, particularmente en secuencias cortas, regiones compartidas entre taxones cercanos o grupos con alta similitud genética.

Estos enfoques muestran que la reducción del tamaño y la redundancia de las bases de datos de referencia es un problema central en metagenómica. Sin embargo, la mayoría de las estrategias existentes reducen la redundancia con fines principalmente computacionales, ya sea para disminuir el tamaño de los índices, reducir los requerimientos de almacenamiento o acelerar la clasificación. En contraste, la metodología propuesta en esta tesis busca generar bases de datos virales con redundancia reducida que conserven una organización taxonómica explícita, distinguiendo entre regiones específicas de especie y regiones compartidas entre especies del mismo género. De esta manera, los pangenomas virales propuestos no solo buscan optimizar el uso computacional de las bases de datos, sino también facilitar una clasificación taxonómica más interpretable en el contexto de la metagenómica viral.

1.3 Planteamiento del problema y justificación

Las secuencias que pertenecen a un mismo grupo taxonómico específico, como género o especie suelen tener regiones similares entre sí cuando comparten un origen evolutivo (Wood, Lu y Langmead 2019). Esta similitud puede interpretarse como redundancia en una cuando existen regiones de nucleótidos idénticas o altamente conservadas que no aportan nueva información en una base de datos. En el contexto de la clasificación metagenómica viral, esta redundancia genera dos problemas principales.

A. El primero ocurre cuando existen regiones conservadas entre genomas de especies diferentes que pertenecen al mismo género, o entre variantes y subespecies de una misma especie. Este fenómeno es más frecuente en niveles taxonómicos específicos, como género y especie, debido a la presencia de regiones genómicas conservadas resultantes de la menor divergencia genética. En la Figura 1.1 se muestra un ejemplo del genoma de MERS-CoV y las 18 primeras coincidencias de BLAST (Altschul et al., 1990). En dicha comparación se observan regiones conservadas, asociadas con las proteínas NSP7–NSP14, que están presentes en todos los genomas analizados.

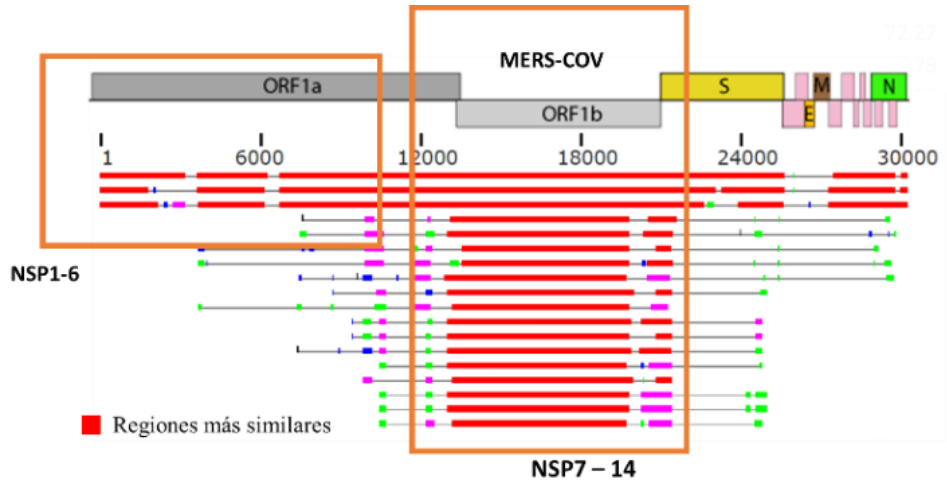

Figura 1.1 Comparación de 19 genomas del género Betacoronavirus*. Se observan regiones conservadas con más del 95% de identidad entre genomas del mismo género, particularmente en las regiones asociadas con las proteínas NSP7–NSP14.*

B. El segundo problema se relaciona con la divergencia genética interespecie. Debido a la alta tasa de mutación de algunas especies virales y a que los clasificadores taxonómicos no siempre responden a una distancia filogenética fija, pueden existir genomas pertenecientes a la misma especie con mayor divergencia con respecto a una única secuencia representativa. Por ello, cuando se utiliza una sola referencia por especie,

como ocurre en RefSeq, no es posible representar de manera completa la variabilidad genética de una especie viral. Este fenómeno puede observarse en especies como *Protoambidensovirus lepidopteran1* y *Tetraparvovirus primate1*, cuyas similitudes interespecie varían entre 84% y 100%, y entre 83% y 100%, respectivamente. Asimismo, en la Figura 1.2 se muestra una comparación genomas de la especie *Erythroparvovirus primate1* contra su genoma de referencia, en la cual se observa que algunos genomas presentan regiones cercanas a los extremos 5' y 3' con baja similitud respecto a la secuencia representativa.

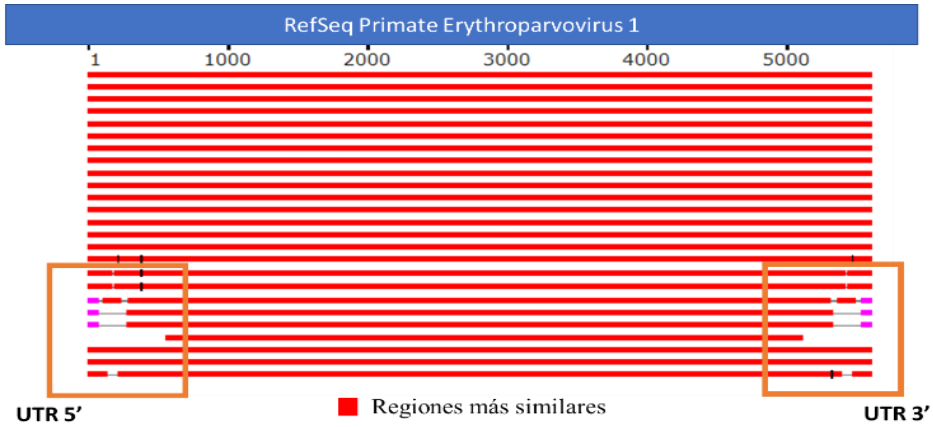

Figura 1.2. Comparación de varios genomas de la especie Erythroparvovirus primate1 contra el genoma de referencia de RefSeq. *Se observan regiones con baja similitud respecto al genoma de referencia, principalmente hacia los extremos 5' y 3', lo que evidencia variabilidad genética intraespecífica o incluso errores de secuenciacion.*

En el primer caso, cuando un *read* se alinea con una región conservada a nivel género, redundante o compartida por varias especies, la asignación taxonómica puede ser ambigua y requiere procedimientos adicionales, como la identificación del ancestro común más bajo, para establecer el nivel taxonómico más específico que puede asignarse con confianza. En el segundo caso, cuando se utiliza una única secuencia representativa, los *reads* que no comparten homologías provenientes de regiones variables que no están representadas en dicha referencia pueden no ser anotados correctamente o permanecer sin clasificación.

La creación de una estructura de información capaz de integrar estos dos tipos de regiones permitiría abordar dos problemas centrales de las bases de datos de referencia virales: la redundancia entre especies filogenéticamente cercanas y la falta de representación de la variabilidad intrataxon. Esta estructura podría reducir comparaciones redundantes durante la asignación taxonómica y, al mismo tiempo, facilitar la interpretación de los alineamientos que ocurren en regiones compartidas entre varias especies.

A esta estructura, que contiene regiones genómicas junto con su anotación taxonómica, se le denominará “pangenoma viral”. El concepto está inspirado en la propuesta de Tettelin et al. (2005), en la cual el término “pangenoma” se utiliza para describir el conjunto de genes comunes, dispensables y únicos de un grupo bacteriano. Sin embargo, a diferencia del pangenoma bacteriano, el pangenoma viral propuesto en esta tesis se basa en regiones de nucleótidos, no en genes completos. En la Figura 1.3 se presenta un diagrama que ilustra cómo distintas regiones genómicas pueden relacionarse de acuerdo con su similitud.

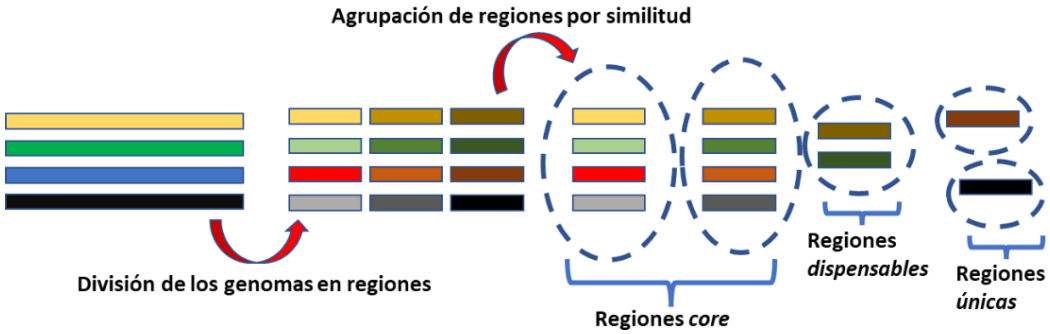

Figura 1.2 Representación conceptual de un pangenoma viral. *El esquema ilustra la organización de regiones genómicas compartidas y específicas entre secuencias virales de un mismo grupo taxonómico.*

1.4 Objetivo general

Desarrollar una base de datos viral con redundancia reducida, basada en pangenomas virales que integren regiones específicas de especie, regiones compartidas entre especies de un mismo género y regiones dispensables a nivel de género, con el fin de apoyar la clasificación taxonómica de secuencias metagenómicas virales.

1.5 Objetivos específicos

- Descargar y organizar una base de datos que incluya todas las secuencias de las especies virales disponibles en la ISND.
- Filtrar la base de datos para eliminar secuencias cortas, incompletas o de baja calidad.
- Asignar anotaciones taxonómicas a nivel de familia, género y especie a las secuencias incluidas en la base de datos viral.
- Desarrollar y ejecutar un algoritmo para generar pangenomas virales a partir de secuencias pertenecientes a una misma especie.

- Desarrollar y ejecutar un algoritmo de reducción de redundancia para integrar pangenomas virales pertenecientes a un mismo género.
- Realizar el posprocesamiento de los pangenomas generados para incorporar los metadatos necesarios para su uso en pipelines metagenómicos.
- Calcular y evaluar las características de los pangenomas virales generados.

1.6 Hipótesis

Es posible modelar la variabilidad genética de secuencias virales pertenecientes a un mismo nivel taxonómico, como especie o género, mediante la identificación y agrupamiento de regiones de alta y baja similitud. Estas regiones pueden organizarse en pangenomas virales con redundancia reducida y utilizarse como bases de datos de referencia para la anotación taxonómica de virus en datos metagenómicos.

Capítulo 2 Metodología

En este capítulo se describe la metodología desarrollada para generar pangenomas virales a nivel de especie y género. El objetivo general del método es transformar una base de datos viral convencional, compuesta por genomas completos o secuencias virales anotadas taxonómicamente, en una base de referencia reducida y organizada en regiones pangenómicas no redundantes. La novedad principal del enfoque radica en que la reducción de redundancia no se realiza únicamente para compactar la base de datos, sino para conservar explícitamente la información taxonómica asociada a cada región genómica. De esta manera, el método permite distinguir entre regiones específicas de una especie y regiones compartidas entre especies de un mismo género, lo cual es relevante para la clasificación de reads metagenómicos.

El paso clave de la metodología es la identificación y etiquetado de regiones redundantes mediante agrupamiento de subsecuencias. A partir de secuencias virales fragmentadas en ventanas de longitud definida, el método identifica subsecuencias idénticas o altamente similares y las organiza en clústeres. Cada clúster conserva la información taxonómica de las secuencias que lo originaron, lo que permite determinar si una región es específica de una especie, compartida entre secuencias de la misma especie o conservada entre distintas especies de un mismo género. Así, el resultado no es solo una base de datos más pequeña, sino una representación estructurada de la diversidad genómica viral.

El procedimiento se organiza en cuatro etapas principales (Figura 2.1). En la primera etapa, se prepara la base de datos viral de entrada. Para esta tesis, las secuencias fueron obtenidas de NCBI Virus mediante su servidor FTP, junto con la información taxonómica necesaria para organizar los genomas por especie y género. Como salida de esta etapa se obtiene un conjunto depurado de secuencias virales anotadas y agrupadas taxonómicamente.

En la segunda etapa, se generan pangenomas virales a nivel de especie. Para cada especie viral, las secuencias correspondientes se fragmentan en subsecuencias y se agrupan de acuerdo con su identidad. Este proceso permite identificar regiones redundantes dentro de la misma especie y seleccionar secuencias representativas. Como resultado, se obtiene una versión reducida del conjunto de secuencias de cada especie, conservando regiones representativas de su diversidad genómica.

En la tercera etapa, se generan pangenomas virales a nivel de género. En esta fase, los pangenomas generados a nivel de especie se utilizan como entrada para identificar regiones compartidas entre especies pertenecientes al mismo género, así como regiones que permanecen como específicas de una especie. Como salida, se obtienen piezas pangenómicas

etiquetadas de acuerdo con su nivel de especificidad taxonómica: regiones asociadas a una especie particular o regiones conservadas a nivel de género.

Finalmente, en la cuarta etapa se realiza la evaluación y el posprocesamiento de los pangenomas generados. Esta etapa incluye la revisión de la reducción obtenida, la conservación de la información taxonómica, la generación de archivos de salida compatibles con herramientas de alineamiento y la evaluación del desempeño de los pangenomas como bases de referencia para clasificación metagenómica.

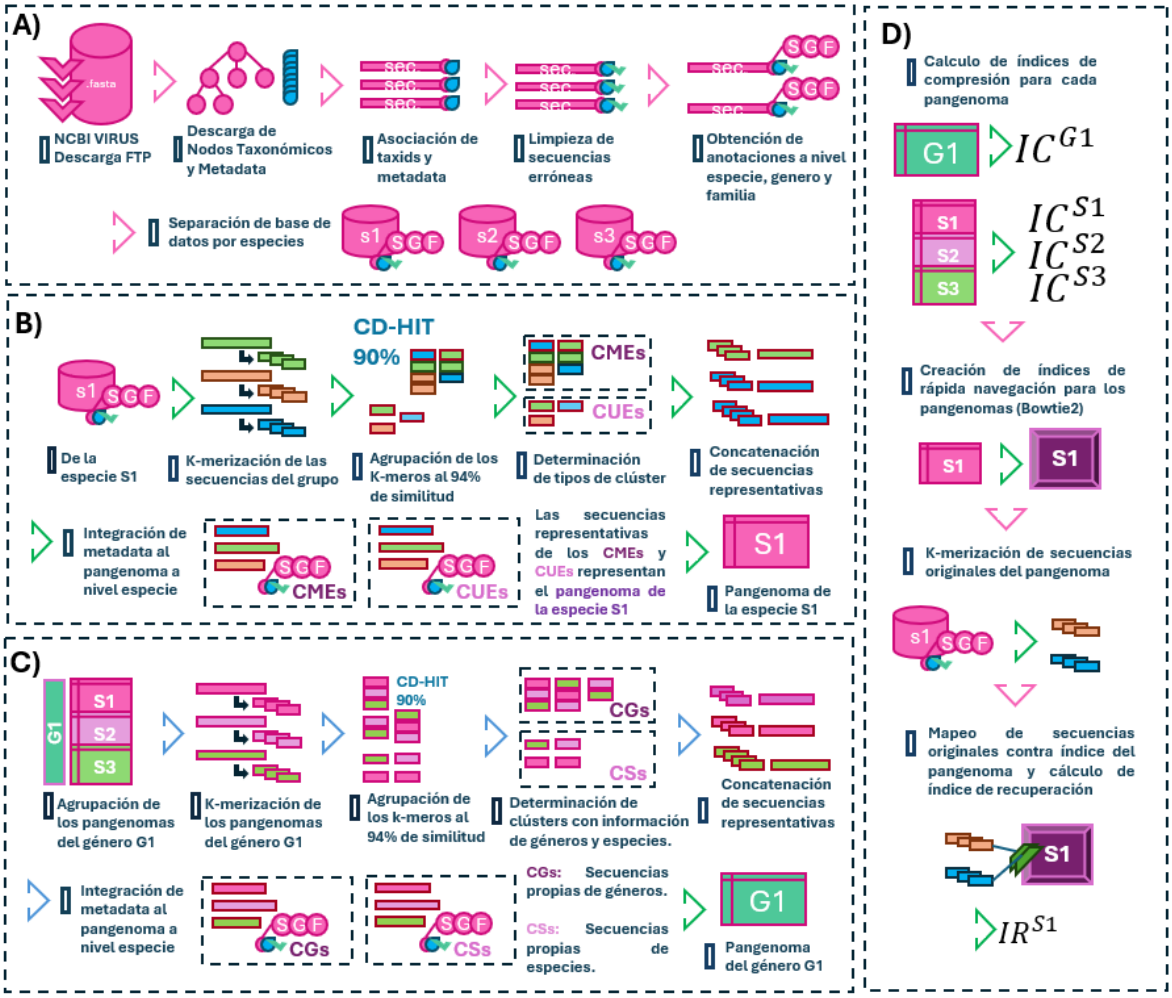

Figura 2.1 Metodología general para la generación de pangenomas virales. *El flujo resume las etapas principales del procedimiento: preparación de la base de datos viral, generación de pangenomas a nivel de especie, generación de pangenomas a nivel de género, y evaluación y posprocesamiento de los pangenomas generados. Las abreviaturas S, G y F indican especie, género y familia, respectivamente, CMEs: clústeres de múltiples elementos, CUEs: clústeres de un elemento, CG: clústeres asociados a géneros, CSs: clústeres asociados a especies.*

2.1 Preparación de base de datos de virus

2.1.1 Asociación de metadata para la base de datos de virus

La base de datos de virus de NCBI se encuentra alojada en los servidores ftp de dicha institución, en el directorio https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/AllNucleotide/ (1 de septiembre 2025). El archivo AllNucleotide.fa, corresponde a un archivo FASTA que contiene todas las secuencias de virus disponibles en la base de datos de nucleótidos. Adicionalmente, NCBI proporciona archivos de metadatos taxonómicos en el directorio https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/, los cuales se describen en la Tabla 2.1.

Tabla 2.1 Descripción de archivos utilizados para la preparación de la base de datos.

Archivo	**Descripción**
AllNucleotide.fa	Archivo FASTA que contiene los encabezados y las secuencias de nucleótidos virales disponibles en la base de datos de NCBI virus (INSD).
nodes.dmp	Archivo tabular que contiene los nodos taxonómicos de NCBI. Incluye, para cada nodo, su *taxid*, el *taxid* del nodo parental y su rango taxonómico.
names.dmp	Archivo tabular que contiene los nombres asociados con los nodos taxonómicos. En este estudio se utilizó el nombre estándar correspondiente al campo "scientific name". Este archivo permitió relacionar cada *taxid* con su nombre taxonómico oficial.
nucl_gb.accession2taxid.dmp	Archivo tabular que relaciona los identificadores de acceso o *accessions* de las secuencias de nucleótidos con su *taxid*. Este archivo permitió asociar cada secuencia con su anotación taxonómica.

Cada una de las secuencias de la base de datos viral de la INSD contiene un identificador llamado *accession number* o *accession*. Este identificador se puede asociar con el archivo nucl_gb.accession2taxid.dmp para obtener el *taxid* correspondiente a cada secuencia (García-López, 2018). El *taxid* es un número identificador numérico asociado con el nodo taxonómico que corresponde a la anotación de dicha secuencia. Los nodos taxonómicos contienen la información de sus nodos padres, por lo que es posible construir el linaje

taxonómico de cada secuencia y obtener sus anotaciones a distintos niveles conectando los nodos taxonómicos.

