Protocolo computacional para la asignación taxonómica de virus en metadatos genómicos.

Valentina Cobo Paz

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C, Colombia
Año 2020

Protocolo computacional para la asignación taxonómica de virus en metadatos genómicos.

Valentina Cobo Paz

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Bioinformática

Director(a):Ph.D. Clara Isabel Bermúdez Santana.
Codirector(a): Ph.D Jose Aldemar Usme Ciro.

Línea de Investigación:
Tecnologías computacionales en Bioinformática.
Grupo de Investigación:
Rnomica Teórica y computacional

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C, Colombia
Año 2020

"Of living things, my son, some are made friends with fire, and some with water, some with air, and some with earth, and some with two or three of these, and some with all."
Hermes Trismegistus

Agradecimientos

A mis padres, por su apoyo constante y por haberme dado el privilegio de disfrutar de la educación pública por 20 años, que me dejo experiencias y personas maravillosas. A mis hermanos Natalia y Luis Miguel por su cariño y apoyo incondicional.

A Jose Alejandro por sus sugerencias siempre tan precisas, por su apoyo en los conceptos matemáticos y por su compañía en los momentos más difíciles.

A la profesora Clara Bermúdez por adentrarme en el mundo de la bioinformática, guiarme por mas de 5 años con paciencia y esfuerzo.

Al profesor Jose Aldemar Usme Ciro por su contribución en los aspectos virológicos de este trabajo.

Al señor Ernesto Parra, por su apoyo incondicional en la administración del laboratorio de Biología Computacional.

A mis compañeros y amigos del grupo de investigación Rnomica teórica y computacional.

A Colciencias por el financiamiento de este trabajo, que se desarrollo bajo el marco del proyecto "Expedición Virológica en Ecosistemas Representativos de Colombia: Selva Húmeda Tropical de la Sierra Nevada de santa Marta"(No. 201010029276).

A los esfuerzos económicos del Sistema de Intercambio Alemán DAAD y a la facultad de Ciencias por facilitar el funcionamiento del Laboratorio de Biología Computacional de la Facultad de Ciencias de la Universidad Nacional de Colombia en donde se realizó el proceso computacional.

Resumen

Los virus están ampliamente distribuidos en todos los ecosistemas naturales y son el grupo de entidades biológicas más diverso conocido. Aunque su biodiversidad biológica estimada es de 10^{31} órdenes de magnitud, nuestro conocimiento es menor al 1 %. Además, debido a su capacidad de impacto a la salud humana, como lo ha sido la reciente pandemia de Sars-cov-2, es esencial la búsqueda de estrategias que sean rápidas y fiables al clasificar nuevos virus usando los datos disponibles como referencia de manera eficiente. Nuestro objetivo es encontrar métodos flexibles para filtrar y clasificar secuencias víricas utilizando diversos recursos como el aprendizaje de máquina principalmente con una resolución adecuada, una alta eficiencia y buena precisión, manteniendo la flexibilidad del modelo a secuencias víricas diversas. Seleccionamos las máquinas de soporte vectorial y los árboles de gradiente potenciado como los métodos que más nos favorecían en términos de recursos computacionales, rendimiento y predicción, los datos usados fueron descargados del NCBI Virus para entrenar los modelos. Las secuencias virales fueron filtradas cuidadosamente para el entrenamiento del modelo. Después del filtrado de los datos, 19 familias tuvieron el número de secuencias más representativas. Finalmente, de este conjunto de datos, 80 % fueron usados para entrenar las máquinas de aprendizaje y 20 % fue utilizado para validar las clases taxonómicas. Las secuencias víricas se transformaron a una representación numérica a través de el método *count vectorizer* en *k-mers* de diferentes tamaños, incluyendo 3k-mers con el fin de preservar la información de los marcos abiertos de lectura (*ORF's*) y evitar el sobreajuste. En este trabajo, nuestros métodos permiten encontrar asociaciones a nivel taxonómico de familia entre las secuencias virales y la taxonomía, por medio de recursos computacionales eficientes de predicción y a diferencia de métodos convencionales de comparación de secuencias. Sin embargo, es importante señalar que en el aprendizaje de máquina la calidad de la predicción recae directamente en la calidad de la base de datos de entrenamiento y la definición de la clase, por lo tanto descripciones débiles de las familias de virus son la mayor limitación para construir un modelo coherente de clasificación de secuencias. Finalmente, el modelo de árboles de gradiente potenciado tiene la mejor probabilidad de predicción, encontramos que 8 familias que fueron predichas para los datos experimentales concuerdan con los reportes científicos para *Culex sp.* y *Aedes sp.*

Palabras clave: metaviromica, ensamblaje, aprendizaje de máquina, árboles de gradiente potenciado, maquinas de soporte vectorial.

Computational methodology for taxonomic characterization of virus in genomic metadata

Abstract

Viruses are widely distributed in all the natural ecosystems and belong to one of the most diverse groups of biological entities. Though their estimated biodiversity is 10^{31} orders of magnitude, our current knowledge is still less than 1 %. Besides, due to the capacity to impact human health dramatically, as it has been seen in outbreaks like the current pandemic, it is essential to search for strategies that fast and reliable classify new viruses by using the available data efficiently as reference.

Then, our goal is to search for flexible methods to filter and classify viral sequences from diverse sources using machine learning (ML) principles with a proper resolution, high efficiency, and accuracy, but with flexibility. We have chosen support vector machine and gradient boosting as ML method that are more favorable in terms of computational resources, performance and prediction and the data used was downloaded from the viral NCBI database to train our approach.

Viral sequences from the databases were carefully filtered to train the model. After the filtering of the data, 19 families had more representative number of sequences. Finally, from this set of data, 80 % was used to train the machine, and 20 % was used to validate the taxonomic assignment. Viral sequences was change to numeric representation throught count vectorizer method into k-mers of varied sizes include 3 k-mers to preserve open reading frames (ORF's) information and avoid overfitting.

In this approach, our method allowed to find associations in family taxonomic level between the viral sequences and the viral taxonomy by using inference computational resources efficiently and unlike other conventional methods for sequence comparison. Nevertheless, it is essential to point out that ML approaches rely directly on the quality of the input dataset, and the class definition so weak description of some families of viruses are the major limitation to construct a coherent model to classify their sequences.

Finally, the gradient boosting model have the highest prediction probability, we found 8 families predicted in the experimental data that agree with the scientific reports in different studies for Culex sp and Aedes sp.

Keywords: Metaviromics, assembly, machine learning, boosting trees, support vector machine)

Esta tesis de maestría se sustentó el 03 de Diciembre de 2021 a las 09:00 am,
y fue evaluada por los siguientes jurados:

Andres Mauricio Pinzon Velasco (Msc - Ph.D.)
Instituto de Genética
Universidad Nacional de Colombia

Juan David Garcia Arteaga (Ph.D.)
Facultad de Medicina
Universidad Nacional de Colombia

Contenido

Introducción

En la presente tesis se atiende la necesidad de desarrollar un protocolo de referencia para la clasificación de información vírica encontradas en el secuenciamiento de muestras de comunidades biológicas complejas, lo cual es conocido como metavirómica, como es el caso de los vectores de diversos tipos de RNA que se encuentras alojados en diferentes artrópodos. El trabajo presenta el análisis de datos derivados del secuenciamiento de nueva generación que fue validado usando bases de datos biológicas de referencia en virología con el fin de alcanzar la clasificación de metadatos en categorías taxonómicas a la mejor resolución posible. Se realizó la validación del protocolo usando el metaviroma de *Drosophila suzukii* [1] y los datos metavirómicos del proyecto financiado por Colciencias "Expedición virológica en ecosistemas representativos de Colombia: selva húmeda tropical de la Sierra Nevada de Santa Marta". Se automatizaron diferentes etapas del análisis metavirómico, el pre-procesamiento de productos de secuenciamiento, el filtrado de contaminación de procedencia biológica y el uso de máquinas de aprendizaje que son capaces de clasificar secuencias de virus al mejor nivel de resolución y finalmente proponer metadatos candidatos a fuentes posibles de virus emergentes que aún no han podido ser categorizados.

Esta propuesta se presenta como apoyo metodológico complementario a los estudios tradicionales en estudios de genomas virales y de su detección, donde se implementa para el diagnóstico, principalmente métodos cualitativos, como la observación de cambios fenotípicos de cultivos y métodos cuantitativos como la cuantificación de las partículas virales existentes en una muestra o la serotipificación. Debido a que muchos experimentos son diseñados para un serotipo en particular, un tipo de virus o un tipo de célula huésped usando este tipo de modelos de caracterización convencionales, existe un contraste diferente a estudios que tratan con casos en donde coexiste mas de un tipo de virus como por ejemplo en los vectores. En estos estudios la complejidad en la caracterización se incrementa y la posibilidad de existencia de diversidad de virus coexistentes, en su mayoría posiblemente desconocidos o no categorizados, dificultan la posibilidad de aislar e identificar la totalidad de virus en la muestra.

Es por ello que hoy en día el uso de nuevas aproximaciones tecnológicas como el análisis de metainformación, producto del secuenciamiento masivo de la información genómica y transcriptómica en muestras complejas y diversas, hace que desde el punto de vista computacional existan algoritmos que permitan la clasificación de secuencias, sin embargo presentan limita-

ciones en muestras que divergen de los virus de referencia contenidos en las bases de datos. Es por ello que en este trabajo se utilizaron diferentes métodos de clasificación de secuencias usando diferentes algoritmos usados en el aprendizaje de máquinas como árboles de decisión y máquinas de soporte vectorial que fueron entrenados con las bases de datos de virus de referencia para estudios de clasificación viral.

Finalmente, en este trabajo se implementó diferentes métodos de clasificación basados en aprendizaje y se evaluó su desempeño y eficiencia en la clasificación de secuencias víricas derivadas de la información metavirómica en categorías taxonómicas, los cuales se validaron con la caracterización del viroma de *Drosophila suzukii* por Medd *at al.* [1] y se emplearon para la caracterización de muestras provenientes de diferentes morfotipos de la familia Culicidae, muestreados en la Sierra Nevada de Santa Marta, lo que permitió conocer cual método de acuerdo a su precisión y exhaustividad puede aportar con mayor confianza información sobre que familias se encuentran presentes en estos dos tipos de muestra.

1. Capítulo 1: Cacería de virus en un contexto de emergencia viral.

1.1. Virus: aspectos biológicos de interés

Los virus desde su descubrimiento han generado curiosidad en el entorno académico; en el año 1892 Dimitry Ivanovsky publicó los resultados de los estudios que había hecho en relación a la enfermedad de mosaico en las plantas de tabaco, esta generaba grandes pérdidas económicas a la industria y que por ende se invertía gran cantidad de recursos en su investigación. En su artículo Ivanovsky describía: *"la savia de las hojas infectadas por la enfermedad del mosaico mantenían sus propiedades infecciosas incluso después de la filtración a través del filtro Chamberland"* [2], (instrumento conocido por separar bacterias del agua). Este descubrimiento abrió las puertas a algo completamente desconocido y del cual en el presente todavía sabemos muy poco. De los virus actualmente se conoce que son agentes infecciosos no vivos que tienen la capacidad de propagarse en una célula hospedera determinada. Cada virus tiene un genoma codificado en moléculas de RNA o DNA con una cadena simple o doble, positiva ó negativa [3] que compactan de manera extraordinaria la información para una infección exitosa en un tamaño de genoma muy reducido, el cual se encuentra entre 3400 nucleótidos (nt) (Enterobacteria fago BZ13) a 31000 nt (Coronavirus), en donde contiene las proteínas esenciales para la envoltura protéica y se conoce que el 80 % del genoma viral codifica para genes específicos del virus, de los cuales la mayoría no tiene homólogos conocidos o su función es desconocida [4].

De acuerdo a estudios realizados en virus marinos en 2007, se estima que los virus son las entidades biológicas más diversas conocidas, debido a que han conquistado todos los ecosistemas biológicos [5]. Se estima que su diversidad puede ser del órden de 10^{30}, pese a esto, nuestro conocimiento respecto a la diversidad vírica es mucho menor al 1 % [5]. De acuerdo a los registros actuales de la base de datos del `NCBI` (*National Center of Biotechnology Information* de los Estados Unidos `Genbank`) solamente el $0,036$ % de la información (en bases nucleotídicas) pertenecen a virus (Tabla **1-1**).

Por otro lado, Wolf *et al.* reportan en el 2018 que la mayor diversidad de virus con hospedero Eucariota tiene un genoma tipo RNA, en donde prevalecen los patógenos en humanos, animales y plantas [7]. Como también se menciona, que el incremento de la ganadería ex-

proporción de datos RefSeq del NCBI			
Categoría	información viral	información total	Porcentaje de información viral (%)
Nucleótidos	12,179	33,037,650	0,036
Proteínas	373,129	157,639,958	0,236
Especies	9,350	97,407	9,598

Tabla 1-1.: Información registrada en el NCBI que pertenece a Virus, comparado con la información completa almacenada en esta base de datos. Tomado de: NCBI-statistics.

tensiva, el aumento de la tasa de pérdida de hábitat y el aumento de la migración humana brindan un ambiente favorable para la actual emergencia pandémica debido a los eventos de cambio de hospedero [8], como es el caso del ébola, el síndrome respiratorio de el medio oriente (MERS por sus siglas en inglés *Middle East Respiratory Syndrome*), la influenza A(H1N1) [9] y la actual pandemia causada por Sars-CoV-2. La influencia del hombre y su acceso a reservorio naturales de los virus genera preocupación para la salud pública no solo nacional sino global, en particular en los trópicos donde el ambiente húmedo y selvático, la temperatura constante y la población de vectores en equilibrio brinda un ambiente óptimo para la dispersión y oportunidad cuando estos ambientes son perturbados [10].