Por ejemplo, una secuencia anotada al *taxid* 413041*,* se encuentra asociada con un nodo cuyo rango taxonómico es "*no rank*". Esto indica que no tiene un rango taxonómico perteneciente a una clasificación, por lo que probablemente esta secuencia pertenece a algún subtipo o a una subdivisión especifica de una especie. Debido a que este nodo no corresponde a un rango taxonómico, es necesario recorrer sus nodos parentales conectados uno a no dentro del linaje taxonómico hasta encontrar los rangos de interés. En este caso, el nodo parental correspondiente a especie es 3052464, anotado como *Orthoflavivirus denguei*; el nodo parental correspondiente a género es 3044782, anotado como *Orthoflavivirus*; y el nodo parental correspondiente a familia es 11050, anotado como *Flaviviridae* (García-López, 2018).

Después de analizar los nodos taxonómicos, se puede determinar que la secuencia corresponde a la especie *Orthoflavivirus denguei* y es del serotipo 2. Es importante mencionar que, aunque dengue virus type 2 se reconoce comúnmente como un serotipo, este nodo no tiene un rango taxonómico formal en la taxonomía de la ICTV. Esta distinción es relevante, ya que los análisis taxonómicos más específicos que al nivel de especie requieren interpretaciones específicas para cada grupo viral pue sen algunos casos indican cepas, tipos, subtipos u otras anotaciones y no pueden generalizarse a todas las especies.

El metadato taxonómico obtenido para cada secuencia incluye los nombres de especie, género y familia. En este trabajo, el proceso de asignar estos rangos taxonómicos a cada secuencia se denomina limpieza de nodos terminales. Este procedimiento consiste en recorrer el linaje taxonómico desde el nodo más específico asociado con cada secuencia hasta encontrar nodos con rangos taxonómicos reconocidos. Esto es necesario porque las secuencias suelen estar asociadas con nodos terminales que, en muchos casos, no tienen un rango taxonómico formal. En la Figura 2.2 se presenta un esquema del flujo utilizado para la limpieza de nodos terminales (García-López, 2018). Las secuencias que no cuentan, como mínimo, con una anotación a nivel de especie no fueron consideradas para la generación de pangenomas.

Inicialmente, la base de datos viral contenía 14,331,967 secuencias (1 de septiembre 2026). De estas, 9,143,575 pertenecían a la especie *Betacoronavirus pandemicum*, lo cual refleja la sobrerrepresentación de este virus debido a su importancia médica durante la pandemia de COVID-19 de 2020 (Gozashti & Corbett-Detig, 2021). Debido a esta sobrerrepresentación, en la sección 2.1.2 se describe el procedimiento de submuestreo aplicado a esta especie y a otras cuatro especies con alta representación (*Lentivirus humimdef1*, *Betacoronavirus*

pandemicum, *Orthopoxvirus monkeypox*, *Alphainfluenzavirus influenzae* y *Hepacivirus hominis)*, con el fin de disminuir su efecto sobre la generación de pangenomas.

Antes de seleccionar las especies virales candidatas para la generación de pangenomas, se eliminaron las secuencias menores de 350 nucleótidos y solo se consideraron especies con al menos 10 secuencias disponibles. Después de aplicar estos filtros, la base de datos utilizada para generar los pangenomas quedó conformada por 2,048,719 secuencias, distribuidas en 3,897 especies, 608 géneros y 175 familias.

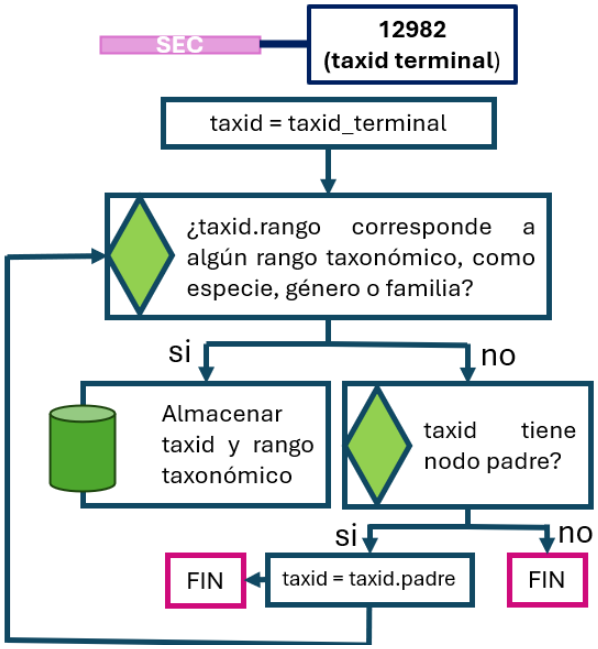

Figura 2.2 Procedimiento de limpieza de nodos terminales. *El esquema representa la asignación de especie, género y familia a partir del recorrido de los nodos parentales del linaje taxonómico de cada secuencia.*

La distribución del número de secuencias por especie se muestra en la Figura 2.3. En esta distribución se observa que una gran cantidad de especies cuenta con pocas secuencias disponibles, lo que evidencia un marcado desbalance en la representación de los virus dentro de la base de datos. Este desbalance es particularmente evidente al comparar especies ampliamente estudiadas con especies poco representadas. Por ejemplo, *Norovirus norwalkense* y *Enterovirus alphacoxsackie* cuentan con 28,198 y 38,578 secuencias, respectivamente, mientras que especies como *Soybean crinkle leaf virus* y *Channel catfish reovirus 730* cuentan únicamente con cinco secuencias.

Dentro de la base de datos viral completa, 60,859 especies cuentan con menos de seis secuencias y 44,773 especies están representadas por una sola secuencia. Este patrón muestra

la alta desigualdad en la disponibilidad de secuencias entre especies virales y justifica el uso de filtros mínimos de longitud y número de secuencias, establecidos en 350 nucleótidos y 10 secuencias por especie, respectivamente. En la Figura 2.3 se muestra la distribución del número de elementos en cada uno de los 3,897 grupos de especie incluidos después del filtrado.

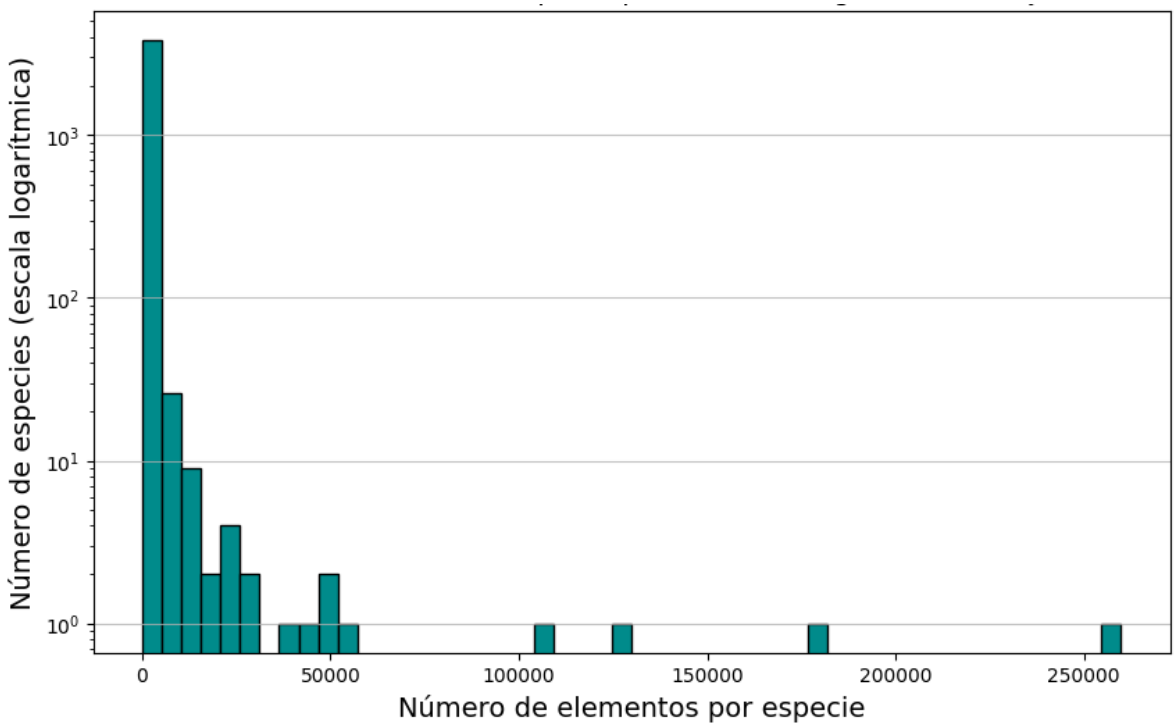

Figura 2.3 Distribución del número de secuencias por especie en la base de datos viral. *Se observa un desbalance en la representación de especies, con muchas especies poco representadas y pocas especies con un gran número de secuencias.*

En las Figuras 2.4A y 2.4B se presentan las distribuciones de la cantidad de elementos por especie para virus que infectan organismos eucariontes y bacterias, respectivamente. En ambos casos, la distribución es similar a la observada en la Figura 2.3, ya que existen numerosos grupos con pocas secuencias disponibles. El objetivo de esta comparación es mostrar que el desbalance en la representación de secuencias ocurre tanto en virus que infectan organismos eucariontes como en virus que infectan bacterias (bacteriófagos).

Entre los virus que infectan organismos eucariontes con mayor número de secuencias se encuentran *Orthoflavivirus denguei*, con 49,347 secuencias, y *Alphainfluenzavirus influenzae*, con 259,812 secuencias. En contraste, entre las especies de fagos con mayor número de secuencias se encuentran *T4virus environmental sample*, con 794 secuencias, y *Efquatrovirus EF3*, con 441 secuencias.

Finalmente, en las Figuras 2.5 y 2.6 se muestra la distribución de los tipos de genoma para virus que infectan organismos eucariontes y para fagos. La mayoría de los virus que infectan organismos eucariontes presentan genomas de RNA, mientras que los bacteriofagos, presentan en su mayoría genomas de DNA.

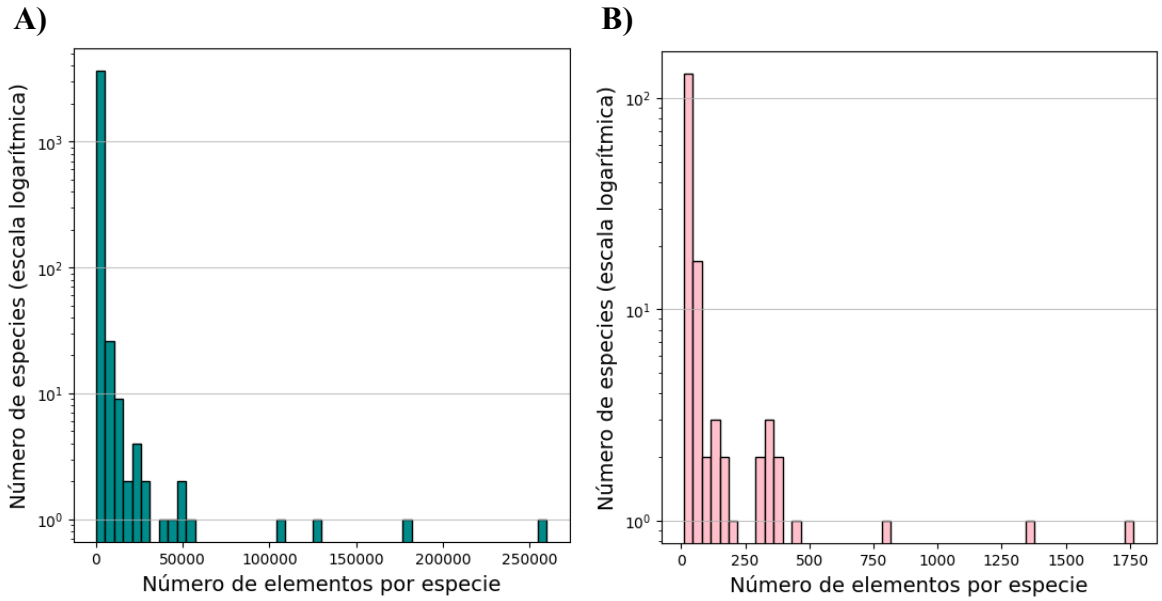

Figura 2.4 Distribución del número de elementos por especie en virus que infectan organismos eucariontes y bacterias. *A) Virus que infectan organismos eucariontes. B) Virus que infectan bacterias.*

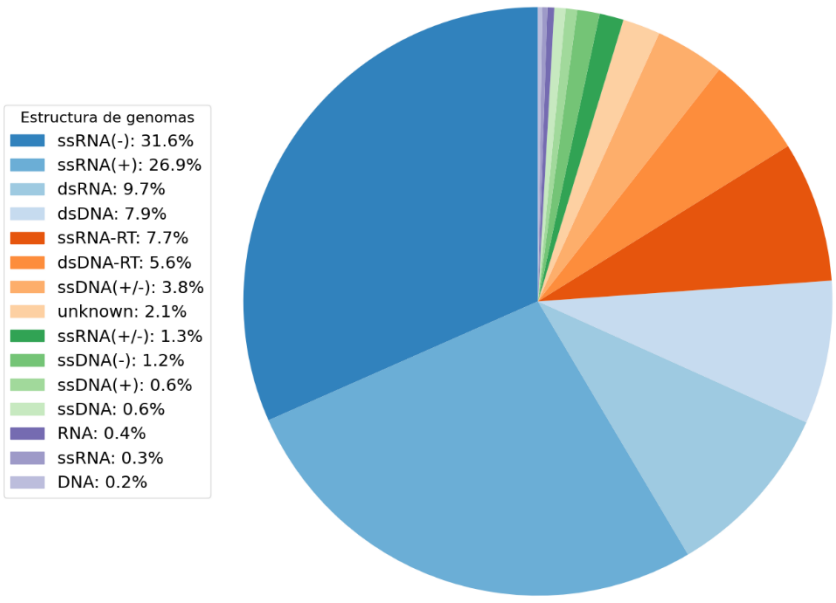

Figura 2.5 Distribución de los tipos de genoma en virus que infectan organismos eucariontes. *La figura muestra la proporción de virus con genomas de DNA, RNA y otros tipos de organización genómica dentro del conjunto analizado.*

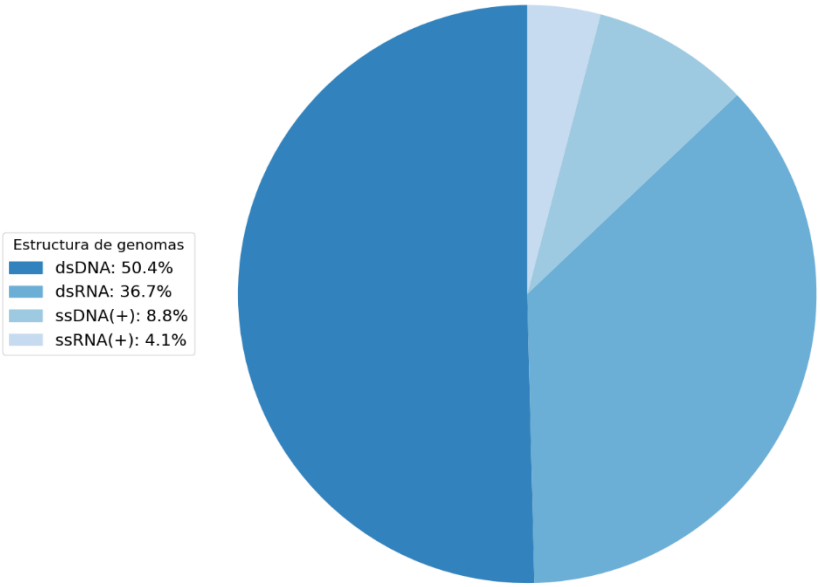

Figura 2.6 . Distribución de los tipos de genoma en virus que infectan organismos procariontes. *Se muestra la proporción de virus con genomas de DNA, RNA y otros tipos de organización genómica dentro del conjunto analizado.*

2.1.2 Submuestreo de especies especiales

Las especies con una cantidad muy alta de secuencias fueron submuestreadas con el objetivo de facilitar la generación de pangenomas y disminuir el efecto de su sobrerrepresentación en la base de datos. En este trabajo, estas especies se denominaron "especies especiales", debido a que sus secuencias recibieron un tratamiento previo distinto antes de ingresar al algoritmo de generación de pangenomas. Las especies consideradas en esta categoría fueron *Lentivirus humimdef1* (HIV), *Betacoronavirus pandemicum* (SARS-CoV), *Orthopoxvirus monkeypox* (mpox), *Alphainfluenzavirus influenzae* (Influenza A*)* y *Hepacivirus hominis* (VHV).

Para conservar la heterogeneidad de estos grupos y evitar que el submuestreo eliminara información relevante, las secuencias se seleccionaron aleatoriamente de acuerdo con combinaciones de parámetros disponibles en los metadatos de cada especie. Por ejemplo, para *Lentivirus humimdef1* se seleccionaron hasta 50 secuencias por año, con el fin de mantener representación temporal dentro del conjunto submuestreado.

Un caso particular fue *Betacoronavirus pandemicum*, debido a que esta especie contiene una gran cantidad de secuencias pertenecientes al subgrupo Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2), con un total de 9,143,575 secuencias. Por esta razón, el submuestreo se aplicó únicamente dentro de este subgrupo. En cambio, las secuencias correspondientes a Severe acute respiratory syndrome-related coronavirus (SARS-CoV), que sumaban 1,721, no fueron submuestreadas. De esta forma, el conjunto final de secuencias de *Betacoronavirus pandemicum* conservó representación de ambos grupos.

Los criterios utilizados para el submuestreo, así como la cantidad de secuencias obtenidas para cada combinación de parámetros, se describen en la Tabla 2.2.

Tabla 2.2 Información del submuestreo aplicado a las especies especiales. *La abreviatura spc indica el número de secuencias seleccionadas por combinación de parámetros.*

Grupo (taxid)	Secuencias iniciales	**Criterios de submuestreo**	**Secuencias finales**
Lentivirus humimdef1 (3418650)	1,189,289	Año: 50 spc	22,311
Severe acute respiratory syndrome coronavirus 2 (2697049*)*	9,143,575	Año/Mes: 50 spc	3,830
Orthopoxvirus monkeypox (3431483)	10,783	Año/Mes: 15 spc	859
Alphainfluenzavirus influenzae (2955291)	1,397,291	Segmento/Año/Subtipo: 10 spc	259,812
Hepacivirus hominis (3052230)	275,985	Año: 50 spc	12,192

2.2 Generación de pangenomas a nivel especie

2.2.1 Definición de grupo candidato a pangenoma

Dada una especie s, para la cual existe un grupo de secuencias de referencia asociadas SEQ_s, se define como grupo candidato a pangenoma al conjunto de secuencias que serán utilizadas para generar el pangenoma $PANG_s$. En este conjunto de secuencias solo se incluyen secuencias con una longitud mínima de 350 nucleótidos. El grupo SEQ_s contiene las secuencias:

$$SEQ_s = \{seq_{s,1}, seq_{s,2}, seq_{s,3}, \dots, seq_{s,n}\}$$

donde n corresponde al número total de secuencias incluidas en el grupo; por lo tanto:

$$|SEQ_s| = n.$$

Cada grupo de secuencias SEQ_s, se utiliza como entrada para el algoritmo de generación de pangenomas. Por ello, cada secuencia $seq_{s,i} \in SEQ_s$, donde $i = (1, 2, 3, \dots, n)$, se divide en

k-meros consecutivos de longitud $l = 200$, con un desplazamiento $r = 50$. Esta combinación genera un sobrelape de 150 nucleótidos entre *k*-meros consecutivos. De esta manera, para cada secuencia $seq_{s,i} \in SEQ_s$ se obtiene un conjunto de *k*-meros:

$$KMERS_i = \{kmer_{i,1}, \quad kmer_{i,2}, kmer_{i,3}, \dots, kmer_{i,j}\}$$

donde j es el número máximo de *k*-meros que se pueden extraer de la secuencia $seq_{s,i}$, calculado de la siguiente manera:

$$j = \frac{len(seq_{s,i}) - l}{r} + 1$$

Cada *k*-mero $kmer_{i,j}$ es evaluado para saber si cumple con un mínimo de calidad. En este trabajo, se considera que un *k*-mero cumple con dicho criterio cuando al menos el 70% de sus posiciones corresponden a nucleótidos no ambiguos; es decir, cuando no más del 30% de sus nucleótidos corresponden a caracteres ambiguos. El conjunto de nucleótidos del *k*-mero se representa como:

$$kmer_{i,j} = \{nuc_1, nuc_2, nuc_3, \dots, nuc_L\}$$

donde $l = 200$.

A partir de esta representación, se define la siguiente variable indicadora:

$$\delta_p = \begin{cases} 1 \; si \; nuc_p \in U \\ 0 \; otro \; caso \end{cases}$$

donde $U = \{A, T, C, G\}$, que corresponde al conjunto de nucleótidos no ambiguos.

La función que determina si el *k*-mero $kmer_{i,j}$ tiene suficientes nucleótidos no ambiguos se define como:

$$f(i,j) = 1 \; si \sum_{p=1}^{L} \delta_p \geq 0.7 * l$$

Finalmente, el conjunto de *k*-meros seleccionados para iniciar el procedimiento de generación de un pangenoma se definen como:

$$K = \{kmer_{i,j} : f(i,j) = 1\}$$

2.2.2 Generación de clústeres por similitud

Para cada grupo de *k*-meros obtenidos a partir de cualquier especie SEQ_s, se realiza un procedimiento ordenado de generación de clústeres utilizando la herramienta CD-HIT 4.8.1 (Fu et al., 2012). Esta herramienta forma clústeres mediante la selección de una secuencia representativa para cada grupo, la cual se utiliza como referencia para determinar si otras secuencias pertenecen al mismo clúster. En este trabajo, el umbral mínimo de identidad para

que una secuencia consulta sea asignada a un clúster fue de 90%. Esta identidad es estimada por CD-HIT mediante la comparación de *k*-meros de menor longitud (Fu et al., 2012).

Como resultado del procedimiento de agrupamiento, se generan dos tipos de clústeres:

(i) Clústeres de múltiples elementos, denominados CME, que corresponden a grupos formados por una secuencia representativa y uno o más *k*-meros que comparten al menos 90% de similitud con ella.

(ii) Clústeres de un solo elemento, denominados CUE, que corresponden a grupos formados únicamente por su secuencia representativa, debido a que esta no alcanza 90% de similitud con ninguna otra secuencia representativa durante el procedimiento de agrupamiento.

Para una especie s, los clústeres de múltiples elementos se definen como sigue:

$$CME_s = \{CME_{1,s}, CME_{2,s}, CME_{3,s}, \dots, CME_{cm,s}\}$$

tal que:

$$\forall CME_{i,s} \in CMES_s, |CME_{i,s}| > 1$$

De manera análoga, para una especie s, los clústeres de un solo elemento se definen como:

$$CUE_s = \{CUE_{1,s}, CUE_{2,s}, CUE_{3,s}, \dots, CUE_{cu,s}\}$$

tal que:

$$\forall CUE_{i,s} \in CUES_s, |CUE_{i,s}| = 1$$

Las secuencias que no corresponden a la secuencia representativa dentro *CME*, se consideran información redundante, ya que representa al menos el 90% de similitud con la secuencia representativa líder de su grupo. Por otro lado, las secuencias representativas obtenidas de los *CUE* representan información única, pues no alcanzan el umbral del 90% de similitud con las secuencias representativas de los demás clústeres generados por la misma especie.

2.2.3 Obtención de piezas pangenómicas mediante la fusión de *k*-meros representativos

Las secuencias representativas de los clústeres se obtienen a partir de los CME y CUE definidos para cada especie (s). En los clústeres de múltiples elementos, cada CME cuenta con una secuencia representativa seleccionada durante el procedimiento de agrupamiento. El conjunto de secuencias representativas de los CME se define como:

$$CMES_s^* = \{CME_{1,s}^*, CME_{2,s}^*, CME_{3,s}^*, \dots, CME_{cm,s}^*\}$$

donde CME_s^* indica la secuencia representativa del clúster $CME_{i,s}$.

En los clústeres de un solo elemento, la única secuencia contenida en cada CUE corresponde a su secuencia representativa, por lo que el conjunto de representantes se define como:

$$CUES_s^* = \{CUE_{1,s}^*, CUE_{2,s}^*, CUE_{3,s}^*, \dots, CUE_{cu,s}^*\}$$

Todas las secuencias representativas obtenidas de la especie s se agrupan en el conjunto:

$$REP_s = \{CMES_s \ , \quad CUES_s\}$$

Estas secuencias representativas corresponden a *k*-meros de 200 nucleótidos derivados de alguna secuencia original $seq_{s,i} \in SEQ_s$. Por ello el j-ésimo *k*-mero de la secuencia $seq_{s,i}$ puede representarse como:

$$kmer_{i,j} \in seq_{s,i}, \quad seq_{s,i} \in SEQ_s$$

Si dos o más *k*-meros representativos provienen de la misma secuencia original, son consecutivos y pertenecen al mismo tipo de clúster, se fusionan eliminando los nucleótidos correspondientes al solapamiento de 150 nucleótidos. Por ejemplo, $kmer_{i,j+1}$ corresponde al *k*-mero consecutivo de $kmer_{i,j}$.

La secuencia resultante de la fusión de dos o más *k*-meros representativos se define como una pieza pangenómica. El conjunto de piezas pangenómicas obtenidas a partir de representantes de los CME de la especie s se define como:

$$PAN_s^m = \{pan_{1,s}^m, pan_{2,s}^m, \dots, pan_{qm,s}^m\}$$

donde qm representa la cantidad total de piezas pangenómicas obtenidas a partir de representantes CME para la especie s y el superíndice m indica que estas piezas provienen de clústeres de múltiples elemento. Estas piezas pangenómicas se denominan "piezas dispensables".

De forma análoga, el conjunto de piezas pangenómicas obtenidas a partir de los representantes de los clústeres de un elemento *CUE* se define como:

$$PAN_s^u = \{pan_{1,s}^u, pan_{2,s}^u, \dots, pan_{qu,s}^u\}$$

donde qu representa la cantidad total de piezas pangenómicas obtenidas a partir de elementos representantes de $CUES_s$ para la especie s, y el superíndice u indica que las piezas pangenómicas provienen de clústeres de un solo elemento. Estas piezas pangenómicas se conocen como "piezas únicas".

De forma que el pangenoma generado de la especie x se expresa así:

$$PGNM_s = \{PAN_s^m, PAN_s^u\}$$

Y su tamaño en número de piezas pangenómicas, expresado como:

$$|PGNM_s| = qm + qu$$

2.3 Generación de pangenomas virales a nivel género

2.3.1 Obtención de clústeres a nivel género y especie

La generación de pangenomas virales a nivel de género se centra en el etiquetado de los clústeres de acuerdo con su distribución entre las especies que conforman un género. En este nivel, el objetivo es distinguir entre secuencias compartidas por varias especies del mismo género y secuencias específicas de una sola especie.

Dado un género g, las secuencias que lo conforman corresponden a distintas especies. Por lo tanto, el conjunto de secuencias asociadas al género puede expresarse como:

$$G_g = \{SEQ_{1,g},\ SEQ_{2,g}, SEQ_{3,g}, \dots, SEQ_{ns,g}\}$$

donde $SEQ_{s,g}$ corresponde al conjunto de secuencias de la especie s dentro del género g y ns representa el número total de especies incluidas en dicho género.

Sin embargo, para generar el pangenoma a nivel de género no se utilizan directamente todas las secuencias originales de cada especie, sino los pangenomas previamente generados a nivel especie. De esta manera, el conjunto de pangenomas de las especies pertenecientes al género g se define como:

$$GEN_g = \{pgnm_{1,g}, pgnm_{2,g}, pgnm_{3,g}, \dots, pgnm_{ns,g}\}$$

donde $pgnm_{s,g}$ corresponde al pangenoma de la especie s) dentro del género g, y ns es el número total de pangenomas asociados al género, equivalente al número de especies consideradas.

El primer paso para generar el pangenoma a nivel de género consiste en dividir cada pangenoma $pgnm_{s,g} \in GEN_g$ en *k*-meros consecutivos de longitud $l = 200$, con salto $r = 50$. Esto produce un sobrelape de 150 nucleótidos entre *k*-meros consecutivos. De esta forma, el conjunto GEN_g puede representarse como un conjunto de *k*-meros:

$$KGEN_g = \left\{kmer_v^{s,g} \middle| s = 1, \dots, ns; v = 1, \dots, nk_s\right\}$$

donde $kmer_v^{sg}$ representa *k*-meros v generados del pangenoma de la especie s dentro del género g y nk_s corresponde al número de *k*-meros generados para dicha especie.

Al igual que el procedimiento a nivel especie, los *k*-meros obtenidos a partir de los pangenomas del conjunto GEN_g se agrupan utilizando CD-HIT 4.8.1 (Fu et al., 2012), con un umbral de 90% de similitud. Como resultado, se obtienen clústeres de múltiples elementos y clústeres de un solo elemento.

Los clústeres de múltiples elementos se definen como:

$$CMS_g = \{CM_{1,g}, CM_{2,g}, CM_{3,g}, \dots, CM_{cm,g}\}$$

tal que:

$$\forall CM_{i,g} \in CMS_g, \left|CM_{i,g}\right| > 1$$

Cada clúster $CM_{i,g}$ está formado por *k*-meros que pueden provenir de una o más especies del género g. Por ejemplo:

$$CM_{i,g} = \{kmer_{v_1}^{s_1,g}, kmer_{v_2}^{s_1,g}, \ldots kmer_{v_m}^{s_m,g}\}$$

donde cada $kmer_v^{s,g}$ conserva la información de la especie s de la cual proviene.

Para identificar la especie de origen de un *k*-mero se define la función:

$$sp\left(kmer_v^{s,g}\right) = s$$

De esta forma, el conjunto de especies representadas en un clúster $CM_{i,g}$ se expresa como:

$$sp\left(CM_{i,g}\right) = \{sp(km) \mid km \in CM_{i,g}\}$$

Con base en esta información, los clústeres de múltiples elementos pueden clasificarse en clústeres puros y clústeres mixtos. Un clúster se considera puro cuando todos sus *k*-meros provienen de la misma especie, mientras que se considera mixto cuando contiene *k*-meros provenientes de dos o más especies del mismo género:

$CM_{i,g}$ es puro si $\left|sp(CM_{i,g})\right| = 1$

$CM_{i,g}$ es mixto si $\left|sp(CM_{i,g})\right| > 1$

Los clústeres puros contienen información específica de una sola especie. Por esta razón, solo se conserva la secuencia líder de cada clúster puro y esta se incorpora al conjunto de clústeres específicos de especie, junto con los clústeres de un solo elemento. Esta decisión permite que, cuando dichas secuencias se utilicen como referencia en estudios metagenómicos, sean tratadas como información específica de una especie.

Por otro lado, los clústeres mixtos contienen *k*-meros compartidos por dos o más especies dentro del mismo género. Por lo tanto, estos clústeres se denominan clústeres de género, o CGS_g, ya que representan información compartida a nivel de género.

Los clústeres de un género se definen como:

$$CGS_g^* = \left\{CG_{1,g}^*, CG_{2,g}^*, \ldots, CG_{cg,g}^*\right\}$$

donde cada $CG_{1,g}$ corresponde a un clúster mixto derivado de CMS_g, y cg corresponde al número total de clústeres del género obtenidos. El superíndice * indica la secuencia representativa o líder de cada clúster de género.

Las secuencias representativas de los clústeres de genero se obtienen seleccionado la secuencia líder de cada clúster $CG_{1,g}$. Para ello se define:

$$rep\left(CG_{ig}\right) = CG_{i,g}^*$$

donde $CG^*_{i,g}$ corresponde a la secuencia representativa o líder del clúster CG_{ig}.

De esta manera, el conjunto de secuencias representativas de los clústeres de genero se expresa como:

$$CGS^*_g = \{rep(CG_{ig}) \mid CG^*_{i,g} \in CG_{cg,g}\}$$

donde CG^*_g representa el conjunto de secuencias compartidas entre especies del mismo género.

En contraste, los clústeres de un solo elemento y los clústeres puros contienen información específica de una sola especie. En los clústeres de un solo elemento, la única secuencia del clúster corresponde directamente a su secuencia representativa. En los clústeres puros, se conserva únicamente la secuencia líder del clúster.

De esta manera, el conjunto de clústeres específicos de especie se define como:

$$CSS_g = \{CS_{1,g}, CS_{2,g}, \dots, CS_{cg,g}\}$$

donde cada $CS_{i,g}$ corresponde a un clúster especifico de especie, ya sea un clúster de un solo elemento o un clúster puro.

Las secuencias representativas de los clústeres específicos de especie se obtienen seleccionando la secuencia conservada para cada clúster $CS_{i,g}$. Para ello, se define:

$$rep(CS_{ig}) = CS^*_{i,g}$$

donde $CS^*_{i,g}$corresponde a la secuencia representativa del clúster específico de especie CS_{ig}.

De esta manera, el conjunto de secuencias representativas de los clústeres específicos de especie se expresa como:

$$CSS^*_g = \{rep(CS_{ig}) \mid CS^*_{i,g} \in CSS_g\}$$

equivalentemente:

$$CSS^*_g = \{CS^*_{1,g}, \quad CS^*_{2,g}, \dots, CS^*_{cg,g}\}$$

donde CSS^*_g representa el conjunto de secuencias específicas de una sola especie dentro del género g, y cg corresponde al número total de clústeres específicos de especie obtenidos en este nivel.

En conjunto, el pangenoma viral a nivel de género queda representado por dos tipos de secuencias, las secuencias representativas de clústeres compartidos entre especies del mismo género, representadas por CGS^*_g, y las secuencias representativas de clústeres específicos de especie, representadas por CSS^*_g. Esta separación permite conservar tanto la información común del género como la información particular de cada especie.

2.3.2 Concatenación de secuencias representativas a nivel género

Cada secuencia representativa obtenida durante la generación del pangenoma a nivel de genero puede expresarse como:

$$srep_{i,j}^{s,g}$$

donde g indica el género al que pertenece la secuencia representativa, s corresponde a la especie de origen, j indica la j-ésima pieza del pangenoma $pgnm_{s,g}$, e i indica que es el i-ésimo k-mero obtenido de la pieza pangenómica.

El tipo de secuencia representativa se define mediante la función:

$$p\left(srep_{i,j}^{s,g}\right) = \begin{cases} 0 \; si \; srep_{i,j}^{s,g,} \in CGS_g^* \\ 1 \; si \; srep_{i,j}^{s,g,} \in CSS_g \end{cases}$$

donde $p\left(srep_{i,j}^{s,g}\right) = 0$ indica que la secuencia representativa proviene de un clúster de género, es decir, de un clúster mixto compartido por dos o más especies del mismo género. En cambio, $p\left(srep_{i,j}^{s,g}\right) = 1$ indica que la secuencia representativa proviene de un clúster específico de especie, ya sea un clúster puro o un clúster de un solo elemento.

Las secuencias representativas se concatenan siempre que sean consecutivas dentro de la misma pieza pangenómica de origen y pertenezcan al mismo tipo de información. Por lo tanto, dos secuencias representativas $srep_{i,j}^{s,g}$ y $srep_{i+1,j}^{s,g}$ se concatenan si cumplen las siguientes condiciones:

$$j\left(srep_{i,j}^{s,g}\right) = j\left(srep_{i+1,j}^{s,g}\right)$$

$srep_{i,j}^{s,g}$ es consecutiva a $srep_{i+2,j}^{s,g}$

$$p\left(srep_{i,j}^{s,g}\right) = p\left(srep_{i+1,j}^{s,g}\right)$$

Cuando estas condiciones se cumplen, las secuencias representativas se concatenan eliminando el sobrelape de 150 nucleótidos generado durante la fragmentación inicial en k-meros de longitud ($l = 200$) y salto ($r = 50$). De esta manera, se reconstruyen fragmentos más largos a partir de secuencias representativas consecutivas que provienen de la misma pieza pangenómica original.

Como resultado de este proceso, se obtiene un conjunto de piezas pangenómicas concatenadas a nivel de género. Estas piezas conservan la información sobre la especie de origen, la pieza pangenómica de la que derivan y el tipo de clúster del que provienen sus secuencias representativas.

El conjunto de piezas pangenómicas generadas a nivel de género se expresa como:

$$pgnmg_g = \{pieza_{j,p}^{s,g} |\ s \in especies(g); j = 1, \dots ns; p\ \in \{0,1\}$$

donde $pieza_{j,p}^{s,g}$ representa una pieza pangenómica concatenada derivada de la especie s, dentro del género g, asociada a la pieza original j del pangenoma de especie $pgnmg_{s,g}$. El valor de p indica el tipo de información contenida en la pieza:

$$p = \begin{cases} 0, \text{si la pieza proviene de secuencias representativas de clústeres de género} \\ 1, \text{si la pieza proviene de secuencias representativas especificos a especie} \end{cases}$$

Por lo tanto, $pgnmg_g$ conserva dos tipos de piezas pangenómicas: aquellas que representan regiones compartidas entre especies del mismo género y aquellas que representan regiones específicas de una sola especie. Esta concatenación permite recuperar fragmentos pangenómicos más largos, manteniendo la distinción entre información común del género e información específica de especie.

2.4 Evaluación y posprocesamiento de los pangenomas generados

2.4.1 Índice de compresión relativa de pangenomas

Para evaluar la cantidad de información reducida en los pangenomas generados, se calcula un índice de compresión relativa. Este índice permite estimar qué proporción de la información original se conserva después del proceso de generación de pangenomas, ya sea a nivel especie o a nivel género.

Utilizando las piezas pangenómicas generadas a nivel de género, el pangenoma del género s se expresa como:

$$pgnmg_g = \{pieza_{j,p}^{s,g} |\ s \in especies(g);\ \ j = 1,2,..,ns_i; x; p\ \in \{0,1\}$$

donde $pieza_{j,p}^{s,g}$ representa una pieza pangenómica concatenada derivada de la especie s, dentro del género g, asociada a la pieza original j del pangenoma de especie $pgnmg_{s,g}$. El valor de p indica el tipo de información contenida en la pieza:

$$p\ = \begin{cases} 0, Si\ la\ pieza\ representa\ información\ compartida\ a\ nivel\ género \\ 1, Si\ la\ pieza\ representa\ información\ específica\ de\ especie \end{cases}$$

El índice de compresión de la especie s se calcula como sigue:

$$ICS_s = \frac{\sum_{p\in\{0,1\}} \sum_{j=1}^{n_{s,p}} len\left(pieza_{j,p}^{s,g}\right)}{len(SEQ_s)}$$

donde n_s es la cantidad total de piezas pangenómicas asociadas a la especie s, $len(pieza_{j,p}^{s,g})$ representa la longitud en nucleótidos de la pieza pangenómica j, y $len(SEQ_s)$ corresponde a la longitud total de las secuencias originales de la especie s.