1.2. Arbovirus como grupo de estudio

Los Arbovirus son una agrupación artificial nombrada así por sus siglas en inglés ***Ar****thopod* ***Bo****rne* ***virus***, donde han clasificado virus que usan a los Artrópodos como vectores para infectar animales (incluidos los humanos) y plantas [11]. La conexión entre Artrópodos y enfermedad se descubrió cuando el médico Cubano Carlos Finlay propuso en 1881 que la fiebre amarilla era transmitida por mosquitos, en 1901 el Mayor Walter Reed de Estados Unidos lideró la investigación que comprobaba que los mosquitos son los vectores de esta enfermedad infecciosa. Dicha conclusión resulta extraordinaria para la época que fue planteada antes de conocer la existencia de los virus. Este evento es importante porque marcó el inicio de la investigación en epidemiología y biomedicina, disciplinas que actualmente juegan un papel principal en temas de salud pública y que se acentúa en la actual emergencia viral [12].

Como consecuencia de la globalización producto de la esclavitud en los siglos XV y XVI,*Aedes aegypti* fue el primer vector que se propagó globalmente y a mediados del siglo XX el virus del Dengue (DENV) alcanza una propagación global debido a la urbanización, el crecimiento de la población, el incremento de los viajes internacionales y el cambio climático, re-emergiendo como una enfermedad mundial [12].

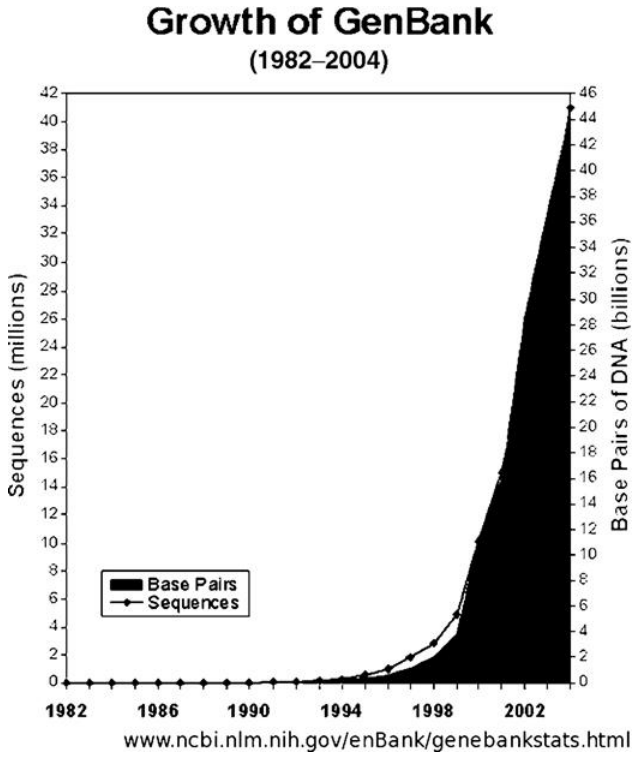


Figura 1-1.: Crecimiento de la base de datos del genBank [6]

Actualmente, se sabe que la categorización de Arbovirus es una agrupación que no está basada en su filogenia, en cambio esta basada en el tipo de transmisión. Esta está compuesta por 7 familias (tabla **1-2**) donde, exceptuando una especie con doble cadena de ADN de la familia *Asfarviridae*, todos sus miembros son exclusivamente virus con genoma tipo RNA, esto podría vincularse con la alta plasticidad en su estructura genómica, que algunos autores relacionan con el éxito infectivo y con la dinámica de las interacciones que tienen los virus con sus hospederos [13]. Para el año 2006, se conocían aproximadamente 100 especies de Arbovirus con capacidad de infectar y causar patologías en el humano y para el año 1988, donde se realizó la primera revisión exhaustiva del phylum *Arthropoda* se encontró que existen 14.000 especies de insectos hematofagos con capacidad de transmisión de este grupo [14].

Los agentes infecciosos y los hospederos han evolucionado conjuntamente. Esto, junto con el crecimiento poblacional humano, ha estado íntimamente relacionado con la dispersión de los artrópodos y, por consecuencia, con la dispersión de Arbovirus. actualmente se ha avanzado mucho en el estudio y conocimiento de este grupo de virus y por ende hace que sean de los más conocidos e investigados por su capacidad de infección y las patologías que se asocian a estos [23]. De los Arbovirus que más han impactado al humano están los virus Dengue, Zika (ZIKV), la fiebre amarilla y el virus del Nilo Occidental (west nile virus) (WN) todos perte-

grupo	tipo de genoma	Taxonomía	tamaño del genoma
Flaviviridae	ssRNA(+) †	4g 100sp	9.0-13 kb [15]
Togaviridae	ssRNA(+) †	1g 31sp	10–12 kb [16]
Bunyaviridae	ssRNA(-) §	9g 330sp	10.5-22.7 kb [17]
Rhabdoviridae	ssRNA(-) † §	18g 130sp	11-15 kb [18]
Reoviridae	dsRNA §	30g 87sp	18.2-30.5 kb [19]
Orthomyxoviridae	ssRNA(-) §	1g 1sp	∼10kb [20]
Asfarviridae	ssDNA †	1g 1sp	170-194 kb [21]

Tabla 1-2.: Familias de virus pertenecientes a los Arbovirus [22], exceptuando *Bunyaviridae* que es un orden. Los simbolos † y § representan la no segmentación del genoma y la segmentación del genoma respectivamente, en la columna de Taxonomía se representa genero (g) y especie (sp).

necientes a la familia *Flaviviridae* han sido asociados a diferentes alteraciones neurológicas como el síndrome de Guillain-Barre, el síndrome de zika congénito y la consecuente microcefalia en neonatos [9].

Dado el amplio conocimiento del grupo, se ha podido reconstruir la relación filogenética de los virus RNA que comparado con sus hospederos, se observa incluso la existencia de transferencias de información genómica entre diferentes hospederos con relaciones muy distantes (entre plantas y animales), en donde los virus actuan como vectores [24], [7], lo cual desde una perspectiva de salud pública sería útil estudiar para encontrar posibles colonizaciones víricas en humanos.

1.3. Virus y las nuevas tecnologías de secuenciamiento

En 1977, el virus bacteriofago MS2 fue la primera entidad biológica secuenciada. A mediados de 1990, se secuenciaron virus de interés por su patogenicidad, como el VIH (Virus de Inmunodeficiencia Humana), grupo para el cual se realizaron los primeros estudios de genómica comparativa en familias de virus [4]. Por las dinámicas de construcción de conocimiento humanas, los virus fueron siendo agrupados y descritos conforme al conocimiento que se tenía sobre ellos y mediante características subjetivas como sucede con el grupo Arbovirus.

Por otro lado, el desarrollo de las tecnologías de secuenciación de próxima generación *(Next-generation sequencing NGS)* ha constituido una revolución en la forma en la que se concebía la construcción de conocimiento y la investigación en la biología, la medicina o la química, ya que ha permitido romper paradigmas o la comprobación de otros, porque nos ha ayudado a descifrar el detalle de la información que esconde el genoma de los virus y de muchos grupos

de especies. Actualmente, la disminución de los costos y el aumento en el desempeño de las máquinas, ha permitido en el área de la virología la rápida identificación y caracterización de diferentes tipos de virus patógenos, permitiendo trazar la evolución de algunos grupos de virus, así como también rastrear cual fue su cadena de transmisión e incluso de dispersión [25]. Por ejemplo, el secuenciamiento por medio de NGS durante los primeros estadios de un brote, permitiría seguir el rastro de donde, cuando y como surgió la epidemia, ayudando al desarrollo de medidas de mitigación y prevención, especialmente de aquellos lugares que pueden estar relacionados con ser focos de enfermedades epidémicas emergentes con etiologías desconocidas aún y por ende que se encuentran sin diagnóstico ni tratamientos efectivos [25].

Con la creciente masificación de los métodos de secuenciamiento, surgen nuevas técnicas y con ellas nuevos desafíos y nuevos campos de estudio, como lo es la metagenómica, la cual se puede definir según la literatura como un conjunto de metainformación que contiene una mezcla de genomas presentes en una muestra que puede ser secuenciada por NGS y analizada por métodos computacionales. En otras palabras puede ser llamada como "*El genoma colectivo de una comunidad específica que fue aislada*" [26], y para el caso de una muestra vírica se denomina como metavirómica. Los análisis de las muestras metavirómicas pueden ser complejos y suponen un reto para la biología computacional que va desde el manejo de la información disponible en las bases de datos y el análisis riguroso de grandes volúmenes de información, hasta la extracción de información de calidad, lo que difiere de los análisis de muestras individuales también ya que la profundidad de las lecturas no es constante, debido a la mezcla misma de los genomas presentes, por lo que no se obtiene la misma información homogénea como si se obtiene cuando se secuencia el genoma de un único organismo, como ocurre en un secuenciamiento genómico individual [9].

En el análisis de datos metavirómicos asociados a vectores artrópodos se han realizado recientemente estudios cada vez más minuciosos y con la posibilidad de obtener un mayor volumen de información. En el año 2018 se realizaron tres publicaciones que se destacan porque se hicieron identificaciones de una gran cantidad de muestras virales, como es el caso del grupo de Atoni *et al.* 2018 [27] . Este grupo realizó un análisis de la diversidad de virus asociados a dos especies diferentes de *Culex sp.* en Kenia y China, o el caso de Xiao *et al.* 2018 [28] quienes realizaron un estudio metavirómico en cerca de 8000 mosquitos, encontrando el serotipo 4 del DENV, el virus de la encefalitis Japonesa (JEV), entre otros. Lo importante de este trabajo es que fueron identificados mediante técnicas bioinformáticas y posteriormente confirmados con técnicas de biología molecular. Así mismo, otros trabajos como los realizados por Sadeghi *et al.* 2018 [29] en 31 municipios de California en Estados Unidos, secuenciaron el metaviróma de *Culex sp.*, ensamblaron los genomas completos de 56 virus y descubrieron 32 virus nuevos. Estos estudios muestran nuevas formas de abordar los grandes datos, mediante la obtención de conocimiento a partir de las relaciones de los organismos que constituyen la muestra, donde se pueden hacer inferencias sobre las dinámicas

poblacionales, relaciones con otros phyla, lo cual contribuiría a un mejor monitoreo de la transmisión de virus como se sugiere en la literatura.

1.4. El genoma viral visto como información

Desde una perspectiva computacional, en muchos estudios se reporta que el análisis de secuencias metavirómicas es uno de los retos algorítmicos mas demandantes en cuanto a la relación desempeño *vs.* costo computacional, tiempo de ejecución, precisión y eficiencia de los resultados. Adicional a esto, los virus son las entidades con mayor abundancia en la biosfera (con una estimación de $1{,}2 \cdot 10^{30}$ en océano abierto y $0{,}25 - 2{,}5 \cdot 10^{31}$ en superficies terrestres) [30]. Existe una alta correlación entre los metadatos, presumiblemente debido a la homología de secuencia que puede derivar de especies virales "cercanas", lo cual viola el supuesto de independencia que siguen varios modelos estadísticos aplicados al estudio genómico[31], así como también en principios básicos del uso de BLAST o de los algoritmos de mapeo a datos de referencia. En general, muchos métodos estándar de la bioinformática no son fácilmente aplicables en el análisis de metagenomas virales ya que tienden a subestimar la existencia de datos nuevos, lo cual es uno de los mayores limitantes si se considera que la mayoría de las secuencias virales no tienen una similaridad significativa con nada conocido, y en donde de una muestra metavirómica puede encontrarse del 60 % al 90 % de secuencias no conocidas [32].

Dado que el análisis metavirómico puede ser más complejo que el mismo análisis virómico, se hace necesario el desarrollo de enfoques diferentes a los tradicionales usados con otros datos genómicos. Debido a que es un área de estudio relativamente nueva, el desarrollo de herramientas para su análisis lleva poco tiempo en desarrollo y que continua siendo un problema abierto computacional y algorítmico en cuanto a optimización [33].

Como es el caso del ensamblaje de secuencias, que siguen principios básicos del análisis genómico convencional, en donde existen dos posibles metodologías, el ensamblaje por referencia, en donde los fragmentos producto del secuenciamiento son mapeados a un genoma que ha sido ensamblado previamente y ha sido sometido a una curación de la información, que se conoce como genoma de referencia, proceso que tiene unos costos computacionales altos y unas limitaciones teóricas dependiendo de la divergencia de la muestra [34]. Segundo, el ensamblaje *de novo*, en donde se construye una aproximación al genoma por medio de los fragmentos secuenciados sin una referencia, esto se logra por medio de algoritmos basados en teoría de grafos (que son estructuras matemáticas usadas para la construcción de relaciones entre objetos). biyeccion una palabra se construye con solo dos fragmentos del conjunto total de fragmentos. Sin embargo, muchos modelos siguen principios básicos del análisis genómico convencional en un inicio. Por ejemplo, en cuanto al ensamblaje, existen dos metodologías

posibles: 1) es el ensamblaje por referencia, en donde los fragmentos producto del secuenciamiento son mapeados a un genoma que ha sido ensamblado previamente y ha sido sometido a una curación de la información, que se conoce como genoma de referencia, proceso que tiene unos costos computacionales altos y unas limitaciones teóricas dependiendo de la divergencia de la muestra [34], y 2) es el ensamblaje *de novo* [35], donde por medio de algoritmos basados en teoría de grafos -que son estructuras matemáticas usadas para la construcción de relaciones entre objetos, las cuales se representan por medio de aristas y nodos- se reconstruyen una aproximación al genoma real, los grafos de *De Bruijn* es uno de los algoritmos más utilizados para este fin.