El índice de compresión relativa para el género g se calcula como:

$$ICG_g = \frac{len(pgnmg_g)}{len(G_g)}$$

donde $en(pgnmg_g)$ corresponde a la longitud total del pangenoma generado a nivel género, calculada como la suma de las longitudes de todas las piezas pangenómicas concatenadas que lo componen, y $len(G_g)$ corresponde a la longitud total en nucleótidos de las secuencias originales incluidas en el género g.

2.4.2 Cálculo del índice de recuperación de pangenomas

El objetivo del índice de recuperación es evaluar la proporción de secuencias de prueba que son reconocidas por un pangenoma generado. Este índice permite estimar qué tan bien el pangenoma conserva la información necesaria para recuperar secuencias provenientes de una especie determinada.

Para una especie s, el índice de recuperación se calcula como:

$$IR_s = \frac{N_{map}}{|N_{sim}|}$$

donde N_{sim}, corresponde al número total de *k*-meros simulados a partir de las secuencias de prueba de la especie s y N_{map} corresponde al número de *k*-meros simulados que logran mapear contra el pangenoma generado.

Las secuencias de prueba de la especie s se definen como:

$$SEQ_s = \{seq_{s,1}, seq_{s,2}, seq_{s,3}, \ldots, seq_{s,n}\}.$$

donde cada $seq_{s,k}$ se divide en *k*-meros consecutivos de longitud m, con salto $r = 50$. La longitud m se define de acuerdo con el sobrelape permitido durante la k-merización de las piezas utilizadas para generar los pangenomas, como se describió en la sección 2.2.1.

De esta forma, el conjunto de *k*-meros simulados a partir de las secuencias de prueba de la especie s se expresa como:

$$SIM(s, m, r) = \left\{kmer_{j,k}^{s} \mid seq_{s,k} \in SEQ_s; j = 1 \ldots . n_k\right\}$$

Donde $kmer_{j,k}^{s}$ representa el j-ésimo k-mero generado a partir de la secuencia de prueba $|\ seq_{s,k}$ y n_k corresponde al número total de k-meros que pueden obtenerse de dicha secuencia utilizando longitud l y salto r.

El número total de k-meros simulados se define como:

$$N_{sim} = |SIM(s, m, r)|$$

Posteriormente, cada k-mero simulado se compara contra el pangenoma generado. El subconjunto de k-meros que logra mapear contra el pangenoma se expresa como:

$$MAP_s = \{km \in SIM(s, m, r) \mid km \text{ mapea contra el pangenoma generaod }\}$$

Por lo tanto, el número de k-meros recuperados se define como:

$$N_{map} = |MAP_s|$$

Valores cercanos a 1 en IR_s indican que el pangenoma generado recupera una alta proporción de las secuencias de prueba de la especie evaluada, mientras que valores cercanos a 0 indican una baja recuperación.

2.4.3 Métricas generales

Para las evaluaciones que involucran clases, especies, géneros o subtipos, se calcularon métricas generales de desempeño con el objetivo de estimar la capacidad del pangenoma para reconocer correctamente las secuencias evaluadas.

La precisión se calculó como:

$$precision = \frac{TP}{TP + FP}$$

La tasa de verdaderos positivos o sensibilidad, se calculó como:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$

La tasa de falsos negativos se calculó como:

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{TP + FN}$$

donde TP representa las predicciones positivas correctas para alguna clase, especie, género o subtipo determinado; FP corresponde a las predicciones positivas incorrectas; FN representa los casos positivos reales que no fueron identificados correctamente.

La métrica TPR evalúa la proporción de positivos reales correctamente identificados, mientras que FNR refleja la proporción de positivos reales omitidos durante la clasificación

realizada con el pangenoma. Por su parte, la precisión indica qué proporción de las predicciones positivas realizadas por el pangenoma corresponde a asignaciones correctas.

Capítulo 3 Resultados y discusión

3.1 Índices de compresión para pangenomas generados a nivel especie

Los pangenomas generados a nivel especie presentaron diferentes índices de compresión, (ICS_s) (sección 2.4.1), los cuales indican la proporción de nucleótidos iniciales que fueron reducidos durante el procedimiento de generación de pangenomas. En la Figura 3.1 se muestra la distribución de estos índices de compresión a través de las 3,897 especies analizadas, pertenecientes a 608 géneros.

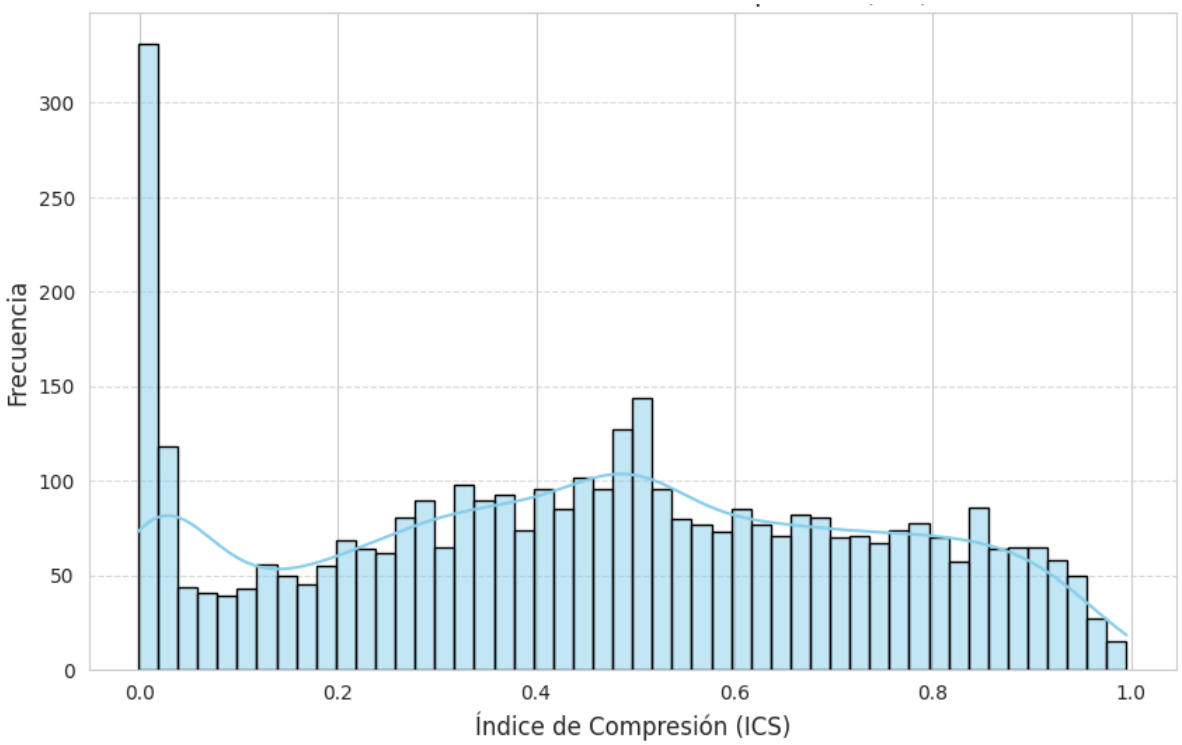

Figura 3.1 Distribución del índice de compresión a nivel especie (ICS_s)**.** *Se muestra la distribución del índice de compresión obtenido para los pangenomas generados a nivel especie. Valores mayores indican una mayor reducción de redundancia durante la construcción del pangenoma.*

Un total de 580 especies presentó menos del 10% de reducción de redundancia, lo que representa el 14.9% del total. En contraste, 3,317 especies, equivalentes al 85.1%, mostraron

índices de compresión mayores al 10%. Además, 1,770 especies presentaron un índice de compresión mayor o igual al 50%, lo que corresponde al 45.4% del total de especies. Estos resultados indican que la mayoría de las especies presentó algún grado de reducción al ser procesada por el generador de pangenomas, aunque la magnitud de la compresión varió ampliamente entre grupos. El valor medio de la distribución fue de 47% de compresión para las 3,897 especies analizadas.

También se observaron algunos casos extremos. Por ejemplo, las especies *Klosneuvirus KNV1* con 16 secuencias, *Pithovirus LCPAC101* con 26 secuencias y *Bat kobuvirus* con 15 secuencias presentaron un índice de compresión de 0%. Este resultado puede deberse a diversos factores, incluyendo posibles problemas de anotación a nivel especie o una elevada divergencia entre las secuencias agrupadas, ya que el algoritmo de generación de pangenomas no detectó similitudes suficientes para reducir la redundancia en estos grupos.

Por otro lado, algunas especies presentaron índices de compresión muy altos. Entre ellas se encuentran *Betainfluenzavirus influenzae* con 178,171 secuencias, *Gequatrovirus G4* con 368 secuencias y *Pistolvirus ichi* con 764 secuencias, con índices de compresión de 99%, 98% y 98%, respectivamente. Estos valores sugieren una alta redundancia entre las secuencias de estas especies, lo que permitió una reducción considerable durante la generación de los pangenomas.

Asimismo, algunos patógenos de interés presentaron niveles de compresión variables, como *Lentivirus humimdef1* con 22,311 secuencias, *Orthopoxvirus monkeypox* con 859 secuencias y *Tenuivirus oryzaclavatae* (RSV) con 803 secuencias, cuyos índices de compresión fueron de 64%, 90% y 82%, respectivamente.

En la Figura 3.2 se presenta un diagrama de barras con las 15 especies que mostraron los índices de compresión más altos. Esto indica que se trata de especies cuyas secuencias presentan una identidad intraespecífica elevada. Es importante señalar que entre estas especies se incluyen tanto grupos muy numerosos, como *Betainfluenzavirus influenzae* con 178,171 secuencias e $ICS_s = 99.5\%$, como grupos relativamente pequeños, como *Pospiviroid impedichrysanth* con 510 secuencias e $ICS_s = 97.7\%$. Este contraste sugiere que el grado de compresión no depende únicamente del número de secuencias, sino principalmente de su nivel de similitud intraespecífica.

Para complementar esta observación, en la Figura 3.3 se muestra un diagrama de barras con las 15 especies más numerosas y sus respectivos índices de compresión. En este caso, se observa que algunas especies grandes, como *Enterovirus betacoxsackie* con 21,067 secuencias, presentan únicamente 58.8% de compresión, lo que sugiere una menor similitud entre sus secuencias de acuerdo con el procedimiento de generación de pangenomas.

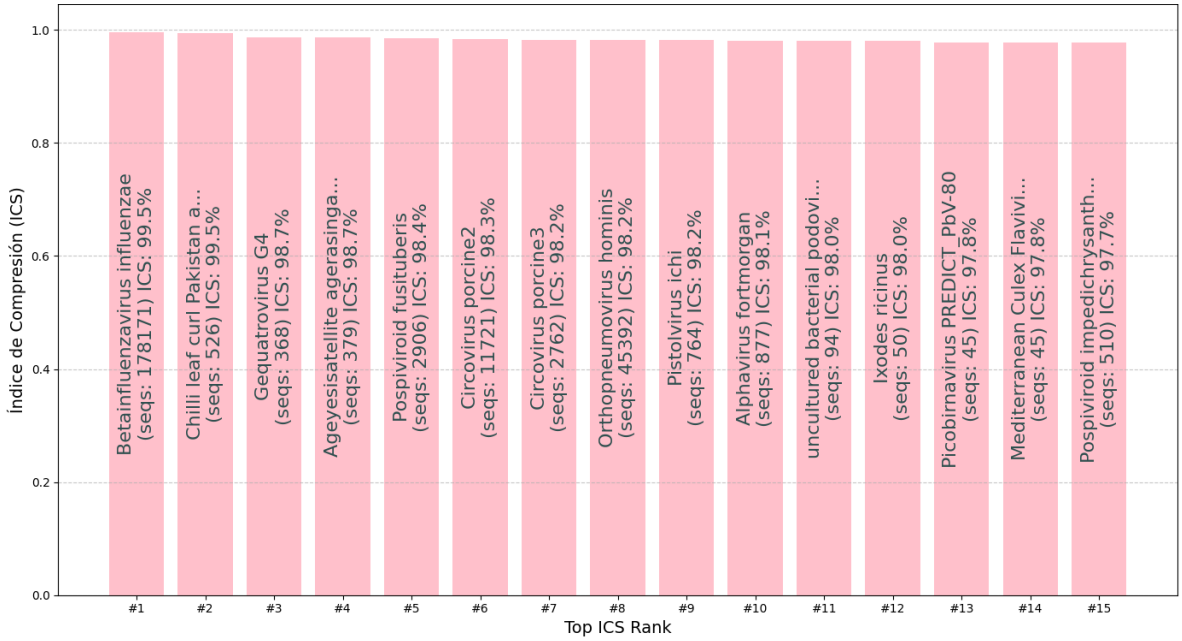

Figura 3.2 Especies con mayor índice de compresión a nivel especie (ICS_s)***.*** *Se muestran las 15 especies con los valores más altos de* ICS_s*, indicando los grupos con mayor reducción de redundancia durante la generación de pangenomas.*

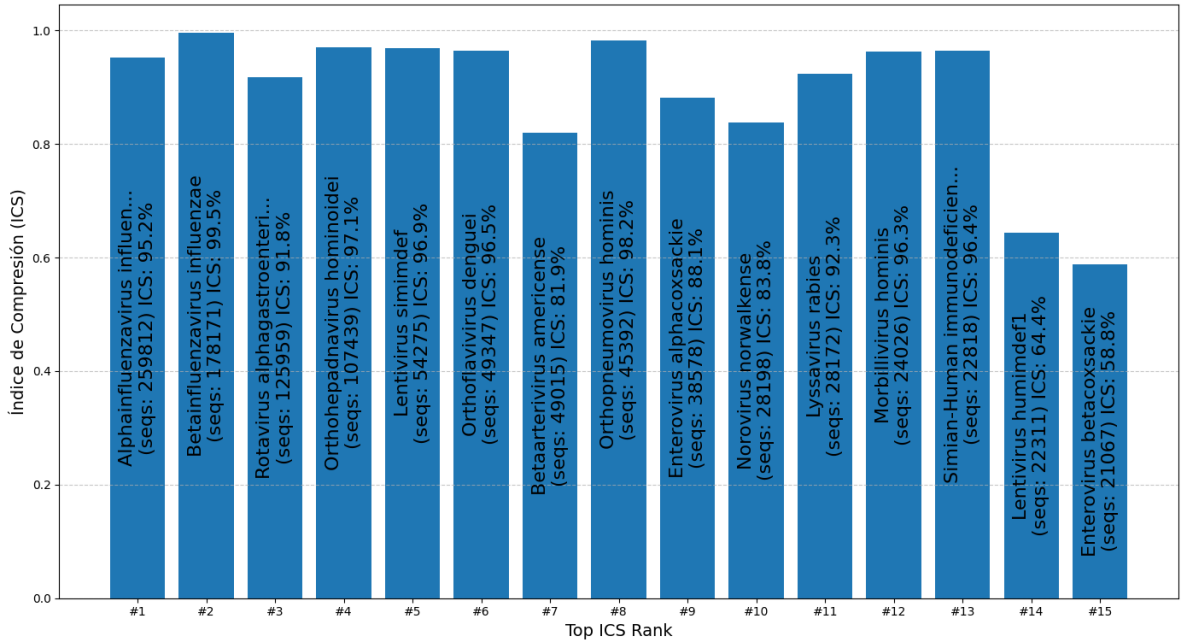

Figura 3.3 Especies más numerosas y sus índices de compresión a nivel especie (ICS_s)*. Se muestran las 15 especies con mayor número de secuencias y sus valores de* ICS_s*, lo que permite comparar el tamaño de cada grupo con su grado de reducción de redundancia.*

En la Figura 3.4 se muestra un gráfico combinado. En la parte superior se presenta un histograma de la distribución del índice de compresión a nivel especie (ICS_s), mientras que en la parte inferior se muestra un diagrama de puntos que relaciona el número de secuencias de cada especie con su respectivo índice de compresión. En conjunto, la figura permite observar que las especies con mayor número de secuencias tienden a presentar índices de compresión más altos. Sin embargo, también se identifican especies con pocas secuencias que alcanzan valores elevados de compresión, lo que indica que el número de secuencias no es el único factor que influye en la reducción de redundancia. De manera general, los grupos más numerosos presentaron índices de compresión superiores al 50%, lo que sugiere que, en estos casos, existe suficiente similitud intraespecífica para permitir una reducción considerable durante la generación de los pangenomas.

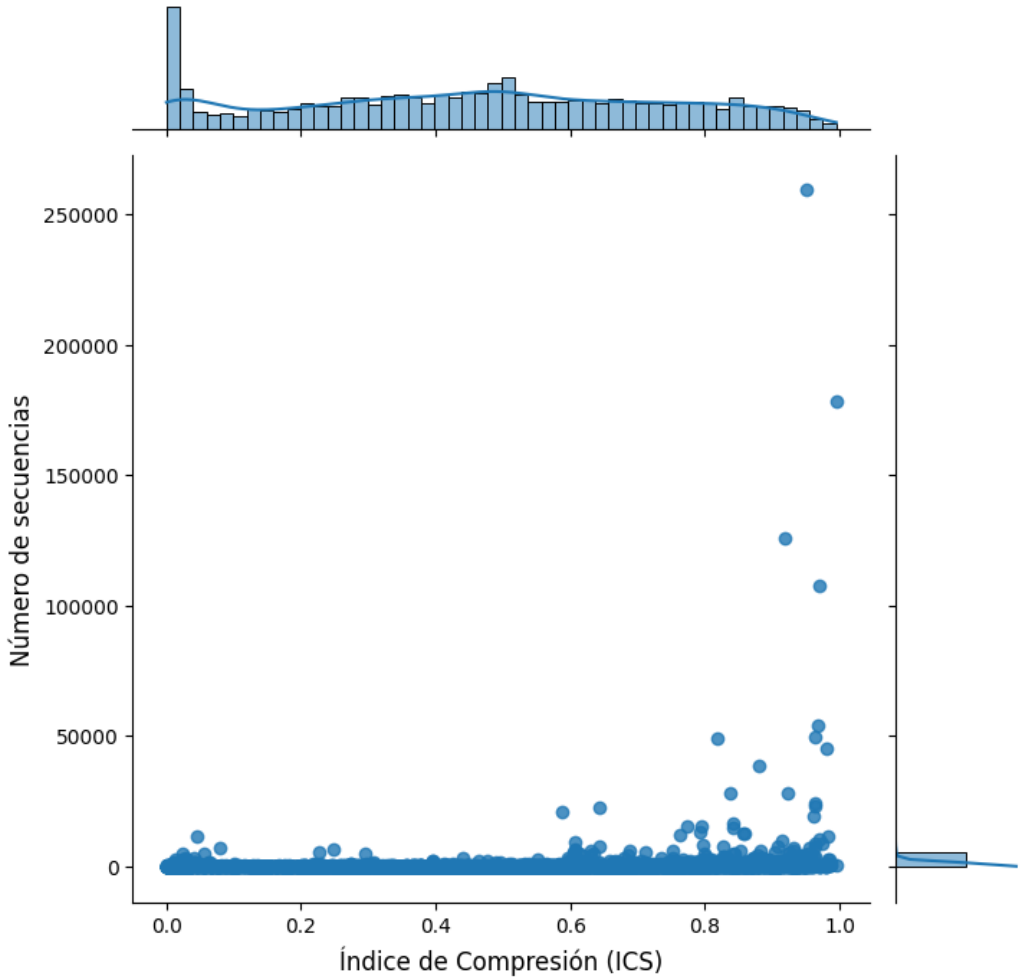

Figura 3.4 Número de secuencias e índice de compresión a nivel especie (ICS_s)**.** *La figura muestra la distribución de* ICS_s *y su relación con el número de secuencias por especie, permitiendo comparar el tamaño de cada grupo con su grado de reducción de redundancia.*

3.2 Índices de compresión para pangenomas generados a nivel género

La compresión a nivel género se llevó a cabo siguiendo el procedimiento descrito en la sección 2.3. Para un género g, las secuencias de entrada corresponden al conjunto de pangenomas generados previamente para las especies hijas de dicho género, es decir:

$$GEN_g = \{pgnm_{1,g}, pgnm_{2,g}, pgnm_{3,g}, \dots, pgnm_{ns,g}\}$$

donde$pgnm_{s,g}$ corresponde al conjunto de secuencias del pangenoma de la especie s dentro del género g y n_s representa el número total de pangenomas de especies asociados a dicho género.