Este tipo de grafos permite la reconstrucción de una secuencia, a partir de la búsqueda de relaciones entre sus fragmentos (o *reads* producto del secuenciamiento) por medio de su sobrelapamiento. Para construir un grafo de *De Bruijn* de orden n y k letras ($B(k, n)$) sobre un alfabeto $\mathcal{A} = \{A, C, G, T\}$, se ubican todas las k^n palabras del alfabeto $\mathcal{A}$ de longitud n sobre los vértices del grafo. Sean x_i y x_j dos vértices del grafo, se dice que el grafo presenta un arco de x_i a x_j cuando la palabra obtenida al suprimir la primera letra de la palabra en el vértice x_i es la misma palabra obtenida al eliminar la ultima letra en el vértice x_j. Cuando hay un arco presente en el grafo hay un sobrelapamiento entre dos palabras de la misma longitud. De esta forma, este tipo de grafos nos permite reconstruir el genoma por medio de múltiples copias de los fragmentos de un genoma, pues al construir el grafo de *De Bruijn* asociado a tales fragmentos, es posible estudiar todas las combinaciones de palabras de una longitud dada y así reconstruir lo que se puede considerar la mejor aproximación al genoma real [36].

En términos generales, se requiere extraer a partir de un conjunto de datos provenientes de una muestra metavirómica, un análisis computable y obtener unos resultados coherentes y con la menor pérdida de información. Este último es uno de los retos actuales de la bioinformática, ya que las muestras genómicas colectivas, debido a las limitaciones de los métodos de secuenciamiento, contienen lecturas con diferentes profundidades de secuenciación y una alta tasa de contaminación [26], lo que ocasiona dificultades en la recuperación de la información de los organismos de forma individual. Por otro lado, la mayoría de ensambladores *de novo* no han sido desarrollados para este tipo de datos, ni las complejidades que representan, como por ejemplo que no exploran la reconstrucción de los genomas con una baja cobertura de secuenciamiento. Por esto, y entre otras razones, se requiere de algoritmos específicos para este tipo de datos, tanto para el ensamblaje, como para la clasificación de virus conocidos y referenciados, como también para el descubrimiento de nuevos virus [37].

Adicional a esto, dado que el análisis de datos metagenómicos de muestras virales es un área nueva de estudio, el desarrollo de software de análisis de este tipo de datos no han sido bien estudiados para el análisis de genomas virales, en el caso de algunas de las herramien-

tas generales que pueden ser empleadas en este tipo de datos, están limitadas a casos muy particulares [25] y no consideran, un largo número de secuencias contaminantes, lecturas que se alinean con el mismo puntaje en diferentes genomas [38] e información insuficiente y desbalanceada en las bases de datos virales [5].

2. Capítulo 2: Pre-procesamiento y análisis de información metavirómica.

2.1. Introducción

El pre-procesamiento de los datos crudos o producto del NGS se considera una parte fundamental para el análisis de datos y es indispensable el uso de datos de alta calidad para los análisis posteriores que se vayan a realizar y para la obtención de resultados con un nivel de confianza alto. Además, se debe realizar una correcta curaduría de las bases de datos que se emplean como referencia, en donde también se debe hacer el filtrado minucioso de los datos, pues se debe tener en cuenta el margen de error producto del secuenciamiento de nueva generación y los errores producto de las depuraciones de datos de referencia.

En el proceso de transformar los datos crudos en conocimiento, se extraen patrones e información útil de los datos, por medio de métodos, algoritmos y flujos de trabajo que permiten que esa información se convierta en conocimiento relevante, de acuerdo a nuestras necesidades [39]. El preprocesamiento de los datos, es la parte más importante de este flujo de trabajo, ya que determina la calidad de los resultados que se obtendrán, por lo cual demanda la mayor cantidad de tiempo del proyecto, errores en la curaduría de los datos puede verse reflejado en valores fuera de rango, relaciones entre los datos sin sentido o pérdida de información [40].

En cuanto a el análisis de datos, en especial de secuencias víricas, el pre-procesamiento de los datos cumple una función fundamental para lograr una identificación eficiente y precisa de secuencias víricas en una muestra metavirómica, lo que permite corriente abajo, el ensamblaje fragmentos con una mayor cobertura del genoma, mejor calidad y con una anotación más confiable [34]. Sin embargo, estos tipos de análisis también dependen de la etapa experimental, que en particular genera limitaciones como por ejemplo las limitantes derivadas del proceso metodológico experimental que, según su rigurosidad, influirá sobre el proceso computacional. Este punto se menciona ya que la pureza y concentración viral en una muestra dependen de múltiples factores, como el origen de la muestra, la complejidad de la muestra o incluso la técnica usada de extracción del material genómico viral. Muchos problemas de estas etapas se convierten en causas de las posibles fallas en la perdida de información. Sin embargo, el uso de métodos computacionales son determinantes para poder dar valor a la calidad de la información.

Las muestras metagenómicas, debido a los métodos que se usan para la recolección de la muestra, tienen un mayor porcentaje de contaminación que una muestra regular. En el caso de las muestras metavirómicas el mayor porcentaje de información genómica pertenece al hospedero, teniendo un contenido significativo de otras comunidades microbiales, hongos o eucariotas [41], pero en otros casos en donde la carga viral tiene una alta diversidad de individuos en comparación con el material genómico del hospedero y del contenido bacteriano en la muestra, debe procesarse esta como una muestra de alta complejidad, para la cual el aislado del material virómico depende en gran parte de la disponibilidad de un genoma de referencia y de una base de datos de contaminación (RNA ribosomal, Bacterias, hongos) curada [42].

Una muestra metavirómica puede contener hasta un 90 % de contaminación, por lo que es fundamental una metodología que incorpore un filtrado estricto de información externa y que cuando esta sea de tipo meta-transcriptómico se puedan evaluar métodos estandarizados para este tipo de datos. Por ejemplo, RNA-QC-Chain ha sido uno de los programas que han mostrado eficiencia tanto en el desempeño computacional, como en las múltiples opciones de filtrado de calidad y contaminación de las secuencias, detectando secuencias con baja calidad, duplicaciones, adaptadores y secuencias RNAr discriminados por subunidades cortas de eucariotas, bacterias y arqueas y subunidades largas de bacterias y eucariotas, identificados por medio de modelos ocultos de Markov [43] que favorece la detección de secuencias *de novo*, lo que lo hace una de las herramientas mas completas para el filtrado de secuencias [44].

Aunque la contaminación por rRNA constituye una gran proporción del contenido de la muestra, es necesario filtrar también la contaminación genómica proveniente del hospedero, de hongos, bacterias y archaeas, que representan una porción considerable de los datos, donde su comparación tiene un alto costo computacional, debido al tamaño de las bases de referencia por un lado y al volumen de los datos de NGS por otro.

Dada la complejidad de la muestra y las características que la definen, como la baja profundidad del genoma viral presente y a la alta diversidad de la muestra metavirómica, los algoritmos de ensamblaje desarrollados para muestras genómicas no son útiles para tipo de muestras metagenómicas, ya que tienen una cobertura dispersa y mucho mas baja que una muestra genómica [42], [45]. En donde una cobertura hace referencia a el número de lecturas promedio que están representando una región del genoma, esta medida es usada como grado de confianza en el descubrimiento de variantes [46].

Estas peculiaridades hacen necesario el uso de algoritmos que sean desarrollados para resolver los conflictos que puedan ocurrir durante el ensamblaje debido a la baja representación de las secuencias, lo que conlleva también a que las métricas estadísticas deban ser más flexibles y se consideren con un peso diferente como se reporta en la mayoría de la literatura.

El volumen de datos y los requerimientos necesarios para el computo de análisis de alta complejidad como las mencionadas en muestras metagenómicas siguen siendo una limitante para la investigación en esta área. Esto fue una de las principales motivaciones para el desarrollo de `Megahit` [47], para esto se enfocaron en satisfacer estas necesidades, y `Megahit` fue el primer algoritmo de ensamblaje desarrollado para muestras metagenómicas que podía ensamblar datos de cientos de Gigabytes, incrementando el desempeño del procesamiento con CPU para no depender de unidades de procesamiento gráfico (GPU) con un aumento en el desempeño de 30 %, sin embargo, también se tiene la implementación del algoritmo para GPU la cual aumenta su desempeño en un 50 %, favoreciendo su uso en servidores con diferentes capacidades computacionales.

2.2. Datos y métodos

2.2.1. Datos de Drosophila suzukii

Se emplearon los datos del artículo *"The Virome of Drosophila suzukii, an invasive pest of soft fruit"* [1], como datos de referencia para evaluación de la metodología de ensamblaje. En este trabajo desde un estudio metatranscriptomico se identifican virus que infectan a *Drosophila suzukii* en sus dos rangos de distribución naturales, nativo en Japón e Invasivo en Inglaterra y Francia, donde se describen 8 nuevos virus RNA, de 8 familias y se discuten las relaciones filogenéticas con virus ya conocidos. Los datos crudos de este estudio fueron muestreados en 5 lugares diferentes y fueron secuenciados usando los kits de preparación de librerías de Illumina NGS y la plataforma de Illumina Hi Seq con fragmentos de 120 a 150 nt de lecturas pareadas. Esta información se encuentran disponible en el `NCBI` y los tamaños de los datos usados en este trabajo se presentan en la tabla **2-1**.
Estos datos fueron escogidos como datos de referencia, para el control y verificación del procesamiento, por tres razones principales, primero, en esta investigación se estudian virus con un genoma RNA, segundo, es un estudio meta transcriptómico en un hospedero del phylum Arthropoda realizado en varios lugares y tercero, se describen nuevos virus.

2.2.2. Datos experimentales, pertenecientes a la familia Culicidae

Los datos metavirómicos experimentales fueron muestreados en el marco del proyecto *"Expedición virológica en ecosistemas representativos de Colombia: Selva húmeda tropical de la Sierra Nevada de Santa Marta"*, se realizaron en tres lugares diferentes en la Sierra Nevada de Santa Marta, Colombia, cercanos a bordes de fragmentos de bosque. El material genético fue extraído con el kit *"Viral RNA Mini Kit"* y fue secuenciado con la tecnología de secuen-

Lectura	Número de bases	Tamaño
SRR6019484	33.3G	12Gb
SRR6019485	19G	7.5Gb
SRR6019486	10G	4Gb
SRR6019487	18.9G	11.3Gb
SRR6019488	27.3G	9.9Gb

Tabla 2-1.: Accesión a los datos crudos de referencia, pertenecientes al metaviroma de *Drosophila sukukii* [1]

cia Illumina, para un total de tres librerías.

Librería	No. corridas	Tamaño	Tamaño de las secuencias	Tipo de adaptador
Primera	12	1.7 Gb	229	illumina universal adapter Solid small RNA adapter
Segunda	10	1.1 Gb	50-251	Solid small RNA adapter
Tercera	12	11 Gb	35-151	Solid small RNA adapter

Tabla 2-2.: Datos crudos de las muestras metatranscriptomicas del proyecto.

2.2.3. Metodología

De acuerdo a las consideraciones anteriores, se construyó una metodología que es altamente restrictiva para evitar que contaminación de adaptadores o agentes externos quedara dentro de la información a analizar y garantizar que hay un alto nivel de pureza en la muestra. Esta etapa fue muy importante ya que para la posterior clasificación, se necesita que las secuencias sean de origen viral en su totalidad y así evitar quimeras de fuente biológica diferente.

El flujo metodológico que se utilizó en este trabajo se muestra en el diagrama de flujo **2-1**. Esta corresponde a un proceso automatizado, que es eficiente en tiempo de computo y en el uso de los recursos computacionales disponibles en el laboratorio. Este preprocesamiento está dividido en 4 secciones, 1) El análisis de calidad inicial de los datos crudos, que muestra un panorama general del estado de calidad de las librerías con el programa **Fastqc** [48], 2) El filtrado de duplicaciones y adaptadores en las secuencias con el programa **RNA-qc-Chain** [44] con la función `RQC-parallel-qc`, 3) El recorte de calidad de las secuencias con el programa **Trimmomatic** [49], 4) El filtrado de secuencias identificadas como RNA ribosomal con el programa **RNA-qc-Chain** [44] con la función `RQC-rRNA-filter`, 5) el mapeo al genoma del hospedero y a la base de datos de contaminantes con el programa **Bowtie2** [50] para la obtención de los datos filtrados de la mayor cantidad de organismos externos

posible y finalmente 6) el ensamblaje metagenómico *de novo* con el programa **Megahit** [47], Aunque generalmente los pasos 2 y 3 son realizados invirtiendo los dos pasos, de acuerdo con las recomendaciones de los autores de **Trimmomatic** [49] la remoción de adaptadores y duplicaciones debe hacerse antes del recorte de calidad para obtener secuencias de mejor calidad.