El índice de compresión a nivel género se calculó con respecto a la cantidad de nucleótidos presentes en los pangenomas de entrada y se denominó ICG. Este índice se calculó únicamente para los géneros que contaban con especies asociadas y con pangenomas generados previamente, de acuerdo con lo descrito en la sección 2.4.1.

En total, se comprimieron 608 géneros que integran los pangenomas de 3,057 especies con asignación de género. Además, se identificaron 840 especies sin género asociado, por lo que sus pangenomas no fueron considerados en este procedimiento de compresión a nivel género. En la Figura 3.5 se presenta la distribución de los índices de compresión ICG obtenidos para todos los géneros analizados.

La mediana del índice ICG fue de 43%. Un total de 249 géneros presentó un índice de compresión mayor o igual al 50%, lo que representa el 41.0% de los géneros analizados. Por otro lado, únicamente 39 géneros presentaron una compresión igual o mayor al 80%, equivalente al 6.4% del total. Estos resultados muestran que, aunque una proporción importante de géneros presentó reducción de redundancia, los niveles de compresión a nivel género fueron, en general, menores que los observados a nivel especie.

Algunos géneros, como *Quadrivirus* con 10 secuencias, *Webervirus* con 44 secuencias y *Nouzillyvirus* con 42 secuencias, presentaron un (ICG) de 0%. Esto indica que los pangenomas de las especies incluidas en estos géneros son altamente diversos entre sí o no comparten similitudes detectables mediante el procedimiento de generación de pangenomas. En contraste, géneros como *Pegunavirus* con 1,154 secuencias, *Nerrivikvirus* con 426 secuencias y *Teseptimavirus* con 369 secuencias presentaron índices de compresión de 94.1%, 91.5% y 91.5%, respectivamente.

Debido a que las secuencias de entrada para este procedimiento ya corresponden a pangenomas previamente reducidos a nivel especie, los índices de compresión esperados a nivel género son menores. En este análisis, el valor máximo observado fue de 94.1%,

correspondiente al índice de compresión a nivel género más alto obtenido durante el procedimiento de generación de pangenomas

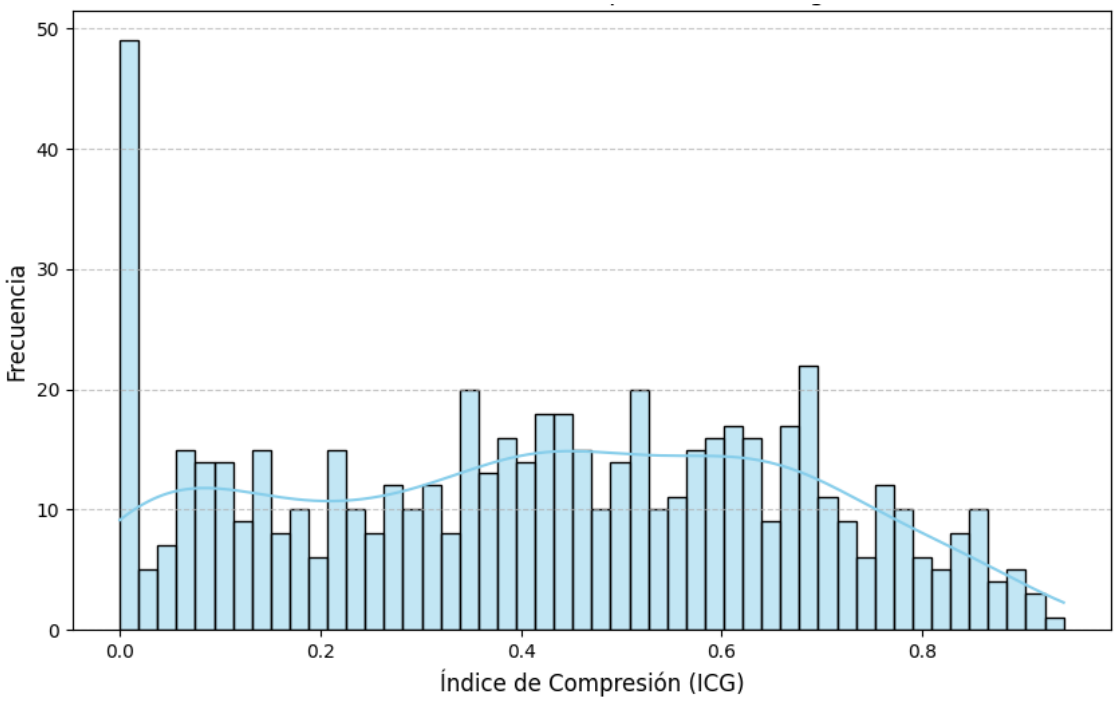

Figura 3.5 Distribución del índice de compresión para pangenomas generados a nivel género (ICG)**.** *La figura muestra la distribución del índice de compresión a nivel género para los 608 géneros analizados. Este índice refleja la proporción de nucleótidos reducidos durante la generación de pangenomas a partir de los pangenomas de las especies hijas de cada género.*

En la Figura 3.6 se presentan los 15 géneros con los índices ICG más altos. Entre ellos se incluyen tanto géneros con un gran número de secuencias pangenómicas, como *Pegunavirus* con 1,154 secuencias e $ICG = 94.1\%$y *Leporipoxvirus* con 1,501 secuencias e $ICG = 90.1\%$, como géneros con un número relativamente bajo de secuencias, como *Acionavirus* con 91 secuencias e $ICG = 86.2\%$. Estos resultados indican que los pangenomas de las especies agrupadas en estos géneros presentan una alta similitud intragenérica.

En contraste, en la Figura 3.7 se muestran, en orden ascendente de ICG, los géneros ubicados entre las posiciones 31 y 45 con los índices de compresión más bajos. Géneros como *Efquatrovirus* con 440 secuencias e $ICG = 0.1\%$y *Galvornvirus* con 3,224 secuencias e $ICG = 1.2\%$presentan una cantidad considerable de secuencias, pero índices de compresión muy bajos, lo que indica una baja similitud entre los pangenomas de las especies que los componen.

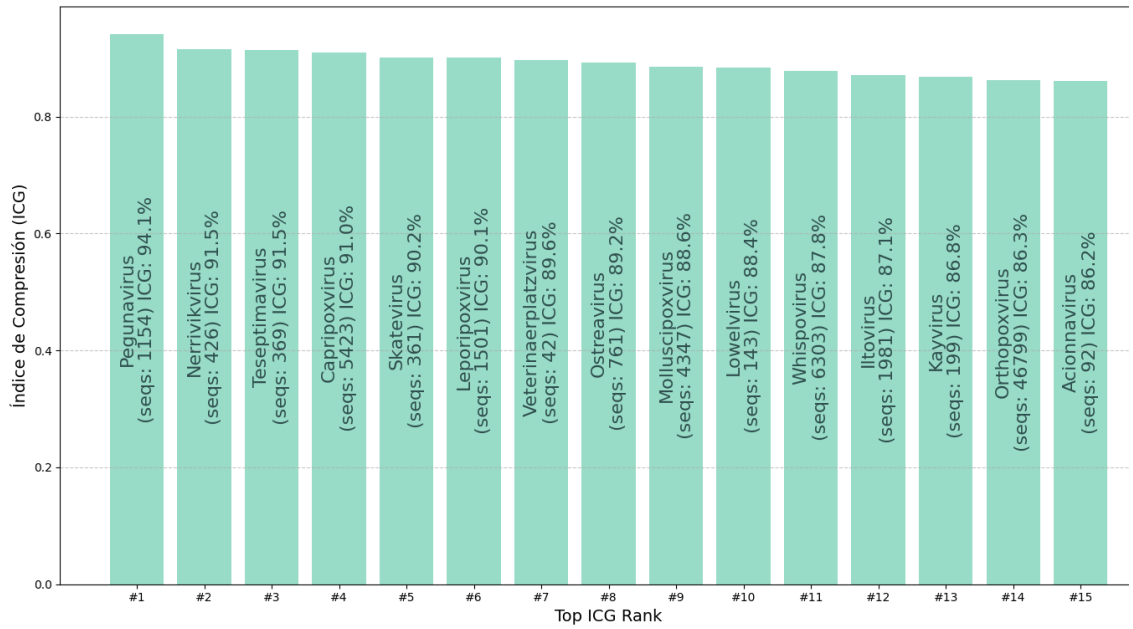

Figura 3.6 Géneros con mayor índice de compresión a nivel género *(ICG).* *Se presentan los 15 géneros con los valores más altos de ICG, correspondientes a los casos con mayor reducción de redundancia durante la generación de pangenomas a nivel género.*

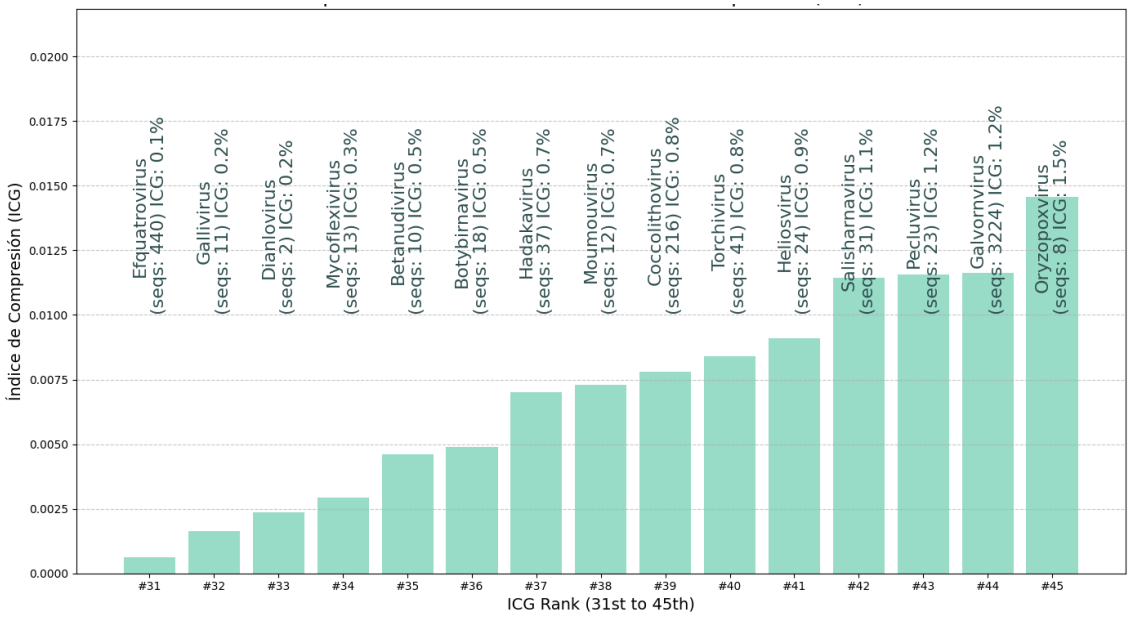

Figura 3.7 Géneros con índices de compresión bajos a nivel género *(ICG).* *Se muestran, en orden ascendente de ICG, los géneros ubicados entre las posiciones 31 y 45 con los índices de compresión más bajos. Estos géneros presentan baja reducción de redundancia entre los pangenomas de las especies que los conforman, lo que sugiere una mayor diversidad intragenérica.*

3.3 Análisis general de índices de recuperación

El pangenoma generado a partir de 2,048,719 secuencias, distribuidas en 3,897 especies y 608 géneros, se evaluó mediante el cálculo del Índice de Recuperación por especie (IR_s) descrito en la sección 2.4.2. A diferencia de los análisis específicos realizados posteriormente para Influenza A, esta evaluación se aplicó de manera global a todas las especies incluidas en la base de datos original. Su objetivo fue determinar si la versión reducida de la base de referencia conserva la información necesaria para recuperar las secuencias representativas de cada especie.

Para este procedimiento, se generaron secuencias artificiales para las 3,897 especies a partir de sus secuencias originales. Estas secuencias artificiales corresponden a k-meros de 150 nucleótidos de longitud, generados con un salto de 100 nucleótidos, como se describe en la sección 2.2.1. Posteriormente, los k-meros se mapearon contra índices de rápida navegación generados con Bowtie 2 (Langmead & Salzberg, 2012). Para las 3,057 especies con género asignado, se utilizó el índice construido a partir del pangenoma a nivel de género, descrito en la sección 2.3.2. Para las 840 especies sin género asignado, se utilizaron los índices generados a partir de sus respectivos pangenomas a nivel de especie, descritos en la sección 2.2.3.

Un IR_s cercano a 1 indica que los k-meros simulados a partir de las secuencias originales de una especie pueden recuperarse utilizando la versión reducida del pangenoma correspondiente. En este contexto, un IR_salto indica que el pangenoma conserva la información necesaria para representar la variabilidad genética contenida en las secuencias originales de esa especie. Por el contrario, un IR_sbajo indicaría pérdida de información durante el procedimiento de generación de pangenomas, debido a que una proporción de los k-meros originales no logra recuperarse a partir de la base reducida.

En la Figura 3.8 se muestra la distribución de los índices de recuperación para las 3,897 especies evaluadas. La mediana de IR_sfue de 100%, mientras que los valores mínimo y máximo fueron de 80.96% y 100%, respectivamente. Además, se muestran líneas verticales correspondientes a los primeros cinco deciles, ordenados de forma ascendente según IR_s, con valores de 99.56%, 99.82%, 99.92%, 99.99% y 100%. Estos resultados indican que, para la gran mayoría de las especies virales incluidas en la base de datos, el pangenoma permite recuperar casi la totalidad de los k-meros derivados de las secuencias originales.

Estos resultados muestran que la reducción de redundancia no implica una pérdida importante de representatividad genómica a escala global. Por ello, el Índice de Recuperación constituye una métrica útil para evaluar la conservación de información en todas las especies virales incluidas en el pangenoma, y no únicamente en grupos virales particulares. Sin

embargo, esta métrica debe interpretarse como una evaluación de recuperación de información respecto a las secuencias originales, no como una validación completa del desempeño taxonómico en muestras metagenómicas reales.

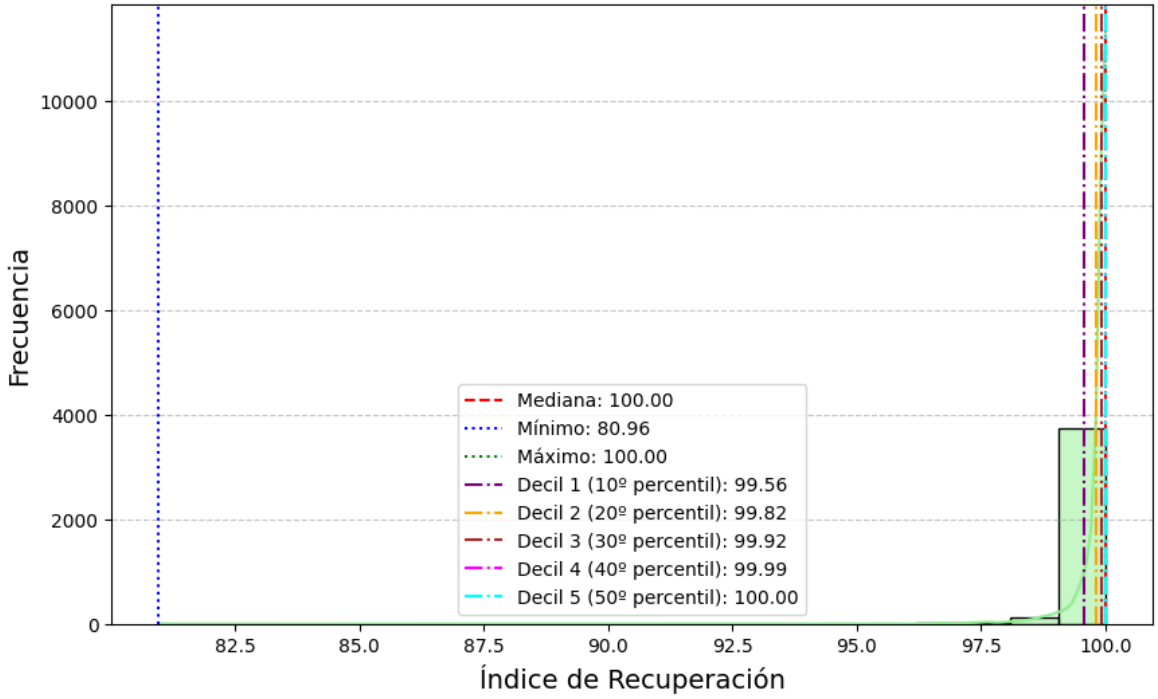

Figura 3.8 Distribución del índice de recuperación IR_s para las 3,897 especies originales. *Las líneas verticales indican los primeros cinco deciles, ordenados de forma ascendente según el valor de IR_s.*

En esta tesis, la validación funcional más detallada se realizó en Influenza A debido a su importancia como patógeno respiratorio de relevancia epidemiológica, a la existencia de subtipos virales bien definidos y a la disponibilidad de datos genómicos reales. Además, este virus forma parte de las actividades de vigilancia genómica desarrolladas actualmente en el laboratorio, lo que permitió contar con experiencia previa, datos adecuados y un contexto biológico relevante para evaluar la utilidad práctica del método. Por esta razón, las siguientes subsecciones se centran en el análisis de Influenza A como caso de estudio para explorar el desempeño del pangenoma en un escenario de clasificación viral más específico. La validación equivalente en otros grupos virales representa una línea futura de trabajo para extender la evaluación práctica del método.

3.4 Generación de pangenomas de subespecies de Influenza A

El Además de su aplicación a nivel de especie y género, el algoritmo de generación de pangenomas puede adaptarse para identificar secuencias únicas y dispensables dentro de grupos de secuencias que pertenecen a una misma especie, siempre que estas cuenten con anotaciones biológicas o taxonómicas por debajo del nivel de especie. Este enfoque permite evaluar si la metodología puede extenderse a escenarios de clasificación más fina, como subtipos, serotipos, linajes u otras categorías intraespecíficas.