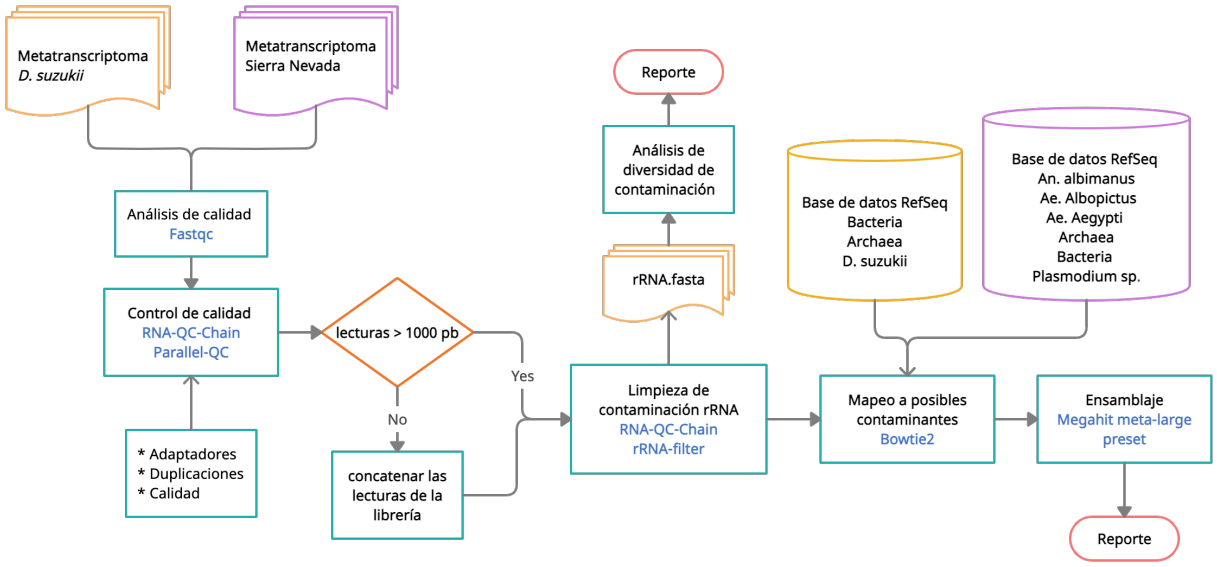


Figura 2-1.: Metodología seguida para el preprocesamiento de los datos crudos, hasta el ensamblaje de los metagenomas, para los dos conjuntos de datos utilizados en este trabajo.

Análisis de calidad

La evaluación de la calidad inicial de la muestra se realizó con el programa **Fastqc** [48], el cual permite hacer un análisis preliminar de los datos crudos provenientes de secuenciación de alto rendimiento, donde a partir de una muestra estadísticamente significativa realiza una serie de análisis basado en la calidad de cada uno de los nucleótidos en las secuencias, como diagrama de caja y bigotes de la calidad de las secuencias por par de base, contenido de adaptadores, distribución del contenido de CG o frecuencia de k-mers, permitiendo a diferencia de los datos estadísticos generados en el momento de secuenciar, un panorama más amplio de las muestras, obteniendo información sobre problemas originados en el secuenciamiento, como también en la construcción de las librerías.

Recorte de calidad y filtrado de secuencias RNAr

En datos RNA-Seq obtenidos con secuenciamiento de próxima generación, se puede tener problemas de calidad que pueden sesgar los resultados analíticos obtenidos como se expuso anteriormente, por lo cual, la remoción de duplicaciones y adaptadores se realizó con el programa **RNA-qc-Chain** con la herramienta `RQC-parallel-qc`, ya que permite de manera eficiente por sus procesos ejecutados en paralelo hacer una revisión exhaustiva de las muestras.

Se realizo el recorte calidad con `Trimmomatic`, únicamente para las muestras de *Droshopila suzukii*, con un mínimo de calidad requerido de 20, este umbral se escogió por debajo de los limites recomendados teniendo en cuenta que por el origen de la muestra los niveles de calidad del conjunto total pueden estar afectados por el alto nivel de contaminación que contiene y un filtro muy estricto puede significar una perdida considerable de información, por otro lado, para las muestras experimentales se decidió no realizar el filtrado porque se encuentran con niveles muy bajos de calidad y este procesamiento significaba una pérdida total de la información.

El recorte de la información se realiza de tres maneras diferentes: por bases de acuerdo a la mínima calidad establecida, por lectura teniendo en cuenta el mínimo de calidad establecido, el porcentaje de secuencias mínimo que lo soporten y el recorte de adaptadores en base a la base de datos de adaptadores de referencia para Illumina. Esto permite tener un amplio espectro de posibilidades que se ajusta a diferentes tipos de datos.

El RNA ribosomal puede ser considerado contaminación interna, si proviene de la misma especie y externa cuando pertenece a una especie foránea, la herramienta `RQC-rRNA-filter` del mismo programa [44], permite la identificación y extracción de secuencias rRNA contaminantes, primero, por medio de modelos ocultos de Markov (HMM) implementados en el programa HMMER [43] identifica las secuencias de rRNA y las extrae. Este modelo fue construido para la identificación de subunidades ribosomales con genoma tipo RNA de 16S y 23S pertenecientes al ribosoma en procariotas y 18S y 28S al ribosoma de eucariotas lo que permite hacer una extracción exitosa de contaminantes de origen ribosomal sin la necesidad de hacer un alineamiento y una anotación previa; posterior a esto se realiza una identificación de las secuencias de procariotas y eucariotas extraídas de rRNA, que fueron mapeadas a la base de datos SILVA [51], teniendo como resultado las librerías filtradas y las secuencias de rRNA debidamente identificados y clasificados taxonomicamente, que permite tener un panorama de la población que se encontraba en la muestra.

Mapeo para identificar contaminación externa a la viral

Debido a que en la muestra también puede haber material genómico adicional al RNA ribosomal encontrado, el cual provenga de organismos no virales, es necesario realizar un mapeo de las muestras a las bases de datos de organismos que pueden estar en la muestra, de acuerdo a el origen de las muestras, el método de extracción del RNAr y de su secuenciamiento. Se decidió entonces, realizar el mapeo a base de datos de Bacterias y Archaeas del `refseq` [52] version 200, como también a la base de datos de los hospederos. Para el primer tipo de datos se empleo el genoma de referencia de *Drosophila suzukii* (código de accesion del genoma GCF_000472105.1) y para las muestras experimentales los genomas de referencia de *Aedes aegypti* [53], *Anopheles albimanus* [54]y *Aedes albopictus* [55].

El alineamiento de las secuencias genómicas, al tener un espacio de búsqueda tan alto, es uno de los procesos que toma más tiempo para un algoritmo de mapeo, por lo cual se han desarrollado diferentes aproximaciones para obtener el mejor resultado. Sin embargo, en muchos casos, la capacidad de computo era una restricción alta, reduciendo la sensibilidad y precisión de los resultados contribuyendo a la pérdida de información. La función `bowtie2-align` del programa **Bowtie2** [50], combina estrategias de otros algoritmos de alineamiento previos, favoreciendo la sensibilidad, la precisión y el alto desempeño del algoritmo. El uso del procesamiento en paralelo en conjunto con algoritmos de recursión que escalen de forma lineal brinda una alta velocidad y un uso eficiente de la memoria.
Pese a todos los esfuerzos de optimización del programa, este sigue siendo un proceso muy costoso computacionalmente en cuando a memoria de almacenamiento utilizada. Por esto fue necesario desarrollar un método de almacenamiento de los datos de salida de **Bowtie2** por medio de diccionarios en el lenguaje de programación `python`, en donde se guardaban únicamente los encabezados que no habían sido mapeados a las bases de datos de contaminación, para luego ser extraídos del archivo de entrada, minimizando el espacio de almacenamiento de la información.

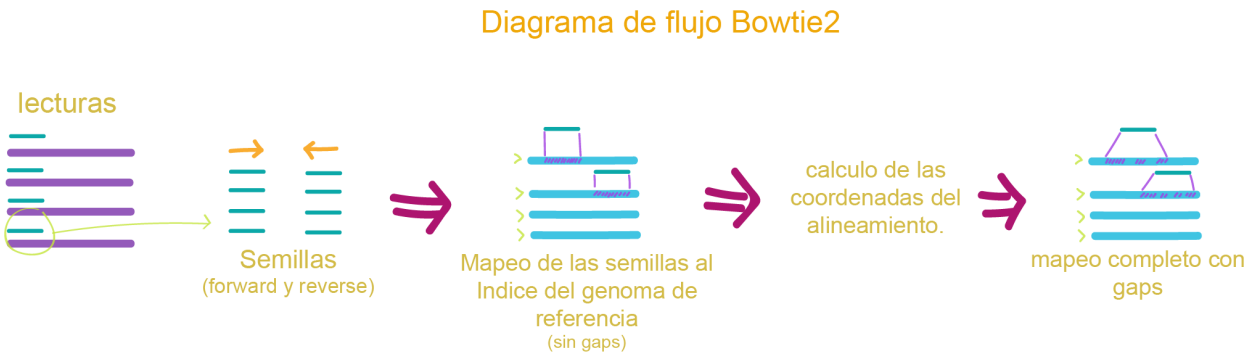


Figura 2-2.: Metodología de Bowtie2, explicada en 4 pasos por sus autores [50]

Ensamblaje

Teniendo en cuenta los requerimientos necesarios para el ensamblaje de una muestra metavirómica, descritos anteriormente, para el desarrollo del programa `MegaHit` [47], los autores por medio de la implementación paralelizada de los grafos *de Bruijn* sucintos [56], que hace referencia a las estructuras de datos sucintas, en la cual el espacio de almacenamiento que requiere se acerca al mínimo espacio teórico de información, desarrollaron un método de ensamblaje eficiente en tiempo, espacio de almacenamiento y en costo computacional, apto para CPU y GPU, por medio de un diseño iterativo con diferentes tamaños de k-mers y una alta flexibilidad para determinar la cobertura, al tiempo que se reducen los k-mers que contienen errores de secuenciamiento y tienen una baja expresión, maximizando el uso de la información disponible.

Adicionalmente, se seleccionó como tamaño mínimo de contig 500 nt, longitud utilizada comúnmente en ensamblajes genómicos que fue seleccionada siguiendo los parámetros usados en [1]. Adicionalmente se uso el parámetro `meta-large` para el ensamblaje de secuencias complejas, descritas como secuencias con un alto contenido de diversidad, para el ensamblaje de los dos conjuntos de datos.

2.3. Resultados y discusión

Control de calidad metaviroma Drosophila suzukii y muestra Sierra Nevada de Santa Marta

Como primera evaluación de los datos crudos pertenecientes al metaviroma de *Drosophila suzukii* [1] y las muestras pertenecientes a la Sierra Nevada de Santa marta, se utilizó el programa `Fastqc` [48], cuyos reportes se encuentran ubicados en la sección de anexos A. Con base en estos reportes se realizaron las gráficas **2-3** para *Drosophila suzukii* y **2-4**, **2-5** y **2-6**, para la primera, segunda y tercera muestra de la Sierra Nevada respectivamente, en donde de forma resumida se puede ver la evaluación que el programa determina para cada muestra en cada una de las categorías, para los cuales se mantiene en todas las muestras un alto contenido de adaptadores, k-mers, duplicaciones y secuencias sobre representadas. Por último, el contenido de GC para las diferentes muestras no sigue una distribución unimodal, comportamiento habitual en muestras de origen metagenómico, ya que muestras que contienen información de diferentes organismos tienden a tener distribuciones de GC multimodales. El compendio de gráficas con esta información esta disponible en el anexo A.

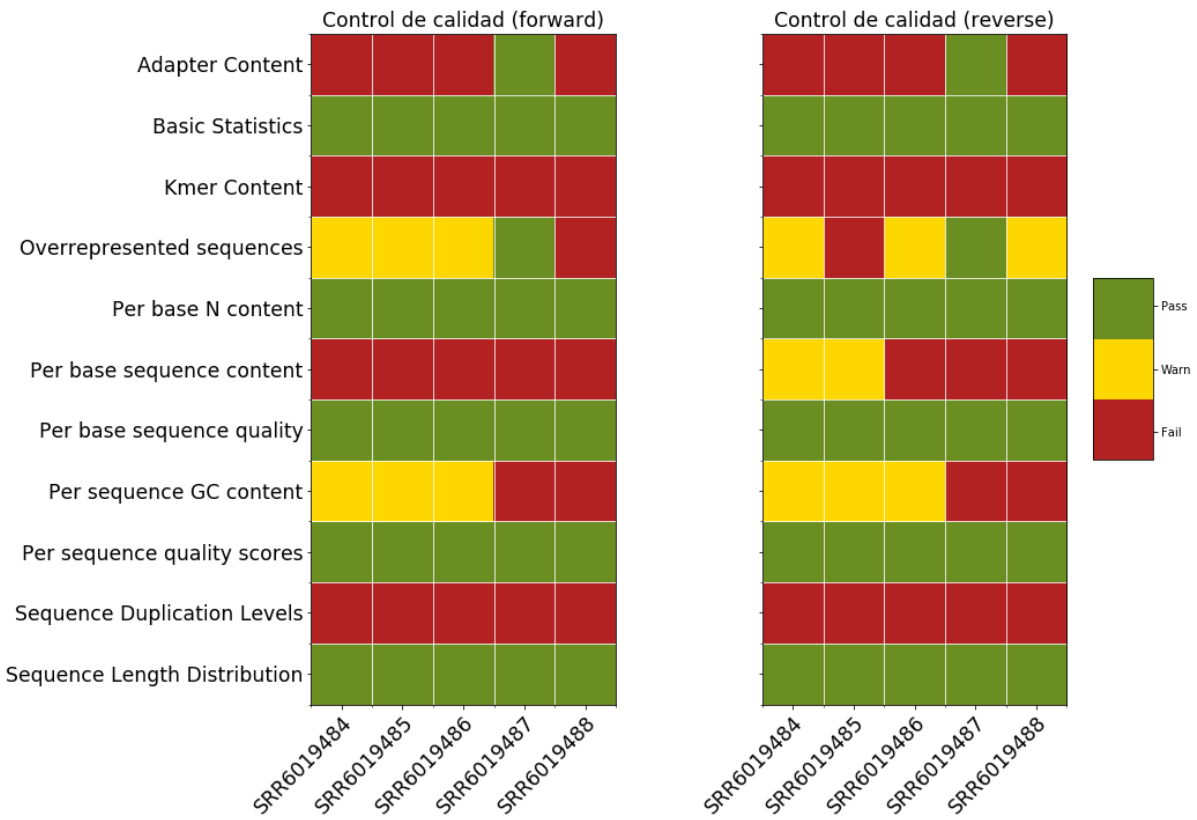


Figura 2-3.: Control de calidad librerías *Drosophila suzukii*.

Limpieza de duplicaciones y adaptadores

La limpieza de duplicaciones, adaptadores y calidad muestra para las muestras de la Sierra Nevada de Santa Marta una proporción de cerca de una cuarta parte del total de la muestra, para la librería 2 se observo una proporción de tres cuartas partes de contaminantes externos. Por otro lado las muestras de *Drosophila suzukii* muestran una proporción mucho más grande, con cerca de un 80 % de contenido contaminante de este tipo. Esta información comprueba que metodologías con un estricto preprocesamiento de los datos son necesario para poder llegar a conclusiones confiables, la limpieza de duplicaciones, adaptadores y de calidad son necesarios debido a la alta proporción de contaminantes provenientesde reactivos moleculares y de los métodos de secuenciación, que pueden conformar cerca del 90 % de la muestra, o como otros casos registrados en diferentes tipos de muestras en animales y ambientales para identificación vírica con proporciones del 75 % de agentes externos de este tipo [57].

Figura 2-4.: Control de calidad primera librería proyecto Sierra Nevada de Santa Marta.

Figura 2-5.: Control de calidad segunda librería proyecto Sierra Nevada de Santa Marta.

Análisis de Contaminación RNA ribosomal

El filtrado de secuencias de RNAr, es uno de los pasos mas importantes en el preprocesamiento de los datos metagenómicos y que muchas veces no es incluido dentro de las pipeline de preprocesamiento ya que se asume que el uso de kit diseñados para la limpieza de RNAr de la muestra son suficientes. De acuerdo con la metodología del programa **RNAqcChain**, se pudo rastrear elementos de 16S, 18S, 23S y 28S e identificarlos taxonómicamente, lo que permitió evaluar un aspecto también importante dentro del análisis de la información genómica, pues permitió ampliar el conocimiento sobre el contexto en el que fue extraída la muestra, lo que cobra mayor sentido en el marco de la metagenómica.

Para la muestra de *Drosophila suzukii* se encontró una proporción similar en las 4 subunidades de RNAr, exceptuando la muestra **SRR6019487** que tuvo un número mucho menor de RNA ribosomal. Por otro lado las muestras de la Sierra Nevada de Santa Marta muestran un contenido de RNAr inferior.

Figura 2-6.: Control de calidad tercera librería proyecto Sierra Nevada de Santa Marta.

Ensamblaje metavirómico

De acuerdo a los resultados del ensamblaje metavirómico, en la gráfica **2-12** se muestran las métricas mas relevantes producto del ensamblaje. Se obtuvo el mayor tamaño de ensamblaje para la muestra *SRR6019485* y un menor tamaño para la muestra *SRR6019488*, lo que demuestra que no es directamente proporcional el volumen de los datos crudos al del ensamblaje final, por un lado el número total de contigs (palabra que proviene de çontiguoz que hace referencia a la unidad de ensamblaje de fragmentos que se sobrelapan) mantiene las mismas proporciones que el tamaño, sin embargo con respecto a su longitud, el mas largo del ensamblaje de *SRR6019485* es el mas pequeño de las 4 muestras, lo que indicaría que aunque sea el ensamblaje que mayor contenido de información capturó, no fue exitosa la recuperación de la información del contig de mayor longitud.

Por otra parte, el N50 que representa el 50 % del ensamblaje que está contenido en contigs iguales o más largos que la longitud de este contig y el L50 que está definido como el número más pequeño de contigs que su longitud sumen la mitad del tamaño del ensamblaje. Teniendo en cuenta estas métricas, se esperaría que un buen comportamiento de un ensamblaje tendría los valores N50 y L50 con diferencias significativas, como lo obtenido en el ensamblaje de *SRR6019484*, sin embargo, para las otras librerías se obtuvo niveles de N50 y L50 muy similares, indicando que estos ensamblajes están compuestos por contigs muy pequeños; Las métricas N90 y L90, tienen el mismo significado de las dos anteriores con la diferencia de que están descritas para el 90 % de los datos, en este caso se observa un comportamiento esperado, donde las dos métricas tienen una diferencia significativa, exceptuando el ensamblaje de *SRR6019488*, que las dos métricas son casi iguales. Finalmente es importante destacar que el contenido de GC en todas las muestras se presenta con su desviación estándar, ya que en

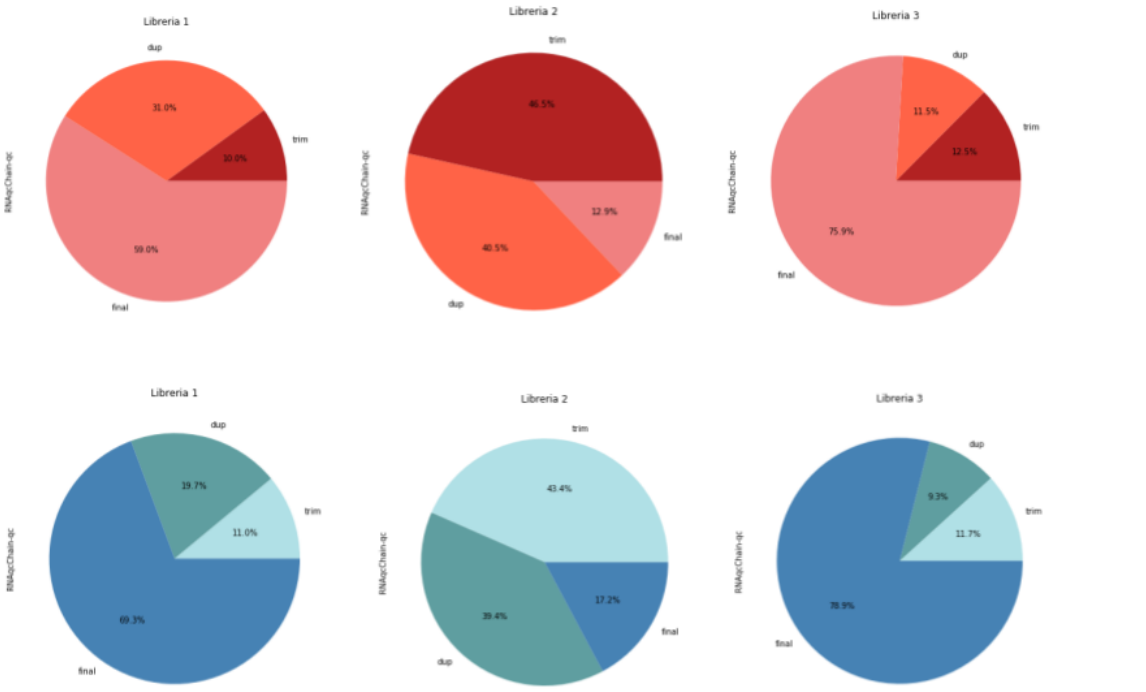


Figura 2-7.: Porcentaje de adaptadores y recorte duplicaciones para los datos del proyecto, en donde los diagramas de colores rojos corresponden a las secuencias en sentido *forward* y los diagramas en azul a las secuencias en sentido *reverse*.

el caso de los análisis metavirómicos la desviación en el porcentaje de GC da muestra de que hay una alta diversidad en la muestra.

Para los ensamblajes de las muestras de la Sierra Nevada de Santa Marta los resultados fueron más heterogéneos, sin embargo, antes de hacer un análisis de este ensamblaje, se debe enfatizar en que debido a que el parámetro de calidad del secuenciamiento no se realizó dados los valores de baja calidad y que es esta parte procedimental fundamental como se ha descrito anteriormente en el análisis del ensamblaje, para este tipo de datos el análisis sobre las métricas del ensamblaje como única evaluación de este es inconclusa. Para comprobar que el contenido que ha sido ensamblado tiene una relación entre ellos debe someterse a posteriores análisis. Por otra parte se observa que el ensamblaje de la tercera librería tiene una mejor calidad de ensamblaje que para las otras dos, como se observa para la gráfica **2-13**, en donde el tamaño del ensamblaje, el N50L50, N90L90 y el número total de contigs, aunque este último no sea proporcional a una buena calidad, incluso se presenta en escala logarítmica para que sean visibles los tres valores, aunque su diferencia sea significativa. Finalmente, para el caso de la tercera librería el N50 alto y el L50 pequeño son buenos indicadores de la calidad del ensamblaje.

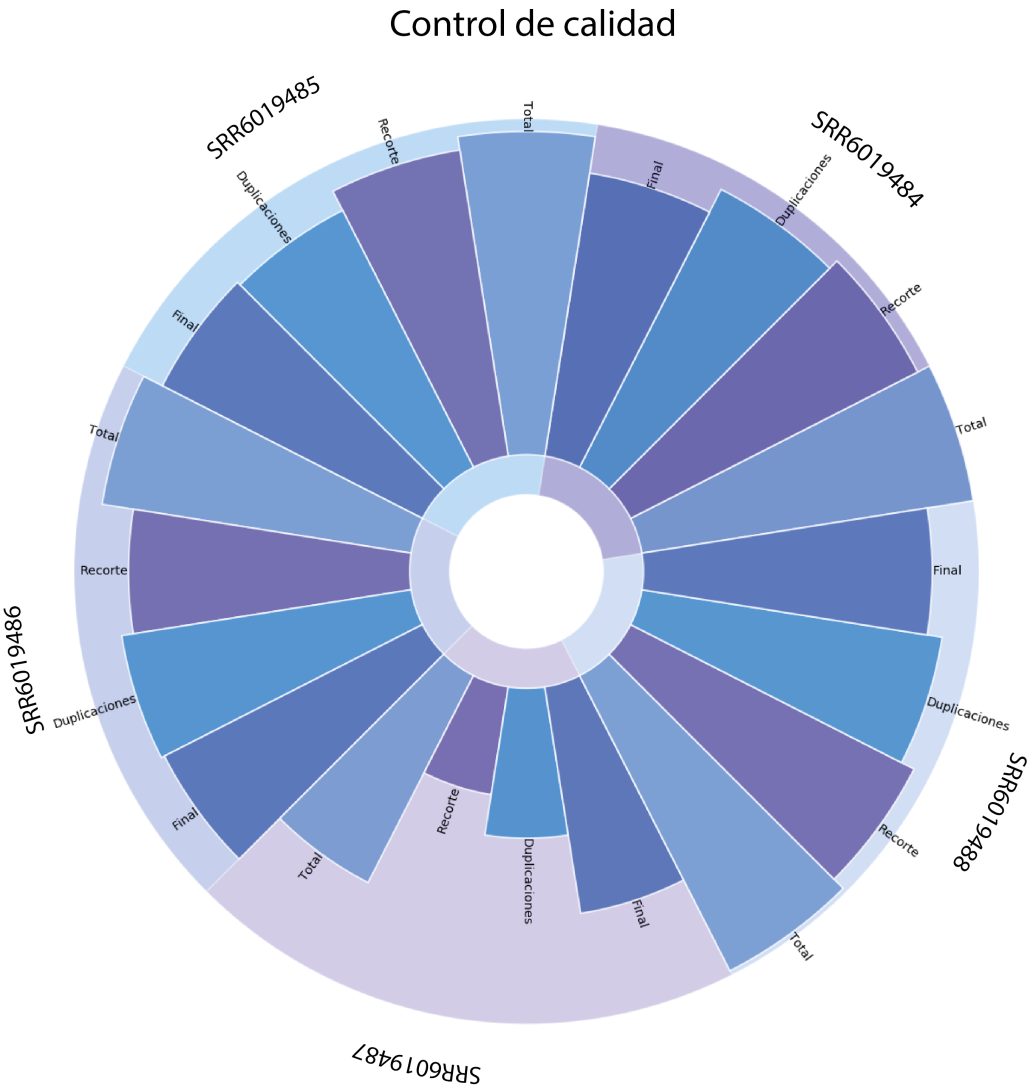


Figura 2-8.: Contaminación rRNA *Drosophila suzukii*

	muestra genómica	Después del filtrado de duplicaciones y adaptadores	Después del filtrado de calidad	Después del filtrado de contaminantes RNAr	Después del filtrado de contaminantes	Ensamblaje
SRR6019484	35 Gb	28 Gb	3.7 Gb	2.6 Gb	860 Mb	9.1 Mb
SRR6019485	23 Gb	3.0 Gb	2.7 Gb	2.0 Gb	487 Mb	6.2 Mb
SRR6019486	15 Gb	3.6 Gb	2.8 Gb	2.5 Gb	162 Mb	2.0 Mb
SRR6019487	24 GB	179 Mb	158 Mb	157Mb	9.5 Mb	133 Kb
SRR6019488	33 Gb	4.1 Gb	2.8 Gb	2.2 Gb	2.2 Gb	82 Mb

Tabla 2-3.: Tamaño de la muestra de *drosophila suzukii* en cada paso de la pipeline.