Como caso de estudio, en esta tesis se aplicó la metodología a Influenza A, un virus de alta relevancia epidemiológica y sujeto a vigilancia genómica continua en el laboratorio. Esta aplicación corresponde al trabajo principal derivado de esta tesis, presentado en Uscanga Junco et al., 2025, en el cual el generador de pangenomas se utilizó para facilitar la clasificación de subtipos de Influenza A. A partir de este desarrollo se propuso K-FluDB, una base de datos de referencia pangenómica diseñada para conservar secuencias informativas a nivel de subtipo y reducir la redundancia de las bases de referencia convencionales.

La base de datos utilizada para la generación de estos pangenomas se construyó a partir de 895,900 secuencias de Influenza A descargadas de NCBI Virus el 8 de mayo de 2022. Para crear los pangenomas, se tomaron 30 secuencias aleatorias de 8,998 grupos definidos a partir de su metadata, considerando el año de recolección, segmento genómico y anotación de subtipo. Como resultado, el número final de secuencias consideradas para la generación de pangenomas fue de 81,262. En la Tabla 3.1 se muestra la cantidad de secuencias y nucleótidos por cada segmento de influenza A en la base de datos de entrada utilizada para la generación de pangenomas.

Tabla 3.1 Conteo de secuencias inicial para entrada de pangenoma de Influenza A.

Segmento	Conteo de secuencias iniciales	Conteo de nucleótido iniciales
1 (PB2)	9,811	21,962,041
2 (PB1)	9,946	22,239,834
3 (PA)	9,868	21,121,062
4 (HA)	11,287	18,222,760
5 (NP)	9,872	14,841,330
6 (NA)	10,478	14,547,376
7 (M1, M2)	10,056	9,902,780
8 (NS1, NEP)	9,944	8,479,502

El algoritmo de generación de pangenomas se aplicó de manera independiente para cada uno de los segmentos genómicos de influenza A, con el objetivo de mejorar la diferenciación de subtipos. La filosofía general del algoritmo sigue las directrices de la generación de pangenomas a nivel de género; sin embargo, en este caso no se utilizan anotaciones interiores al nivel de especie, sino las anotaciones correspondientes a los subtipos de influenza A. Actualmente se reconocen 18 subtipos de hemaglutinina (HA) y 11 subtipos de neuraminidasa (NA), asociados exclusivamente con la clasificación de los segmentos 4 y 6 del genoma de influenza A, respectivamente.

El procedimiento para la generación de pangenomas inicia con la generación de k-meros (sección 2.2.1). Aquellos *k*-meros que contienen nucleótidos ambiguos son eliminados. Posteriormente, los *k*-meros se agrupan en clústeres de múltiples elementos $CMES$ o clústeres de un elemento $CUES$ (sección 2.2.2), y se procede a la concatenación de k-meros representativos para generar las piezas pangenómicas (sección 2.2.3).

La principal diferencia en la generación de pangenomas a nivel de subtipo reside en la validación de las piezas pangenómicas únicas. Debido a que las secuencias iniciales utilizadas para la generación de pangenomas están submuestreadas, es necesario validar que las piezas pangenómicas únicas sean realmente específicas de cada subtipo. Para ello, se utiliza la base de datos completa de influenza A. A partir de estas secuencias se generan k-meros de 150 nucleótidos de longitud, con un salto de 50 nucleótidos. Estos *k*-meros se mapean contra el índice FM del pangenoma de influenza A, construido para Bowtie 2, y se calcula la huella de mapeo de subtipo o *subtype mapping footprint* (SMF), de la siguiente manera:

$$SMF_{k,t} = \frac{stype}{hits}$$

donde k corresponde al índice de Bowtie 2 asociado a la pieza pangenómica $pang$, t es el subtipo asociado a dicha pieza pangenómica, $hits$ corresponde al número total de *k*-meros que mapean contra $pang$, y $stype$ corresponde al número de *k*-meros que mapean contra $pang$ y que, además, pertenecen al subtipo t.

Si una pieza pangenómica mantiene una huella de mapeo de subtipo (SMF)mayor o igual al 95%, se conserva como una secuencia específica de ese subtipo. En caso contrario, la pieza pangenómica se clasifica como dispensable, debido a que su señal de mapeo se distribuye entre varios subtipos, y se utiliza únicamente como referencia general para Influenza A. Es importante señalar que este umbral no corresponde al criterio de redundancia utilizado para la generación de las piezas pangenómicas, el cual se estableció en 90%, sino a un criterio posterior de consistencia taxonómica. El valor de 95% se eligió como un punto de corte conservador para permitir una proporción baja de asignaciones discordantes, de hasta 5%,

atribuibles a ruido en la anotación taxonómica de las secuencias originales. Durante la revisión exploratoria de las huellas de mapeo, se observó que algunas piezas con una señal dominante para un subtipo presentaban una fracción muy baja de mapeos asociados a otros subtipos, lo cual era más compatible con posibles inconsistencias de anotación que con una señal biológica compartida entre subtipos.

Por esta razón, exigir una SMF de 100% habría descartado piezas altamente específicas afectadas por un número reducido de anotaciones discordantes. En contraste, utilizar un umbral menor al 95% podría aumentar el riesgo de clasificar como específicas de subtipo piezas que en realidad son compartidas entre varios subtipos. Así, el umbral de 95% se utilizó como un criterio operativo para conservar piezas con una señal taxonómica claramente dominante, minimizando al mismo tiempo la incorporación de regiones ambiguas como específicas de subtipo.

El procedimiento de generación de pangenomas se realizó en tres modalidades para influenza A, considerando tamaños de *k*-mer de 75, 150 y 300 nucleótidos, que corresponden a longitudes comunes en plataformas de secuenciación. En la Tabla 3.2 se muestran los índices de compresión calculados con respecto al pangenoma completo, que incluye piezas únicas y dispensables, así como con respecto a la fracción única del pangenoma, integrada únicamente por piezas específicas de subtipo.

Tabla 3.2 Índices de compresión de los pangenomas de influenza A generados con k-meros de 75, 150 y 300 pares de bases. *Se muestran los valores calculados para el pangenoma completo y para la fracción única específica de subtipo.*

Tipo de pangenoma		**Conteo secs**	**Conteo nt**	**IC**
Base de datos de entrada		81,262	131,316,685	-
75 pb	**completo**	22,765	4,373,590	96.67%
	único	10,008	2,273,708	98.27%
150 pb	**completo**	15373	4,412,173	96.64%
	único	6705	2,196,255	98.33%
300 pb	**completo**	8273	3,788,073	97.11%
	único	3592	1,828,642	98.60%

En la Tabla 3.3 se presenta el conteo detallado de nucleótidos para todas las versiones del pangenoma de influenza A, generadas con *k*-meros de 75, 150 y 300 nucleótidos, para cada

uno de los segmentos genómicos. Esta tabla permite observar el porcentaje de nucleótidos necesarios para clasificar información a nivel de subtipo después del procedimiento de generación de pangenomas, en comparación con el conteo inicial de nucleótidos. Para los segmentos 4 y 6, únicamente el 6.61% y 5.64% de los nucleótidos, respectivamente, representan el tamaño del pangenoma con respecto a la cantidad de nucleótidos en las secuencias iniciales de dichos segmentos, que corresponden a 18,222,760 y 14,547,376 nucleótidos. Estos segmentos son los más variables y presentan los conteos de nucleótidos más altos en comparación con los demás segmentos, lo cual es consistente con el papel de HA y NA en la infectividad viral y la evasión del sistema inmune (Bhatt et al., 2011; Poon, 2024; Uscanga Junco et al., 2025).

Tabla 3.3 Desglose de las versiones del pangenoma de Influenza A generadas para diferentes tamaños de k-mero. *Se incluyen las versiones de 75, 150 y 300 pares de bases, así como información específica para cada segmento genómico.*

Segmentos			1 (PB2)	2 (PB1)	3 (PA)	4 (HA)	5 (NP)	6 (NA)	7 (M1, M2)	8 (NS1, NEP)
# Inicial de secuencias			9,811	9,946	9,868	11,287	9,872	10,478	10,056	9,944
# Inicial de nucleótidos			21,962,041	22,239,834	21,121,06	18,222,760	14,841,330	14,547,376	9,902,780	8,479,502
75 pb	dispensable	#secs	3,474	2,818	2,840	103	1,839	116	720	847
		#nt	574,550 (2.62%)	447,300 (2.01%)	451,200 (2.14%)	27,325 (0.15%)	309,275 (2.08%)	24,500 (0.17%)	112,150 (1.13%)	140,825 (1.66%)
	único	#secs	219	141	144	5,316	137	3,895	58	98
		#nt	55,875 (0.25%)	34,275 (0.15%)	40,050 (0.19%)	1,237,450 (6.79%)	30,475 (0.21%)	836,375 (5.75%)	11,650 (0.12%)	17,550 (0.21%)
150 pb	dispensable	#secs	2312	1,978	1,920	71	1,300	62	477	546
		#nt	603,350 (2.75%)	487,050 (2.19%)	484,700 (2.29%)	26,000 (0.14%)	330,950 (2.23%)	19,050 (0.13%)	114,700 (1.16%)	139,400 (1.64%)
	único	#secs	146	87	72	3,613	64	2,627	36	60
		#nt	53,850 (0.25%)	31,600 (0.14%)	32,650 (0.15%)	1,204,350 (6.61%)	20,400 (0.14%)	820,000 (5.64%)	10,600 (0.11%)	16,100 (0.19%)
300 pb	dispensable	#secs	1,235	1,197	978	41	683	35	245	267
		#nt	531,600 (2.42%)	482,850 (2.17%)	409,750 (1.94%)	21,300 (0.12%)	284,650 (1.92%)	16,250 (0.11%)	97,350 (0.98%)	111,000 (1.31%)
	único	#secs	67	49	32	1946	29	1,432	16	21
		#nt	36,900 (0.17%)	29,100 (0.13%)	26,100 (0.12%)	1,006,950 (5.53%)	14,450 (0.10%)	693,100 (4.76%)	8,250 (0.08%)	10,200 (0.12%)

En la Figura 3.9 se evalúa la calidad de la clasificación local de los pangenomas para diferenciar los subtipos de influenza A, considerando los 18 subtipos de hemaglutinina (HA) y los 11 subtipos de neuraminidasa (NA). Para ello, se utilizaron la tasa de verdaderos positivos (true positive rate, TPR) y la tasa de falsos positivos (false positive rate, FPR).

En las Figuras 3.9a y 3.9b se observa que los valores de TPR se concentran en la diagonal, lo que indica una clasificación precisa de los subtipos evaluados. En contraste, los valores de FPR fuera de la diagonal reflejan asignaciones incorrectas entre subtipos. La baja proporción de estos valores fuera de la diagonal indica una alta precisión del pangenoma para identificar los subtipos de influenza A.

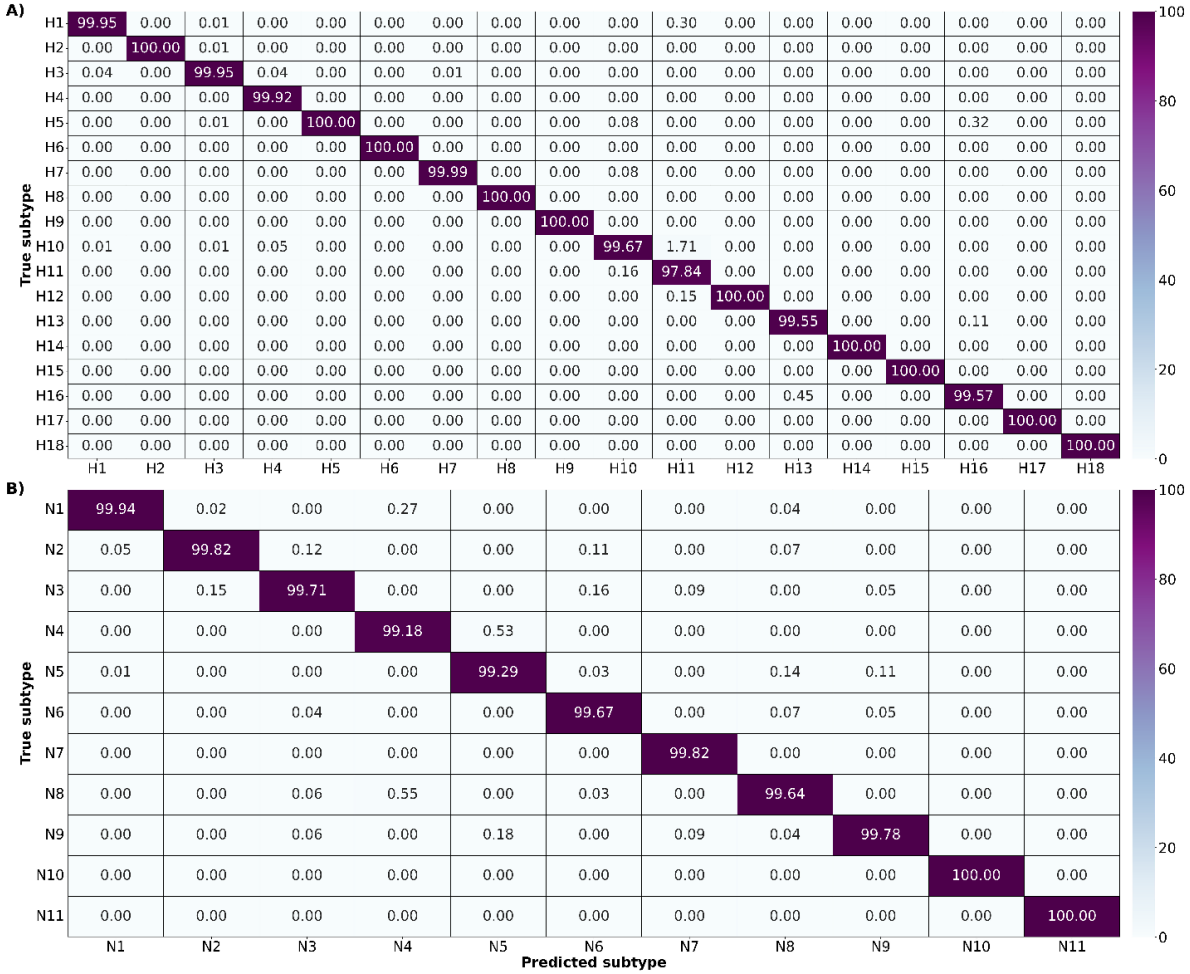

Figura 3.9 Matrices de confusión basadas en TPR y FPR para la clasificación específica de subtipos de influenza A*. A) Clasificación de subtipos HA en el segmento 4, correspondiente a hemaglutinina. B) Clasificación de subtipos NA en el segmento 6, correspondiente a neuraminidasa.*

Para los subtipos HA, la mediana de TPR fue de 99.77%, mientras que para los subtipos NA se alcanzó una mediana de TPR de 99.4%. En conjunto, estos resultados muestran la robustez general del pangenoma para la clasificación local de los subtipos de influenza A.

3.5 Evaluación de pangenomas a nivel subespecie de influenza A con secuencias desconocidas

Para evaluar estos pangenomas se utilizaron dos bases de datos de prueba independientes: una compuesta por secuencias nuevas de Influenza A y otra integrada por secuencias clínicas de Influenza A confirmadas por PCR. La primera base de datos incluyó secuencias de Influenza A obtenidas entre el 9 de mayo de 2022 y el 28 de mayo de 2024. Estas secuencias no formaron parte de los grupos de entrada utilizados para la generación de los pangenomas, por lo que representan un conjunto independiente para evaluar la capacidad del pangenoma de identificar secuencias nuevas de Influenza A. En la Tabla 3.4 se muestra el desglose detallado de las tres partes que conforman esta base de datos de secuencias nuevas.

Tabla 3.4 Resultados del mapeo de secuencias recientes contra los índices del pangenoma completo y del pangenoma de subtipos de Influenza A.

Base de datos de prueba	**Conteo inicial**		**K-meros mapeados to K-FluDB (IR%)**		**Hx**	**Nx**
	Num de secs	***k*-meros**	**completo**	**único**	**k-meros mapeados (TPR%)**	**k-meros mapeados (TPR%)**
Inf A	261,104	7,903,883	7,865,521 (99.51%)	2,017,726 (25.52%)	-	-
H5N1	21,333	647,036	645,027 (99.69%)	168,502 (26.04%)	H5 98,595 (100.00%)	N1 69,907 (100.00%)
H9N2	4,151	126,227	121,539 (96.28%)	34,788 (27.56%)	H9 23,005 (97.52%)	N2 11,199 (100.00%)

En conjunto, se generaron 7,903,883 k-meros a partir de 261,104 secuencias. Al mapear estos k-meros contra el pangenoma completo, se obtuvo un índice de recuperación de 99.51%, lo que indica una alta capacidad para reconocer secuencias de Influenza A. En contraste, el mapeo contra el pangenoma de subtipos, conformado únicamente por los segmentos 4 y 6, produjo una recuperación de 25.52%. Esta diferencia es esperada, ya que este pangenoma contiene únicamente regiones informativas para la identificación de subtipos y no representa de manera completa todos los segmentos del genoma viral. Por lo tanto, estas anotaciones

permiten identificar secuencias de Influenza A, pero no necesariamente están asociadas con un subtipo específico.

Para el análisis específico de secuencias asociadas con subtipos definidos, se utilizaron 21,333 secuencias de Influenza A H5N1, a partir de las cuales se generaron 647,036 k-meros, y 4,151 secuencias de Influenza A H9N2, que produjeron 126,227 k-meros. Los índices de recuperación contra el pangenoma completo fueron de 99.69% para H5N1 y 96.28% para H9N2, lo que confirma que ambas bases de datos fueron clasificadas correctamente como secuencias de Influenza A.

Al mapear estas secuencias contra el pangenoma de subtipos de los segmentos 4 y 6, los índices de recuperación disminuyeron a 26.04% para H5N1 y 27.56% para H9N2. Esta reducción se explica porque el pangenoma de subtipos está constituido únicamente por piezas específicas que permiten distinguir los subtipos de hemaglutinina y neuraminidasa, y no por regiones conservadas o compartidas por todas las secuencias de Influenza A. A pesar de esta menor recuperación global, las métricas de TPR mostraron una clasificación precisa de los subtipos: 100% para H5, 97.52% para H9, 100% para N1 y 100% para N2. Estos resultados indican que el mapeo con Bowtie 2 contra el pangenoma de subtipos permite identificar de manera eficiente los subtipos correctos de Influenza A.

Los pangenomas de subtipos de Influenza A fueron evaluados con 18 muestras clínicas confirmadas por PCR, identificadas como H1N1 (muestras 1–9) y H3N2 (muestras 10–18). Estas muestras se mapearon contra el pangenoma completo de Influenza A, conformado por piezas únicas y dispensables, y contra los pangenomas de los segmentos 4 y 6, correspondientes a hemaglutinina (H) y neuraminidasa (N), respectivamente. Estos últimos pangenomas fueron utilizados como predictores de subtipo, ya que los segmentos 4 y 6 son los empleados para la nomenclatura de los subtipos de Influenza A. En contraste, los pangenomas de los segmentos 1, 2, 3, 5, 7 y 8 no se consideran para dicha nomenclatura, aunque también contienen piezas únicas informativas (Uscanga Junco *et al*., 2025). En la Tabla 3.5 se muestran los resultados de tipado obtenidos para las 18 muestras clínicas.