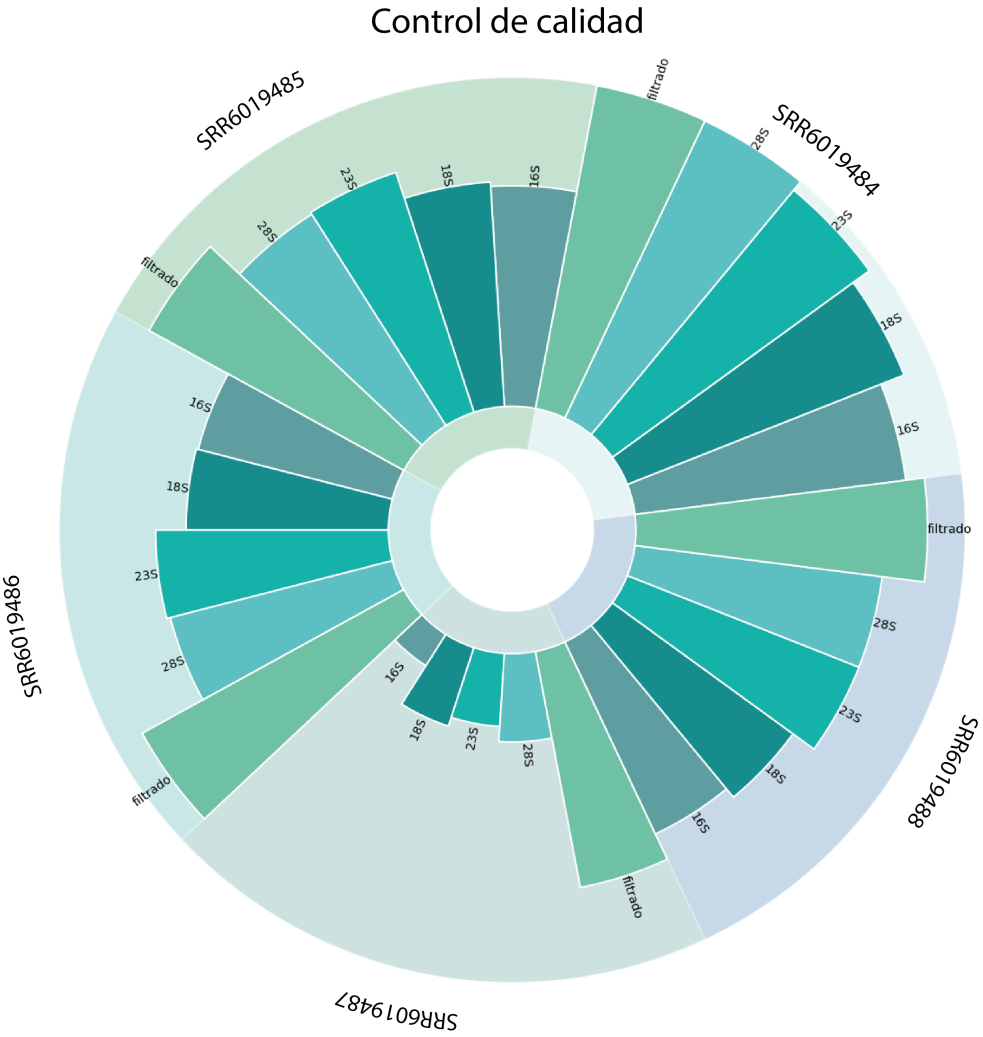


Figura 2-9.: Control de calidad *Drosophila suzukii*.

	muestra genómica	Después del filtrado de duplicaciones y adaptadores	Después del filtrado de contaminantes RNAr	Después del filtrado de contaminantes	Ensamblaje
Primera	822 Mb	377 Mb	179 Mb	167 Mb	119 Kb
Segunda	550 Mb	88 Mb	32 Mb	31 Mb	104 Kb
Tercera	5.4 Gb	4.0 Gb	3.4 Gb	3.0 Gb	9.7 Mb

Tabla 2-4.: Tamaño de la muestra metagenómica de la Sierra Nevada de Santa Marta en cada paso de la pipeline.

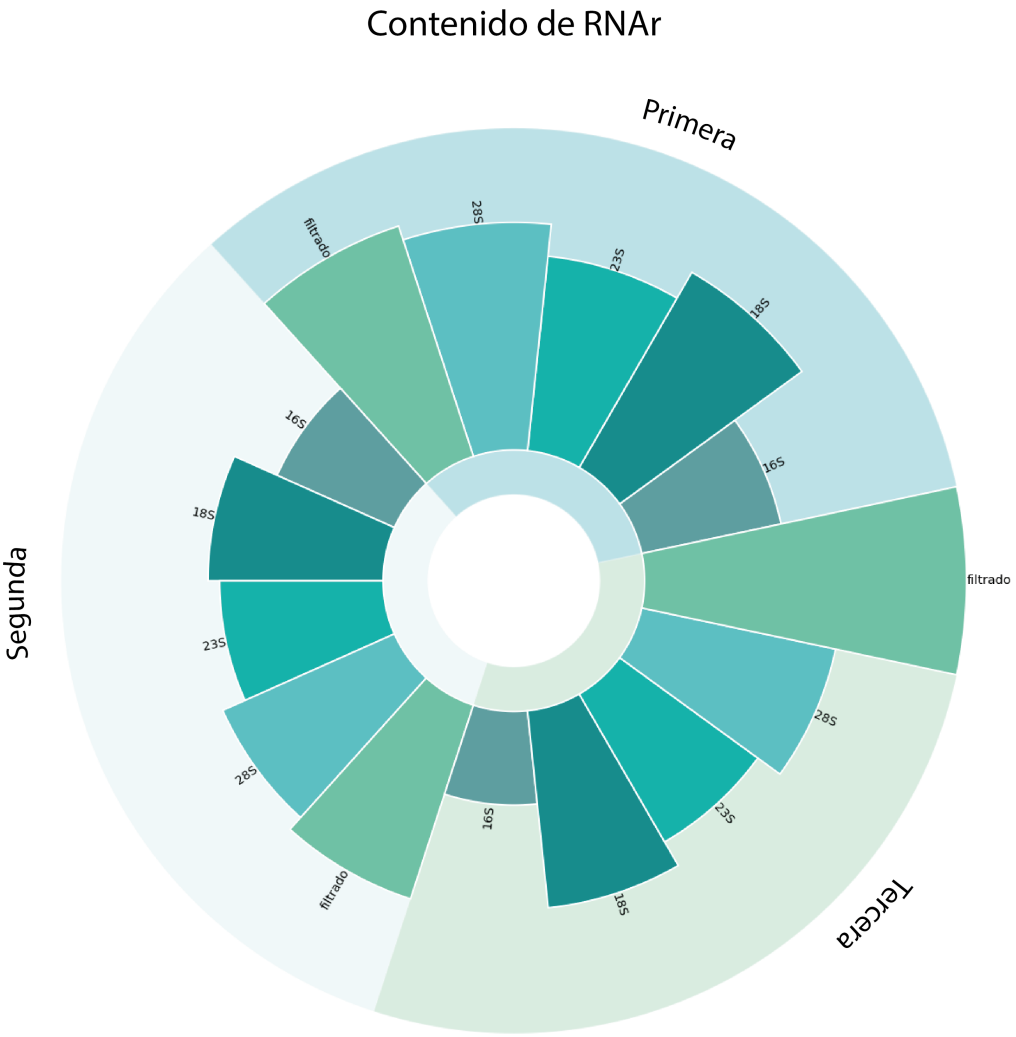


Figura 2-10.: Contaminación rRNA muestra Sierra Nevada de Santa Marta.

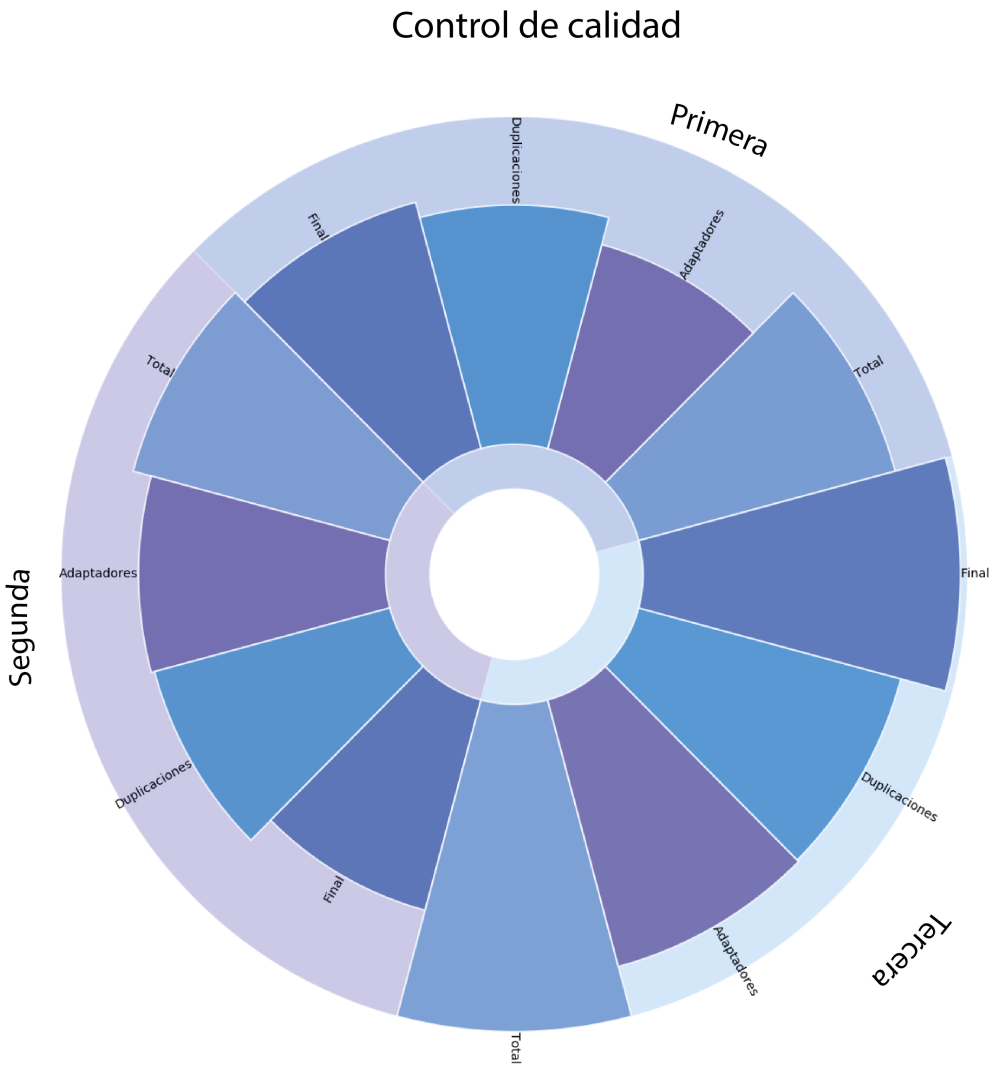


Figura 2-11.: Control de calidad muestra Sierra Nevada de Santa Marta.

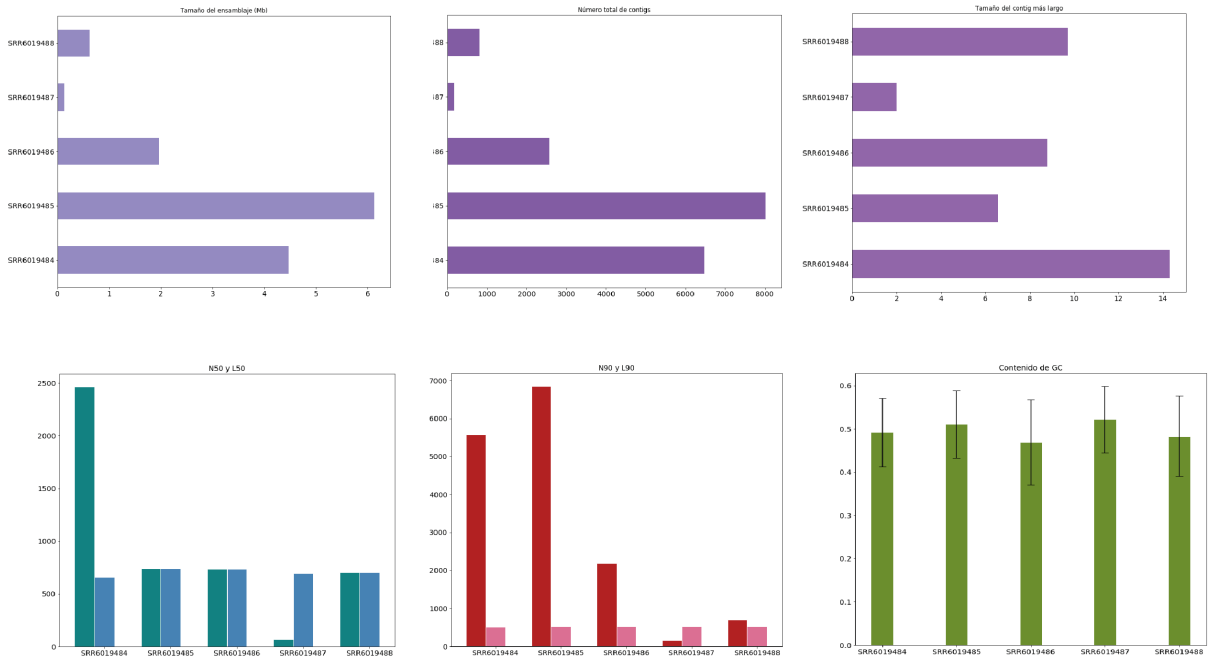


Figura 2-12.: Estadísticas de ensamblaje del metaviróma de *Drosophila suzukii*

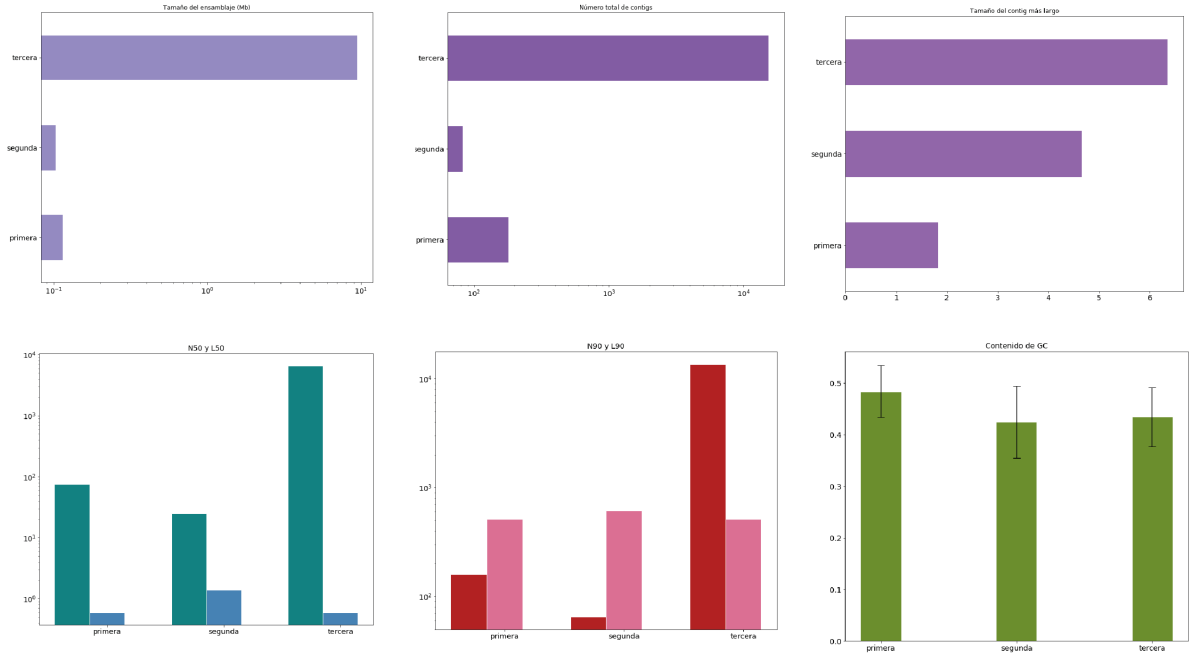


Figura 2-13.: Estadísticas de ensamblaje del metaviroma de la Sierra Nevada de Santa Marta

2.4. Conclusiones

De acuerdo con los resultados obetnidos **2-3** y **2-4**, se observa que con la metodología aplicada se filtró una cantidad considerable de información y que el ensamblaje fue realizado con un porcentaje muy bajo de los datos crudos, esto se debe en primera instancia a la clase de muestra genómica que se esta trabajando, por el método de recolección y extracción de la información con lo cual se ha sido reiterativo y en segunda instancia a que se utilizaron programas que permiten hacer un filtrado exhaustivo de las muestras y que son ampliamente usados en muestras metagenómicas, lo que permitió obtener una mejor calidad y pureza de los datos, como tambien realizar el ensamblaje a partir de muestras con baja profundidad[47].

3. Implementación y validación de los métodos de clasificación.

3.1. Introducción

Los métodos de aprendizaje de máquina se han convertido en la herramienta más usada en la actualidad para la clasificación de datos de diferentes volúmenes y complejidades. Este incremento se debe a la creciente disponibilidad de recursos computacionales de fácil acceso como los servidores en la nube y la progresiva demanda de centros de computo en las instituciones, como también por la elevada inversión en la optimización de algoritmos y librerías por su potencial uso en tecnología. Gracias a la inversión económica en estas metodologías problemas complejos en biología y para este caso en la genómica pueden ser abordados desde otras perspectivas.

Para el caso de la información virómica, el nivel de conocimiento que tenemos sobre el universo de información es muy bajo, el volumen de información es lo suficientemente detallado para tres familias, pero se sigue teniendo un desconocimiento amplio sobre más del 90 % de las 189 familias descritas actualmente[58], sin embargo, aunque este es uno de los grupos biológicos con menor representación en el **Genbank** su clasificación manual se hace cada vez más dificil de realizar [6].

Las necesidades en salud pública que cada año se hacen más notorias, la globalización y el creciente aumento de los estudios genómicos a gran escala, requieren el uso de herramientas efectivas y confiables para la identificación de información genómica, tareas que son altamente demandantes y que en la actualidad no están siendo cubiertas en su totalidad [6]. Por otro lado, el desarrollo de herramientas que sean precisas y al mismo tiempo eficientes computacionalmente se ha convertido en uno de los retos actuales. La implementación de métodos estadísticos robustos que permiten extraer la mayor cantidad de información del conjunto de datos disponibles, es una necesidad que no es nueva en esta área, pero que con la coyuntura actual de la pandemia, se han hecho más evidentes los vacíos en este campo de investigación.

3.1.1. Modelos de clasificación computacionales para el análisis de las secuencias usando aprendizaje de máquina

Los métodos de máquinas de aprendizaje (*machine learning*), permiten construir modelos computacionales mediante el uso de teoría estadística con el fin de hacer inferencias sobre las muestras que se están analizando, con el objetivo de encontrar patrones que describan el conjunto de datos. Ya sea usando métodos supervisados con un entrenamiento basado en una etiqueta o clasificación *a priori*, o no supervisados realizando transformaciones del espacio muestral [59].

La implementación de métodos de aprendizaje de máquina usados como métodos de clasificación, han sido ampliamente discutidos y empleados en diversos campos, más recientemente se ha visto su potencial de aplicación en el área de la genómica, en la clasificación de secuencias de diversos orígenes [60], en la reconstrucción de filogenias utilizando estadística Bayesiana, predicción de estructuras moleculares y funcionalidad, biología de sistemas, entre otros problemas de la biología que antes no podían ser abordados por el tamaño de la información y por los algoritmos basados en comparaciones [6] [61] [62].

En el presente trabajo se realizó una evaluación de algunos de los algoritmos más usados para diversos problemas en biología o en análisis de texto, como Máquinas de soporte vectorial y gradiente potenciado, haciendo una revisión sobre la eficiencia, precisión, demanda computacional entre otros. Previo al entrenamiento de los modelos se realizó una curaduría exhausta y detallada de la base de datos de referencia de genómica viral con preprocesamiento de las secuencias, teniendo así un modelo que puede ser reproducible y mejorado.

Para la implementación de los sistemas de aprendizaje de máquina se debe realizar una transformación de los datos a una representación numérica. Para el caso de análisis de caracteres, dado que los genomas tienen una representación en caracteres, se han desarrollado diferentes métodos de representación enfocado en diferentes problemas y tipos de datos. Con las secuencias genómicas han sido ampliamente utilizadas la representación *One Hot* y la representación en *k-mers*.

Para la representación *One Hot*, se realiza una matriz binaria en donde el número de filas corresponde al número de elementos en el alfabeto y el número de columnas al número de elementos que tiene el *iesimo* dato. Este tipo de transformación es completamente fiel a los datos, su limitación radica en que la matriz del conjunto de datos aumenta su tamaño de acuerdo al tamaño del alfabeto, en este caso 4 veces su tamaño y considerando que el espacio ocupado por esta matriz debe ser contiguo. Estas restricciones hacen que esta representación sea inviable computacionalmente por los recursos de memoria que se necesitan no solo para almacenar la información, sino también para los cálculos posteriores que pueden requerir una capacidad mucho mayor.

Para la representación en *k-mers*, se construye el alfabeto con las posibles palabras que pueden ser construidas con los caracteres que conforman la secuencia, el alfabeto para este caso es de tamaño 4 por el número de nucleótidos y se realiza la extracción de fragmentos o palabras de tamaño k, mediante pasos de tamaño 1, con el cual se hace un conteo de frecuencias por palabra, esta matriz tiene un tamaño igual al número de elementos multiplicado por el número de combinaciones posibles de k con los caracteres disponibles. (para un $k = 3$, número de *elementos* $\times$ 64.)

- **Máquinas de Soporte Vectorial**

  Las máquinas de soporte vectorial (SVM) hacen parte de los modelos de aprendizaje supervisados, fue desarrollado por Vapnik a finales del siglo XX, quien afirma que dependiendo de la representación de los datos en una dimensión mas alta que el espacio de los datos original, dos categorías podrían ser separadas por un hiperplano [63]. Las funciones mas usadas para la construcción de los hiperplanos puede ser linear, polinomial, RBF (*Radial Basis Function*) o sigmoidea [64].

  Este método es uno de los más ampliamente utilizados, por su velocidad y buen desempeño en casos donde el número de muestras es limitado (menos de 1 millón), como tambien, se ha demostrado que tiene una alta efectividad con datos con una alta dimensionalidad [6]. El objetivo es encontrar un hiper plano que separe las dos clases con el mayor margen posible, lo que significa que hay una mejor generalización del clasificador.

  En el caso en el que los datos no pueden ser separados en la dimensión original se realiza una transformación añadiendo una nueva dimensión, que permite obtener una nueva organización de los datos y posiblemente una división del espacio más eficiente.

  Por otro lado, en el caso de las clasificaciones multiclase, se divide el problema en comparaciones binarias, llamado *one-vs-one*, en donde el número de clasificadores que se construirán están dados por la formula: $\frac{n(n-1)}{2}$, con esta combinación de clasificadores se obtiene el clasificador de esta multiclase [65].

  En la representación de la figura **3-1**, se dividen las clases a través de un hiperplano, en la gráfica la línea negra continua, la cual separa los dos conjuntos de datos con el mayor margen posible.

- **Gradiente potenciado**

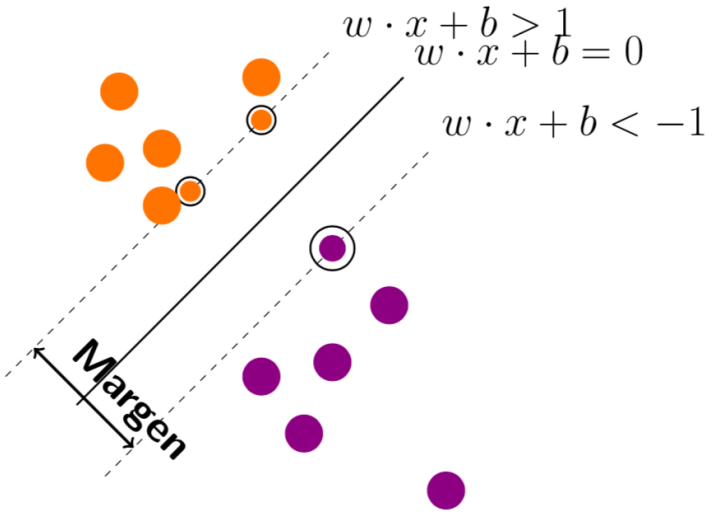


Figura 3-1.: Representación de una Máquina de soporte vectorial, tomado y modificado de El Baúl Del Programador [66]

El método de gradiente potenciado o *Gradient boosting en inglés*, es un método de clasificación en el cual se unen varios árboles de decisión, con el fin de crear un ensamble de pequeños modelos en donde cada árbol corrige los errores del anterior, teniendo como objetivo minimizar el error mínimo cuadrado.

Los árboles de decisión crean un modelo que predice las etiquetas por medio de un árbol de preguntas de verdadero o falso, estimando el mínimo número de preguntas que son necesarias para evaluar la probabilidad de tomar una decisión correcta. Los métodos de ensamblaje combinan múltiples algoritmos para obtener un mejor modelo. A diferencia del algoritmo de bosques aleatorios que construye los árboles de forma paralela, el gradiente potenciado, donde el termino *Boosting* hace referencia a la mejora de modelos sencillos combinandolos, para generar un mejor modelo construido colectivamente [67].

Cada árbol de decisión usa el error residual del modelo anterior para ajustar el siguiente modelo, donde finalmente la útlima predicción es la suma de todas las predicciones de los árboles. Mientras los bosques aleatorios minimizan la varianza y el sobreajuste, los árboles de gradiente potenciado minimizan el sesgo y el subajuste [68].

El termino *Boosting* hace referencia al uso de modelos sencillos de forma secuencial, si se excede el número de estos modelos aumenta la probabilidad de que el modelo se sobre ajuste, lo que quiere decir que el modelo se justa bien a los datos de entrena-

miento impidiéndole predecir con precisión nuevos datos, por esto se fija una taza de aprendizaje con la cual se regula el ritmo con el que aprenden los modelos, con una taza de aprendizaje menor es necesario el uso de un mayor número de árboles, pero hay una menor probabilidad de un sobre ajuste del modelo [68].

3.2. Metodología

3.2.1. Preprocesamiento y construcción de la base de datos Goldstandard

Para la construcción de la base de datos de virus RNA de referencia o *Gold standard*, se descargó la base de datos completa de virus del `NCBI virus` [69]. La familia correspondiente a cada secuencia se uso como encabezado, obteniendo un total de 4,057,472 de secuencias de nucleótidos, las cuales se filtraron con el listado de familias RNA reportado en la lista maestra de especies V.1 del 2019 del `ICTV` (*International Committee on Taxonomy of Viruses*) [58] y se filtraron teniendo en cuenta los siguientes parámetros:

- Secuencias no redundantes.
- Secuencias con un número de nucleótidos mayor a 500.
- Secuencias con un alfabeto de A, C, G, T.
- Familias con una representación de mas de 1000 secuencias.

De acuerdo a lo anterior, se obtuvo un total de 473,492 secuencias de familias RNA no redundantes que corresponden a tan solo el 11.6 % de los datos crudos del `NCBI virus`. Esta base de datos curada se denominó como base de datos *Gold standard*, por contener las secuencias que se emplearon de referencia en este trabajo para el entrenamiento de los sistemas de máquinas de aprendizaje.
Debido a la naturaleza de los modelos es necesario hacer una representación númerica de los datos. Para esto existen diferentes métodos, se escogió el método `Count Vectorizer` implementado en la librería `sklearn` en `python`, ya que además de la representación númerica reduce la dimensionalidad de los datos al ser un conteo de los *kmers*, que se puede observar mejor en el gráfico **3-2**, de esta manera, para un número de 3 *kmers* y en conjunto una matriz de 3, 4, y 5 kmers. Esta serie de combinaciones se realizaron:

- Primero para el tamaño de *kmer* 3, por ser el tamaño en que la naturaleza codifica la información a través de codones.