Los índices de recuperación (IR) contra el pangenoma completo de Influenza A en las 18 muestras clínicas oscilaron entre 74.93% y 96.24%. Estos valores indican la proporción de material genético de Influenza A recuperado en cada muestra mediante el mapeo contra el pangenoma completo. En contraste, los IR obtenidos contra el pangenoma de subtipos fueron considerablemente menores, con valores entre 7.10% y 17.85%. Esta disminución es esperada, debido a que el pangenoma de subtipos contiene únicamente piezas diferenciables entre los subtipos H y N, y no regiones compartidas por todo el genoma de Influenza A.

Tabla 3.5 Resultados del mapeo y tipado de 18 muestras clínicas reales de Influenza A.

ID de muestra	*Reads* iniciales	Tipo por PCR	K-meros mapeados a K-FluDB (IR%)		Subtipo Hx	Subtipo Nx
			Completo	Único sbutipos	K-meros mapeados (TPR%)	K-meros mapeados (TPR%)
01020866UIMY	144,500	H1N1	117,173 (81.09%)	11,514 (7.97%)	H1 5,439 (99.32%)	N1 6,032 (99.90%)
01021509UIMY	225,222	H1N1	209,628 (93.08%)	27,805 (12.35%)	H1 9,447 (99.17%)	N1 18,263 (99.91%)
01021641UIMY	264,098	H1N1	242,483 (91.82%)	44,971 (17.03%)	H1 17,284 (99.60%)	N1 27,611 (99.97%)
01021651UIMY	278,589	H1N1	266,290 (95.59%)	47,879 (17.19%)	H1 18,242 (99.45%)	N1 29,527 (99.97%)
01021863UIMY	222,012	H1N1	166,351 (74.93%)	15,756 (7.10%)	H1 5,100 (99.53%)	N1 10,624 (99.92%)
01021899UIMY	258,869	H1N1	240,944 (93.08%)	44,805 (17.31%)	H1 14,187 (99.27%)	N1 30,501 (99.96%)
01022143UIMY	259,288	H1N1	241,323 (93.07%)	34,246 (13.21%)	H1 10,955 (99.48%)	N1 23,227 (99.97%)
01024183UIMY	124,584	H1N1	116,286 (93.34%)	20,186 (16.20%)	H1 8,180 (99.71%)	N1 11,974 (99.93%)
01024536UIMY	332,814	H1N1	315,018 (94.65%)	56,980 (17.12%)	H1 19,626 (99.41%)	N1 37,218 (99.95%)
01000123UIMY	144,024	H3N2	133,153 (92.45%)	24,030 (16.68%)	H3 9,608 (100.00%)	N2 14,422 (100.00%)
01006970UIBMZ	152,089	H3N2	138,084 (90.79%)	24,470 (16.09%)	H3 9,705 (100.00%)	N2 14,765 (100.00%)
01006974UIBMZ	178,142	H3N2	171,435 (96.24%)	28,087 (15.77%)	H3 10,003 (100.00%)	N2 18,084 (100.00%)
01006991UIBMZ	144,521	H3N2	129,424 (89.55%)	18,699 (12.94%)	H3 6,985 (100.00%)	N2 11,714 (100.00%)
01009323UIBMZ	190,406	H3N2	179,690 (94.37%)	33,994 (17.85%)	H3 13,312 (100.00%)	N2 20,682 (100.00%)
01017109UIMY	223,361	H3N2	191,684 (85.82%)	27,533 (12.33%)	H3 10,794 (100.00%)	N2 16,739 (100.00%)
01040096CIBIN	441,202	H3N2	406,680 (92.18%)	60,308 (13.67%)	H3 22,534 (100.00%)	N2 37,774 (100.00%)
01113643LCE	175,407	H3N2	158,535 (90.38%)	24,505 (13.97%)	H3 8,139 (99.99%)	N2 16,365 (100.00%)
01122196LCE	378,982	H3N2	342,136 (90.28%)	38,227 (10.09%)	H3 16,212 (99.99%)	N2 22,013 (100.00%)

Al evaluar los subtipos sugeridos por el pangenoma, el TPR para el segmento 4, correspondiente al subtipo H, varió entre 99.17% y 100%, mientras que para el segmento 6, correspondiente al subtipo N, osciló entre 99.90% y 100%. Estos valores indican una alta sensibilidad para identificar correctamente los subtipos H1, H3, N1 y N2 en las muestras clínicas evaluadas. En conjunto, estos resultados respaldan la capacidad del algoritmo de

generación de pangenomas para identificar piezas únicas a nivel de subtipo y utilizarlas de manera eficiente para el tipado de Influenza A.

3.6 Análisis comparativo de K-FluDB, RefSeq e INSaFLU para la identificación de subtipos de Influenza A

El pangenoma de Influenza A, denominado K-FluDB, fue comparado con la base de datos de referencia RefSeq (Goldfarb et al., 2025) y con la herramienta INSaFLU (Borges et al., 2018) para la clasificación de muestras de influenza (Tablas 3.6 y 3.7). Para esta comparación se utilizaron *reads* simulados obtenidos a partir de secuencias de Influenza A disponibles en NCBI Virus. Tanto K-FluDB como RefSeq fueron evaluados mediante mapeo con Bowtie 2, mientras que INSaFLU fue utilizada de acuerdo con su estrategia de análisis basada en el ensamblaje de *reads* y la clasificación a partir de contigs.

Tabla 3.6 Comparativa de la detección de subtipos H en el segmento 4 entre las plataformas K-FluDB, RefSeq e INSaFLU.

Subtipo original	Num. de secs	Num. de *k*-meros	*k*-meros mapeados		Subtipos identificados		
			Refseq	Pangenoma	Refseq	Pangenoma	INSaFLU
H1	40,394	560,759	363,804 (64.9%)	55,0260 (98.1%)	H1, H3	H1, H3, H11	H1
H2	702	10,678	2,514 (23.5%)	10,442 (97.8%)	H2	H2, H3	H2, H3
H3	40,258	538,189	404,961 (75.3%)	525,758 (97.7%)	H3, H1, H5	H3, H1, H4, H7, H5	H3, H1
H4	2,118	33,021	0	32,632 (98.8%)	_	H4	H4
H5	6,834	101,550	60,776 (59.9%)	96,999 (95.5%)	H5, H1	H5, H3, H10, H16, H1	H5
H6	1,930	29,541	0	28,580 (96.8%)	-	H6	H6
H7	3,245	50,672	25,286 (49.9%)	49,707 (98.1%)	H7	H7, H10, H3	H7
H8	188	2,864	0	2,835 (99.0%)	-	H8	-
H9	4,337	65,355	13,174 (20.2%)	63,326 (97.9%)	-	H9	H9
H10	1,269	19,794	0	19,338 (97.7%)	-	H10, H11, H3, H1, H4	H10, H11
H11	679	10,650	0	10,202 (95.8%)	-	H11, H10	H11
H12	233	3,673	0	3,618 (98.8%)	-	H12, H11	H12
H13	525	8,500	0	8,144 (95.8%)	-	H13, H3, H16	H13
H14	28	447	0	445 (99.8%)	-	H14	H14
H15	20	330	0	329 (99.7%)	-	H15	-
H16	275	4,472	0	4,334 (96.9%)	-	H16, H13	H16, H13
H17	3	51	0	50 (98.0%)	-	H17	-
H18	2	34	0	34 (100%)	-	H18	-

K-FluDB mostró una mayor sensibilidad en ambos segmentos analizados (Tablas 3.6 y 3.7). Para los subtipos H, correspondientes al segmento 4, K-FluDB obtuvo porcentajes de mapeo entre 95.5% y 100%. Para los subtipos N, correspondientes al segmento 6, los porcentajes de mapeo oscilaron entre 92.31% y 100%. En contraste, los mapeos contra RefSeq presentaron un rendimiento menor. Esto puede explicarse por el diseño compacto y no redundante de RefSeq, que, aunque resulta útil como base de referencia general, puede limitar la recuperación de la diversidad específica de subtipos de Influenza A. Por su parte, INSaFLU pudo detectar varios subtipos; sin embargo, su salida no proporciona el mismo nivel de detalle sobre la calidad de los alineamientos, ya que su clasificación se basa principalmente en contigs derivados del ensamblaje de los *reads*.

Tabla 3.7 Comparativa de identificación de subtipos H entre K-FluDB, RefSeq e INSaFlu.

Subtipo original	**Num. de secs**	**Num. de *k*-meros**	**k-meros mapeados**		**Subtipos identificados**		
			Refseq	**Pangenoma**	**Refseq**	**Pangenoma**	**INSaFLU**
N1	30,870	382,408	252,511 (66.03%)	369,604 (96.65%)	N1, N2	N1, N2, N4, N8	N1, N4
N2	30,427	386,313	291,344 (75.42%)	368,182 (95.31%)	N2, N1	N2, N1, N6, N8, N3	N2, N1
N3	1,624	21,103	594 (2.81%)	20,578 (97.51%)	N2, N9	N3, N2, N6, N9, N7	N3, N2
N4	371	4,775	0	4,688 (98.18%)	-	N4, N5, N2	N4
N5	576	7,435	4 (0.05%)	7,260 (97.65%)	N9	N5, N8, N9, N1, N6, N2	N5, N8
N6	2,912	37,717	1 (0.00%)	36,738 (97.40%)	N9	N6, N8, N9, N3	N6
N7	1,115	14,202	0	13,980 (98.44%)	-	N7	N7
N8	2,806	36,453	0	35,116 (96.33%)	-	N8, N4, N6, N3	N8
N9	1,886	24,056	16,740 (69.59%)	23,512 (97.74%)	N9	N9, N8, N7, N5, N3	N9
N10	3	39	0	36 (92.31%)	-	N10	-
N11	2	26	0	26 (100%)	-	N11	-

En algunos casos particulares, K-FluDB permitió identificar coincidencias con subtipos adicionales de H y N. Por ejemplo, en secuencias de aves originalmente anotadas como H1, se detectaron coincidencias con el subtipo H10. Estos resultados fueron revisados mediante alineamientos locales con BLAST BLAST (Altschul et al., 1990), los cuales mostraron que dichas secuencias estaban incorrectamente anotadas como H1. Este resultado resalta la utilidad de K-FluDB no solo para la identificación de subtipos, sino también como una estrategia para detectar posibles inconsistencias en la anotación de secuencias de referencia pudiendo ser potencialmente utilizada como una herramienta en el control de calidad de las anotaciones.

A lo largo del análisis también se detectaron múltiples subtipos en algunas muestras, lo que podría sugerir la presencia de coinfecciones, señales compatibles con reagrupamiento viral o diferencias en la anotación original de las secuencias. Estos escenarios son relevantes en Influenza A, un virus con genoma segmentado en el que la coinfección de una misma célula por diferentes virus puede favorecer eventos de reagrupamiento genético (Lowen, 2017). En las Tablas 3.6 y 3.7 se muestran los k-meros originalmente identificados para cada subtipo, así como los resultados de clasificación obtenidos con K-FluDB, RefSeq e INSaFLU para los subtipos H y N, respectivamente.

3.7 Análisis de subtipos de Influenza A en muestras genómicas reales del periodo pospandémico en México.

Dentro de los resultados de esta tesis, K-FluDB fue aplicado a datos genómicos reales en un estudio de vigilancia genómica, artículo en el cual el estudiante es coautor (Taboada et al., 2026). En dicho estudio se analizaron 1,041 muestras genómicas provenientes de casos confirmados de Influenza A, correspondientes a los subtipos H1, H3, N1 y N2. El objetivo del trabajo fue analizar la evolución epidemiológica, genómica y evolutiva de los patrones de circulación de Influenza A en México entre 2018 y 2023, con énfasis particular en el periodo pospandémico comprendido entre junio de 2022 y junio de 2023. Los *reads* fueron obtenidas mediante la plataforma Illumina NextSeq 500, en modalidad de secuenciación pareada de 2 × 150 pb.

En este análisis, K-FluDB (Uscanga Junco et al., 2025) fue utilizada como base de datos de referencia para confirmar los subtipos de Influenza A identificados en las muestras. Para ello, se empleó Bowtie 2.3.4.4 (Langmead & Salzberg, 2012) para generar los índices y realizar el mapeo contra los pangenomas únicos de los segmentos 4 y 6, correspondientes a hemaglutinina (H) y neuraminidasa (N), respectivamente. A diferencia de las evaluaciones realizadas con secuencias completas o *reads* simulados, este análisis permitió evaluar el desempeño de K-FluDB en datos reales de secuenciación, los cuales pueden presentar variación en la cobertura, calidad de los *read*s y proporción de material viral recuperado.

En la Tabla 3.8 se presenta un resumen de los subtipos identificados a partir de los *reads* que mapearon contra el pangenoma. Los resultados muestran que las coincidencias obtenidas para el segmento 4 correspondieron a los subtipos esperados H1 y H3, mientras que las coincidencias para el segmento 6 correspondieron a los subtipos esperados N1 y N2. No se detectaron subtipos discordantes de H o de N, lo que concuerda con lo esperado para el conjunto de muestras analizado y confirma la consistencia del pangenoma para la identificación de subtipos en muestras clínicas.

Tabla 3.8 Subtipos de Influenza A identificados en 1,041 muestras genómicas reales mediante K-FluDB.

Subtipos H	**H1**	**H3**	**TOTAL**
Conteo *reads*	689,071	5,757,534	6,446,605
%	10.69	89.31	
Subtipos N	**N1**	**N2**	**TOTAL**
Conteo *read*	1,185,973	9,805,023	10,990,996
%	10.79	89.21	

Estos resultados respaldan el uso de K-FluDB como base de referencia para la confirmación de subtipos de Influenza A en datos genómicos reales. Además, demuestran que los pangenomas únicos de los segmentos 4 y 6 pueden ser aplicados en estudios de vigilancia genómica para apoyar la identificación de subtipos circulantes, incluso cuando se trabaja con *reads* provenientes de muestras clínicas y no únicamente con genomas completos o datos simulados.

3.8 Disponibilidad de recursos

El código base utilizado para la generación de pangenomas se encuentra disponible en el siguiente repositorio:
https://github.com/usjunco/pangen

La base de datos reducida para la identificación de subtipos de Influenza A está disponible en Zenodo:
https://zenodo.org/records/17203072

Este repositorio incluye las versiones de 75, 150 y 300 nucleótidos, diseñadas para ser utilizadas como bases de datos de referencia con *reads* de diferentes tamaños, de acuerdo con la plataforma o estrategia de secuenciación empleada.

El pangenoma general de Influenza A, optimizado para datos generados con instrumentos de secuenciación de 150 pares de bases, se encuentra disponible en Zenodo:
https://zenodo.org/records/20533147

Capítulo 4 Conclusiones

Las bases de datos de referencia cobran cada vez mayor importancia en los pipelines metagenómicos, ya que constituyen el punto de comparación para determinar qué organismos se encuentran presentes en una muestra a partir del análisis de su material genético. Por ello, es fundamental considerar que la calidad de las secuencias de referencia influye directamente en la veracidad de los resultados obtenidos. En este sentido, garantizar la precisión, actualización y correcta anotación de los genomas de referencia es una tarea indispensable para la selección y aplicación de cualquier pipeline bioinformático (Martí et al. 2025).

El procedimiento de generación de pangenomas propuesto en esta tesis permite analizar grupos de secuencias virales y obtener una representación reducida de ellas, conservando al mismo tiempo su anotación taxonómica a nivel de especie y género. Esta funcionalidad es relevante porque, cuando una lectura presenta similitud con una región conservada entre especies del mismo género, no siempre es posible asignarle directamente una taxonomía a nivel de especie. En estos casos, se requieren procedimientos adicionales, como el análisis de ancestro común o estudios de abundancia, para determinar la o las especies que describen con mayor precisión la taxonomía de dicha lectura. Sin embargo, estos procedimientos implican un costo computacional adicional, que puede minimizarse mediante el uso de secuencias de referencia más informativas y mejor estructuradas para la anotación taxonómica.

La reducción de redundancia a nivel de especie permitió generar pangenomas de referencia para virus de importancia médica, tales como *Lentivirus humimdef1* —VIH—, *Alphainfluenzavirus influenzae* —Influenza A— y *Orthopneumovirus hominis* —virus sincitial respiratorio—, los cuales presentaron índices de compresión a nivel especie de 64.4%, 95.2% y 98.19%, respectivamente. Estas versiones compactas permitieron recuperar al menos el 99.12% de los k-meros derivados de las secuencias originales. Además, la mediana de recuperación de 100% observada en las 3,897 especies evaluadas indica que la reducción del tamaño de las bases de referencia no compromete, en términos generales, la recuperación de regiones derivadas de las secuencias originales. Estos resultados respaldan que la metodología conserva información relevante para la recuperación de lecturas y para su posible uso en tareas de anotación taxonómica en análisis genómicos y metagenómicos.

La generación de pangenomas a nivel de género permitió identificar regiones compartidas y regiones exclusivas entre especies de un mismo género. La presencia de piezas pangenómicas dispensables, compartidas entre dos o más especies, permite reconocer regiones que no son suficientemente informativas para una asignación taxonómica a nivel de especie y que, por lo tanto, deben asociarse a un nivel superior, como el género. Un ejemplo importante fue el

género *Pegunavirus*, cuyo pangenoma contiene secuencias únicas que permiten diferenciar cuatro especies: *Pegunavirus Suffolk*, *Pegunavirus Pg1*, *Pegunavirus oline* y *Pegunavirus soto*. Este resultado muestra que los pangenomas no solo reducen redundancia, sino que también organizan la información genómica de acuerdo con su valor taxonómico.

La generación de pangenomas a nivel subespecífico sigue una lógica similar a la empleada para los pangenomas a nivel de género, pero se orienta a obtener secuencias únicas que permitan diferenciar anotaciones más específicas, como los subtipos de Influenza A. Este virus se utilizó como caso de estudio debido a su importancia epidemiológica, a la existencia de subtipos bien definidos, a la disponibilidad de datos genómicos reales y a que actualmente forma parte de las actividades de vigilancia genómica desarrolladas en el laboratorio. En este contexto, la base de datos pangenómica K-FluDB permitió evaluar la utilidad del método en un escenario de clasificación viral más fino.

Al comparar el pangenoma de Influenza A con RefSeq e INSaFlu, los resultados mostraron una mayor sensibilidad para la identificación de subtipos, incluyendo subtipos que no fueron recuperados con las bases o herramientas comparadas, como H15, H17, H18, N10 y N11. Además, un resultado especialmente relevante fue la identificación de secuencias de aves anotadas originalmente como H1, pero clasificadas por el pangenoma como H10. Este caso muestra que el enfoque propuesto no solo puede utilizarse para clasificar secuencias, sino también como una herramienta de control de calidad de anotaciones, al detectar posibles inconsistencias taxonómicas en las bases de datos de referencia. Por lo tanto, los pangenomas virales pueden contribuir no solo a la clasificación metagenómica, sino también a la curación y revisión de bases de datos genómicas.

Finalmente, las anotaciones realizadas con 1,041 muestras genómicas reales y 18 muestras clínicas confirmadas como H1N1 y H3N2 demostraron la viabilidad de los pangenomas como secuencias de referencia para la anotación taxonómica. En conjunto, los resultados de esta tesis demuestran que la generación de pangenomas virales representa una estrategia viable para construir bases de datos de referencia más compactas, informativas y útiles para la clasificación viral. La metodología propuesta permite reducir la redundancia de las secuencias originales, conservar información relevante para la recuperación de lecturas, distinguir regiones compartidas o exclusivas en distintos niveles taxonómicos y detectar posibles errores de anotación.

No obstante, el desempeño de esta estrategia depende de la calidad de las secuencias de entrada, de la precisión de sus anotaciones taxonómicas y de la actualización periódica de las bases de datos de referencia. Aunque el Índice de Recuperación permitió evaluar la conservación de información en todas las especies incluidas en la base de datos, la validación funcional más detallada se realizó en Influenza A. Por ello, la aplicación y evaluación

equivalente en otros grupos virales representa una línea futura de trabajo. En este sentido, los pangenomas virales generados en esta tesis constituyen tanto un recurso bioinformático como una base metodológica para futuros desarrollos en clasificación viral, vigilancia genómica, control de calidad de anotaciones y análisis de secuencias de referencia.

Perspectivas

Como parte de los trabajos de mantenimiento y actualización necesarios para la generación de pangenomas de los virus disponibles en las bases de datos de NCBI, el pipeline podría adaptarse para incorporar *reads* de diferentes longitudes, por ejemplo, de 75 y 300 nucleótidos, como se realizó para los pangenomas de Influenza A. Estas actualizaciones permitirían ampliar la compatibilidad de los pangenomas con distintas plataformas y configuraciones de secuenciación, favoreciendo su uso en diferentes contextos experimentales.

Este tipo de bases de datos de referencia requiere actualizaciones constantes, ya que la taxonomía viral se modifica de manera continua. Un ejemplo de ello son los ajustes realizados en diciembre de 2025 por el Comité Internacional de Taxonomía de Virus (ICTV) (Simmonds *et al*., 2025), los cuales no solo incluyeron la incorporación de nuevas especies, sino también la reorganización de grupos taxonómicos y la adición de nuevas etiquetas de nomenclatura. Por lo tanto, el mantenimiento periódico de estas bases de datos es indispensable para conservar la coherencia taxonómica y la utilidad de las secuencias de referencia.

Asimismo, como se realizó en Uscanga Junco et al., 2025 con K-FluDB para obtener secuencias únicas a nivel subespecie, este flujo de trabajo podría replicarse en otros virus patógenos que requieren diferenciación por categorías subespecíficas. Un ejemplo relevante es *Orthoflavivirus denguei*, virus dengue, que comprende cuatro tipos virales. No obstante, para determinar la viabilidad de este procedimiento en otros grupos virales, sería necesario realizar análisis preliminares de preprocesamiento que permitan evaluar si la identidad entre los subtipos o tipos virales es compatible con la generación de pangenomas específicos.

Finalmente, el procedimiento de generación de pangenomas podría aislarse y desarrollarse como una herramienta independiente para construir índices de rápida navegación, compatibles con herramientas populares de alineamiento como Bowtie 2 (Langmead & Salzberg, 2012) y BWA (Li & Durbin, 2009). Esta herramienta podría aplicarse a cualquier grupo de secuencias de interés, siempre que se encuentren correctamente anotadas. Sin embargo, para consolidar esta contribución sería necesario adaptar el pipeline a lineamientos

formales de ingeniería de software, con el fin de generar una herramienta robusta, documentada y de fácil acceso mediante línea de comandos.

Productos académicos derivados

Artículos de investigación arbitrada

Artículo de primer autor: "K-FluDB: a novel K-mer-based database for enhanced genomic surveillance of Influenza A viruses" por Alejandro Uscanga Junco, Lorena Díaz-González y Blanca Taboada. Publicado en Bioinformatics Advances, Volumen 6, número 1, 2026.

DOI: https://doi.org/10.1093/bioadv/vbaf254

"Post-Pandemic Genomic Diversity and Lineage Turnover of Influenza Viruses in Mexico During 2022–2023" por Blanca Taboada, Selene Zárate, José Esteban Muñoz-Medina, Joel Armando Vazquez-Perez, Alejandro Sanchez-Flores, Angel Gustavo Salas-Lais, Alejandro Uscanga Junco, Alida Zárate, Larissa Fernandes-Matano, Luis Antonio Uribe-Noguez, Enrique Mendoza-Ramírez, Fidencio Mejía-Nepomuceno y Carlos F. Arias. Publicado en Viruses 2026, 18(6), 609.

DOI: https://doi.org/10.3390/v18060609

Artículos de difusión

"Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales" por Blanca Itzel Taboada Ramírez, Lorena Díaz-González, Oscar Alejandro Uscanga Junco, Alida Esmeralda Zárate Jiménez y Edna Cruz-Flores. Publicado en TIES, Revista De Tecnología E Innovación En Educación Superior, (15), 29–44.

DOI: https://doi.org/10.22201/dgtic.26832968e.2026.15.158

Participación en congresos

"The Viral Pangenome" por Oscar Alejandro Uscanga Junco, Lorena Díaz González, Claudia Selene Zárate Guerra y Blanca Itzelt Taboada Ramírez. En el 5º Congreso Internacional de Tecnología y Ciencia Aplicada (CITCA) llevado a cabo del 26 al 28 de noviembre del 2025 en Cuernavaca, Morelos, México.

Referencias

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bhatt, S., Holmes, E. C., & Pybus, O. G. (2011). The genomic rate of molecular adaptation of the human influenza A virus. *Molecular Biology and Evolution*, *28*(9), 2443–2451. https://doi.org/10.1093/MOLBEV/MSR044

Borges, V., Pinheiro, M., Pechirra, P., Guiomar, R., & Gomes, J. P. (2018). INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. *Genome Medicine 2018 10:1*, *10*(1), 46-. https://doi.org/10.1186/S13073-018-0555-0

Cadenas-Castrejón, E., Verleyen, J., Boukadida, C., Díaz-González, L., & Taboada, B. (2023). Evaluation of tools for taxonomic classification of viruses. *Briefings in Functional Genomics*, *22*(1), 31–41. https://doi.org/10.1093/BFGP/ELAC036

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/BIOINFORMATICS/BTS565

García-López, R. (2018). Construction of a Comprehensive Database from the Existing Viral Sequences Available from the International Nucleotide Sequence Database Collaboration. *Methods in Molecular Biology (Clifton, N.J.)*, *1838*, 231–243. https://doi.org/10.1007/978-1-4939-8682-8_16

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D. H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., … Murphy, T. D. (2025). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, *53*(D1), D243–D257. https://doi.org/10.1093/NAR/GKAE1038

Gozashti, L., & Corbett-Detig, R. (2021). Shortcomings of SARS-CoV-2 genomic metadata. *BMC Research Notes 2021 14:1*, *14*(1), 189-. https://doi.org/10.1186/S13104-021-05605-9

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/GR.210641.116

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 Biology 2004 5:2*, *5*(2), R12-. https://doi.org/10.1186/GB-2004-5-2-R12

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Lowen, A. C. (2017). Constraints, Drivers, and Implications of Influenza A Virus Reassortment. *Annual Review of Virology*, *4*(Volume 4, 2017), 105–121. https://doi.org/10.1146/ANNUREV-VIROLOGY-101416-041726/CITE/REFWORKS

Martí, J. M., Kok, C. R., Thissen, J. B., Mulakken, N. J., Avila-Herrera, A., Jaing, C. J., Allen, J. E., & Be, N. A. (2025a). Addressing the dynamic nature of reference data: a new nucleotide database for robust metagenomic classification. *MSystems*, *10*(4). https://doi.org/10.1128/MSYSTEMS.01239-24;PAGE:STRING:ARTICLE/CHAPTER

Martí, J. M., Kok, C. R., Thissen, J. B., Mulakken, N. J., Avila-Herrera, A., Jaing, C. J., Allen, J. E., & Be, N. A. (2025b). Addressing the dynamic nature of reference data: a new nucleotide database for robust metagenomic classification. *MSystems*, *10*(4). https://doi.org/10.1128/MSYSTEMS.01239-24;PAGE:STRING:ARTICLE/CHAPTER

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics 2015 16:1*, *16*(1), 236-. https://doi.org/10.1186/S12864-015-1419-2

Poon, A. F. Y. (2024). Prospects for a sequence-based taxonomy of influenza A virus subtypes. *Virus Evolution*, *10*(1). https://doi.org/10.1093/VE/VEAE064

Simmonds, P., Adriaenssens, E. M., Lefkowitz, E. J., Oksanen, H. M., Siddell, S. G., Zerbini, F. M., Alfenas-Zerbini, P., Aylward, F. O., Dempsey, D. M., Dutilh, B. E., Freitas-Astúa, J., García, M. L., Hendrickson, R. C., Hughes, H. R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., … Varsani, A. (2024). Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2024). *Archives of Virology*, *169*(11). https://doi.org/10.1007/S00705-024-06143-Y

Song, L., & Langmead, B. (2024). Centrifuger: lossless compression of microbial genomes for efficient and accurate metagenomic sequence classification. *Genome Biology 2024 25:1*, *25*(1), 106-. https://doi.org/10.1186/S13059-024-03244-4

Taboada, B., Zárate, S., Esteban Muñoz-Medina, J., Armando Vazquez-Perez, J., Sanchez-Flores, A., Gustavo Salas-Lais, A., Uscanga Junco, A., Zárate, A., Fernandes-Matano, L., Antonio Uribe-Noguez, L., Mendoza-Ramírez, E., Mejía-Nepomuceno, F., & Arias, C. F. (2026). Post-Pandemic Genomic Diversity and Lineage Turnover of Influenza Viruses in Mexico During 2022–2023. *Viruses 2026, Vol. 18, Page 609*, *18*(6), 609. https://doi.org/10.3390/V18060609

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit Y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950–13955. https://doi.org/10.1073/PNAS.0506758102;PAGEGROUP:STRING:PUBLICATION

Uscanga Junco, A., Díaz-González, L., & Taboada, B. (2025). K-FluDB: A Novel K-Mer Based Database for Enhanced Genomic Surveillance of Influenza A Viruses. *Bioinformatics Advances*. https://doi.org/10.1093/BIOADV/VBAF254

Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., & Pedersen, M. W. (2022). ngsLCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods in Ecology and Evolution*, *13*(12), 2699–2708. https://doi.org/10.1111/2041-210X.14006;REQUESTEDJOURNAL:JOURNAL:2041210X;WEBSITE:WEBSITE:BESJOURNALS;WGROUP:STRING:PUBLICATION

Wood, D. E., Lu, J., & Langmead, B. (2019a). Improved metagenomic analysis with Kraken 2. *Genome Biology 2019 20:1*, *20*(1), 257-. https://doi.org/10.1186/S13059-019-1891-0

Wood, D. E., Lu, J., & Langmead, B. (2019b). Improved metagenomic analysis with Kraken 2. *Genome Biology 2019 20:1*, *20*(1), 257-. https://doi.org/10.1186/S13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology 2014 15:3*, *15*(3), R46-. https://doi.org/10.1186/GB-2014-15-3-R46

Zhang, L., Chen, F. X., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y. J., Hao, H. X., Yi, W., Li, M., & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in Microbiology*, *12*, 766364. https://doi.org/10.3389/FMICB.2021.766364/BIBTEX

Zheng, A., Shaw, J., & Yu, Y. W. (2024). Mora: abundance aware metagenomic read re-assignment for disentangling similar strains. *BMC Bioinformatics 2024 25:1*, *25*(1), 1–18. https://doi.org/10.1186/S12859-024-05768-9

Cuernavaca, Mor., a 16 de Junio de 2026

DRA. LINA ANDREA RIVILLAS ACEVEDO
COORDINADORA DEL POSGRADO EN CIENCIAS
P R E S E N T E

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la tesis titulada **"El pangenoma viral: bases de datos de referencia mejoradas para genómica viral"** que presenta el M. en C. Oscar Alejandro Uscanga Junco (10033422) para obtener el título de Doctor en Ciencias.

Director de tesis: Dra. Blanca Itzelt Taboada Ramírez
Codirección de tesis: Dra. Lorena Díaz González.
Unidad Académica: Instituto de Investigación en Ciencias Básicas y Aplicadas (IICBA)

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Armando Hernández Mendoza CIDC – UAEM	APROBADO	
Dr. Jorge Hermosillo Valadez CInC – UAEM	APROBADO	
Dr. Ramón Antonio González García Conde CIDC – UAEM	APROBADO	
Dra. Blanca Itzelt Taboada Ramírez IBT – UNAM	APROBADO	
Dra. Sonia Dávila Ramos CIDC – UAEM	APROBADO	
Dr. José Alberto Hernández Aguilar FCAeI – UAEM	APROBADO	
Dr. Rodrigo García López IBT - UNAM	APROBADO	

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento con firma electrónica UAEM, soportada por el certificado vigente a la fecha de su elaboración y con efectos plenos de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS PUBLICADOS en el ÓRGANÓ INFORMATIVO UNIVERSITARIO "ADOLFO MENÉNDEZ SAMARÁ" número 117 de fecha 20 de abril de 2021.

Sello electrónico

ARMANDO HERNANDEZ MENDOZA | Fecha:2026-06-17 11:12:14 | FIRMANTE
NZhLFl5eZmCb4qHaBFKnL350wLE/Mg5S5awXN/7F+8FmmLHzPDzad+Xhw+w4Emxw8RF/0aYd9qNeyrRWn5dF5o0yFQqyggpZUfAtaqtgx+YWVBEAe3Y9LXdiuvegRWL+bPkb
SjIn58eZeYIcLygfRJKC1NEdlkn0vFwSDrhvl4degXru0qXgY3RSot1siqINVtncLzRO2XXSbcOIKTI+VUusdOmrgz6LyQzF3eydChfo4tGKZYaZw+iY34iynIACV3S1JXLd2jLY5HWlU
Qn+o0+lkdRkKu9ntatu5xHyWAgzHhrGAY0SSlVowK55GJ1l/h30RfTGulRNdD54h688nQ==

BLANCA ITZELT TABOADA RAMÍREZ | Fecha:2026-06-17 11:46:42 | FIRMANTE
aUbaAqFC1F9+pUbSUf3Ro2nWA2MkKZdcdp3+fIBVBrdX3bTxOCT5ReLtMf+Q4YzsVhL+LZAbTP3+QldN2DT+fDYSAiUDYJtYfX3ODWy5qeeCvmtemF4nQvHKCdTzgiinMn5pW
yk642tGYJeHRAPgm/7kJwVRDIQiypx+xJlQa3Q2dKIrUaFBrlnyqmw2lpo1c4qBy0GeZPqZf2ba49OPQAqyHGuq1Nm4Mi0LwQfd3DtlS/HxQudTo4m+pK72r430yR4bme+tGvERsc
svLX5PwD48yPOF59PCALRGJRMeAh90CWsnv774KHWn82UKfLTo8stPcy5rEqgdVzT2UyKPdw==

SONIA DAVILA RAMOS | Fecha:2026-06-17 12:03:51 | FIRMANTE
TfgmENYDp8wMGo17tRTciU8yHG7ZGccCLBy9/sKn/M/ePP2f9TTIKZtSgOsBRpAFtkYgq/mHc9zEVCWeK6GLj0vCxU92Evra00uBhlq4M7AyyjtC2q5Rd54w4o8AcWZMJS7p69si
ypLvMBoBtPMhnZqulhyrFc4clRFsp7xlSL66e7fNVWVZlxRR1nJl7Dixqa6Y9ZybKhZ+HcWTOTd5lTFHIBnnWWDJdA+a9cbtt7lUlrYkR6Wknt0lHTzs8RvY+xRf4y4hu44CnJ0NgJzO
NwarUVpQg5hnOdfQwcivCqLLVjg3XSpQ4jbcypeOqt3IRIyATu1f/tDrIfyc8YrEGw==

RAMON ANTONIO GONZALEZ GARCIA CONDE | Fecha:2026-06-17 12:12:35 | FIRMANTE
g0ZDhaWNbe3OuAUr+72OEOr1h0uYfp7RXE2wpKniluPpREtr+6O5Bscr747Fe/EyGcplfOYDHh2gx3t7wb3JsVy19LBPzTXlj/E9V5u9oLvNpnGkWPfYxuZu+Y3ENwv9PfT3qPKIuP
lzGLUq1joz4WdULMBbxOHmpyYeviAJo2+T9ZNDl6PXE2xsbDYRcKzbpAta5eP36Ru9KKGU5xBMMooKpIUeRjfY+Y22WdQRmunLiX6fhZG4cmk8DlLgL4/PEn4fQYRAQCZ8Wb
NqLFD4shzRfwSqByjGhDqpuiuozF4l1SL3dFS8WhIET3W7QCF3A0s8dCsLETD14Y8XctMjdQ==

JORGE HERMOSILLO VALADEZ | Fecha:2026-06-17 12:40:42 | FIRMANTE
ceLem2j69vKpy4kCHFkdHtizZblpmetPqtfpl3uJPhWpeQvehHgQY0TRexvsnXe9YKsnjyIQ/JM5HFEoZhxvhh2+H2dlmVKlVCPPOd7Ym0NMqh8uwYyeuOd6CoLt4z8uRUhNPOrb6
N+r67tyB6cVr87HI0jQSI4II90pLjXeQ49kVcIMQ5JZ8AOMHGM1/F0/ck/YS1E7hBNV9XqoY88V2ElhXtTfXXinHk8j2Apj7bRPR2UNoZT4zjjECE4OzWChvJIQXhe5vkZhpUwuWHZ
ZRivbm1E7k14f6g1hMFWMpzcDPoN03rT8rFHd0OzN5BRcAuTs8Fat42RgZhqFH3IGKg==

JOSE ALBERTO HERNANDEZ AGUILAR | Fecha:2026-06-17 13:16:58 | FIRMANTE
CQHhOW5fj78naUVMgzWc6SvwNqEVtQFpMkfa4yN8Kyu9GHu4693DjzUw0IaFBhwHrrsiUNkezUVOCo9b2A9xXN8kl6ggENunI6s/8IZXLAcxkRzlm/HPOBf3BLDcCUCPQvWzuA
3RKU9aYc9HWdLbtWUQnB16y2U/YJIAV8ShgdBn3aH6WZ8zGEPAc1yCN+LFesB3UmL/8GmT3Y38xENV/g9EGA1dyh2Ej76tX19mdor7qJb5rjZ/qlaO7D/l766aVlyggdAHItJYue
HSeGIiJO2uqFtT2oAGY5qyUK2xb/gAgkikavpT03itMsZ5QX3ag6SbNIIKSbFCynO5CBAk8w==

RODRIGO GARCÍA LÓPEZ | Fecha:2026-06-17 13:37:53 | FIRMANTE
cwsW339CoGEW1WCNtYcTwTZNTkXsopFyybHknQ6htMzPp/Mm6zlrICU49GKWKKpn86DUspYYUzcTbUkf8MJrnftV2dySQ85j/9IRiZPhTjM+FbNP+X8sW9LEBdgpRGIQEcHW
1E7TcVo8ztjEq1bMRwok66INhp32DtNOa3WaG1n5x+zhmu1SalXLZyMEKkq+EMQDNRINe6Vpa6AY9Repe/6szJW0FDzkYmuS2eTx/lween/+WVoDr4nd0w455lIaBTnB69THkU
1rxdKWwiH6gz+PAyubM5HEoLZ8/5ydlqVuU/bA92eFrg7TGhyvwgQkSYoJNUo55mIV8raGad+2iA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

WyPCbVXQD

https://efirma.uaem.mx/noRepudio/YDqoucghUqBfPmUeUlvklV7XnB4Gxyz0