- Segundo, se realizó el conteo para 4, 5 *kmers* y para el conjunto de 3, 4 y 5 *kmers* con el propósito de encontrar si existía una diferencia en el patrón de las secuencias y si se encontraba otro tipo de información con diferentes patrones y que permitiera tener una buena clasificación de las secuencias, como también reconocer si se encuentran patrones de correlación entre los tamaños de *kmers* o redundancia entre las diferentes matrices.

- Tercero, este método favorece la predicción de secuencias que contienen caracteres que no se encuentran incluidos en el alfabeto estándar de codificación de información genética derivada de un alfabeto de caracteres A, C, G, T. Estás regiones no se tienen en cuenta para el conteo de *kmers*, aprovechando al máximo la información relevante.

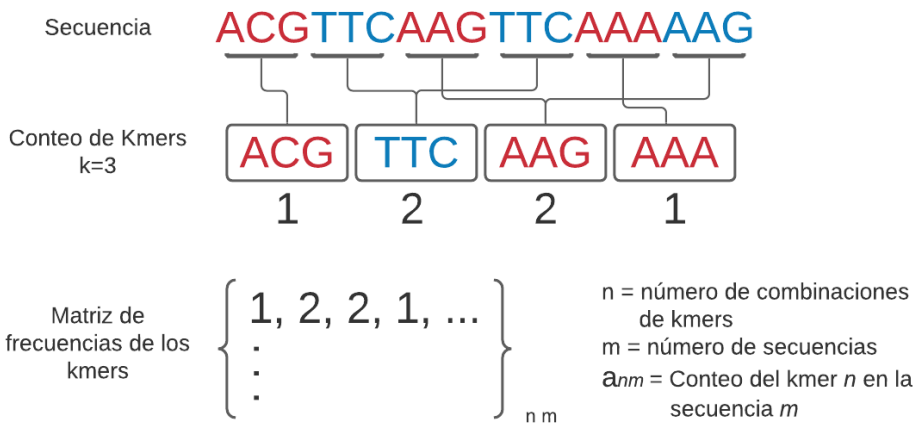


Figura 3-2.: Representación del método *Count vectorizer* para un $kmer = 3$.

Adicionalmente, debido a la alta dimensionalidad de los datos generados, se hace necesario un análisis del peso de cada variable, que en una mayor medida se realiza por medio de un análisis de componentes principales. Esta reducción de la dimensionalidad favorece el desempeño de los procesos de computo. Finalmente, para analizar como cambian las diferentes matrices de representación númerica, se graficó la distribución de los datos en un espacio tridimensional.

3.2.2. Métricas asociadas a los modelos de máquinas de aprendizaje

- Certeza (*Accuracy*), generalmente la mas usada por ser la más intuitiva, ya que mide si un resultado ha sido predicho correcta o incorrectamente, mide el porcentaje de clasificaciones correctas, esta métrica no permite reconocer si el modelo está clasificando

correctamente, por ejemplo, no se puede detectar el sobre ajuste teniendo únicamente esta métrica, por ejemplo cuando el conjunto de datos es desbalanceado.

$$Ac = \frac{VP + VN}{VP + VN + FP + FN}$$

en donde VP son los verdaderos positivos, VN los verdaderos negativos, FP los Falsos positivos y FN los falsos negativos.

- Precisión (*Precision*), se refiere a la dispersión de los valores repetidos predichos para un mismo dato. Esta métrica permite reconocer si un modelo es consistente, pues indica que se están obteniendo datos similares cuando se reproduce el modelo.

$$\frac{VP}{VP + FP}$$

- Sensibilidad (*Sensitivity*), hace referencia a que tan sensible es el clasificador de detectar los verdaderos positivos, siendo la relación entre las predicciones positivas correctas y el número total de predicciones positivas.

$$\frac{VP}{VP + FN}$$

- puntuación F1 (*Score F1*), es una métrica que representa la relación entre la sensibilidad y la precisión, siendo el promedio ponderado entre las dos métricas, teniendo en cuenta los falsos positivos y negativos.

$$\frac{VP}{2VP + FP + FN}$$

- mlogloss, esta métrica toma en cuenta las probabilidades debajo del modelo, no solo la clasificación final, favoreciendo modelos que tienen una diferencia entre clases alta. Este valor es asociado con la entropía por medir la incertidumbre de un modelo, un *log-loss* bajo, significa una baja incertidumbre y baja entropía del modelo, por lo cual se buscan los parámetros que permitan minimizar el *mlog-loss*.

$$\frac{-1}{N}\sum_{i=1}^{N}[y_i log p_i + (1 - y_i) log(1 - p_i)]$$

Donde, p_i es la probabilidad del elemento *iesimo* de ser positivo.

- Raíz del error promedio cuadrado (RMSE), esta métrica corresponde a la raíz cuadrada del promedio de las diferencias entre el valor real y el valor predicho por el modelo. La raíz cuadrada permite la penalización de los errores de acuerdo a su tamaño.

$$\sum_{i=1}^{D}(x_i - y_i)^2$$

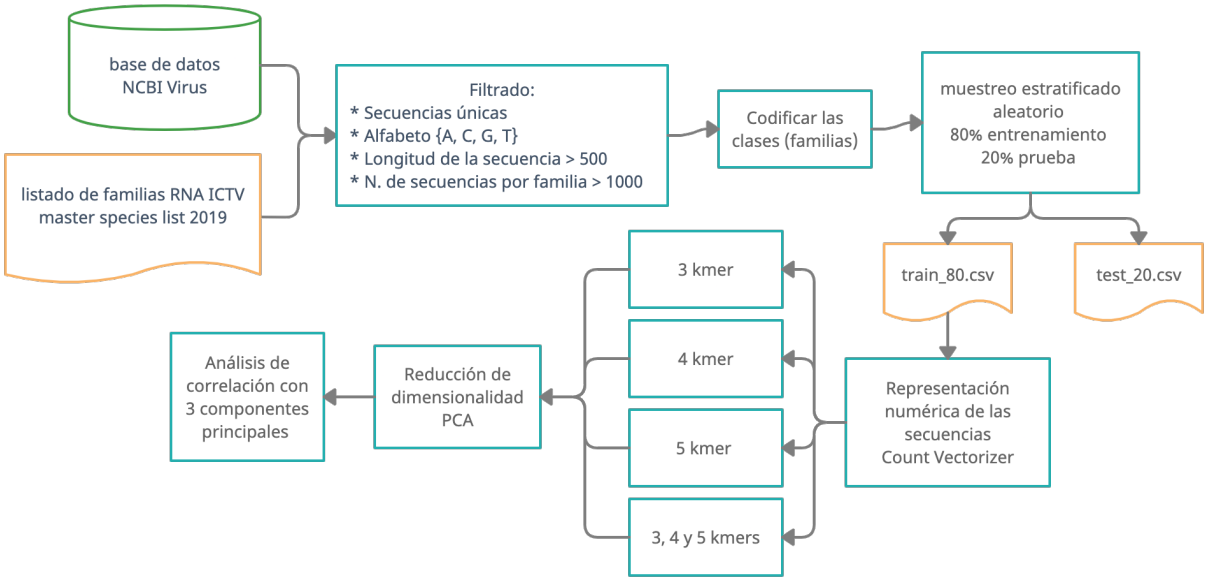


Figura 3-3.: Metodología para el preprocesamiento de la base de datos.

3.2.3. Árboles de decisión con aumento del gradiente (`Gradient boosting`)

En la tabla **3-1**, se describen los parámetros que fueron usados para la construcción del modelo. Se iteró seleccionando variables aleatoriamente para cada árbol en diferentes proporciones (*colsample bytree*), diferentes tasas de aprendizaje y diferentes proporciones de muestras aleatorias, obteniendo como resultado una matriz de tamaño $n - muestras * n - clases$ con las probabilidades de predicción (*multi:softprob*).

3.2.4. Clasificación de Soporte Vectorial

Para el modelo de clasificación de soporte vectorial se emplearon los parámetros descritos en la tabla **3-2**, para una búsqueda aleatoria de los parámetros que se realizó para 5000 iteraciones, cada una con diez particiones de los datos de manera aleatoria, se realizó una validación cruzada y se aplicaron difenretes parámetros para la función de regularización, ambas estrategias para evitar el sobre ajuste, con la función de estratificación aleatoria y manteniendo el balance entre las clases, proceso que tuvo una demanda computacional alta.

Algoritmo: Gradient Boosting	Librería: :XGboost
Parametros:	
objective	multi:softprob
num class	num class
colsample bytree	0.25, 0.5, 0.75
learning rate	0.01, 0.1, 0.5
max depth	3, 4, 5, 6, 7, 8, 9, 10
subsample	0.25, 0.5, 0.75
Parametros Cross Validation:	XGboost
nfold	10
num boost round	num boost
Stratified	True
early stopping rounds	num boost/2
metrics	mlogloss, merror
show stdv	True
seed	123

Tabla 3-1.: Parámetros empleados para la validación cruzada de los modelos entrenados con Gradient Boosting.

Algoritmo: Linear Support Vector classification	Librería: `sklearn.svm`
Parametros:	
`max iter`	5000
`tol`	$1e-5$
`class weight`	`balanced`
`param distributions`	`grid param`
`n iter`	100
`scoring`	`balanced accuracy`
`n jobs`	-1
`random state`	123
Parametros `Cross Validation`:	`sklearn.model_selection` `Random Stratified Kfold`
`C`	`loguniform`$(1e-5, 10)$
`n splits`	10
`n repeats`	10
`random state`	123

Tabla 3-2.: Parámetros empleados para la validación cruzada de los modelos entrenados para la clasificación soportada en vectores. (*log uniform* corresponde a la distribución logarítmica uniforme.)

3.3. Resultados y Discusión

3.3.1. Análisis de la base de datos Gold Standard

De acuerdo con la metodología propuesta, para las cerca de 4 millones de secuencias presentes en la base de datos del NCBI, se encontró que de las 30 familias RNA taxonómicas reportadas por el ICTV [58], mas de la mitad de las secuencias pertenecían a dos familias únicamente: *Orthomyxoviridae* y *Pararnaviridae*, las familias restantes tienen en general una baja representación. Dados estos primeros resultados de análisis se decidió filtrar, excluyendo las familias que tuvieran una representación menor a 1000 secuencias, para obtener finalmente después de los filtros descritos en la metodología una base de datos de referencia *Gold standard* compuesta de un total de 19 familias representadas en 473,492 secuencias y que ocupan un volumen de almacenamiento de 3.6 Gb **3-4**. También en la gráfica **3-5**, se observa la distribución de las secuencias de cada familia con los datos crudos de partida.

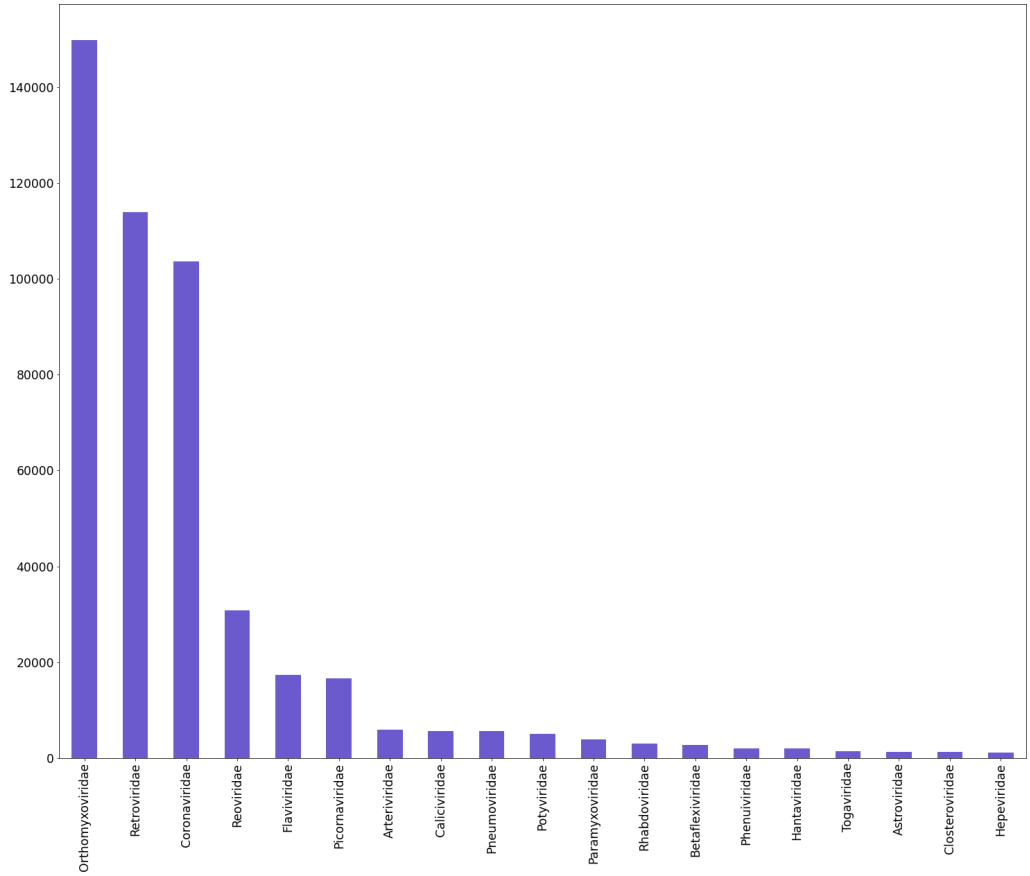


Figura 3-4.: Frecuencias de las familias de virus RNA

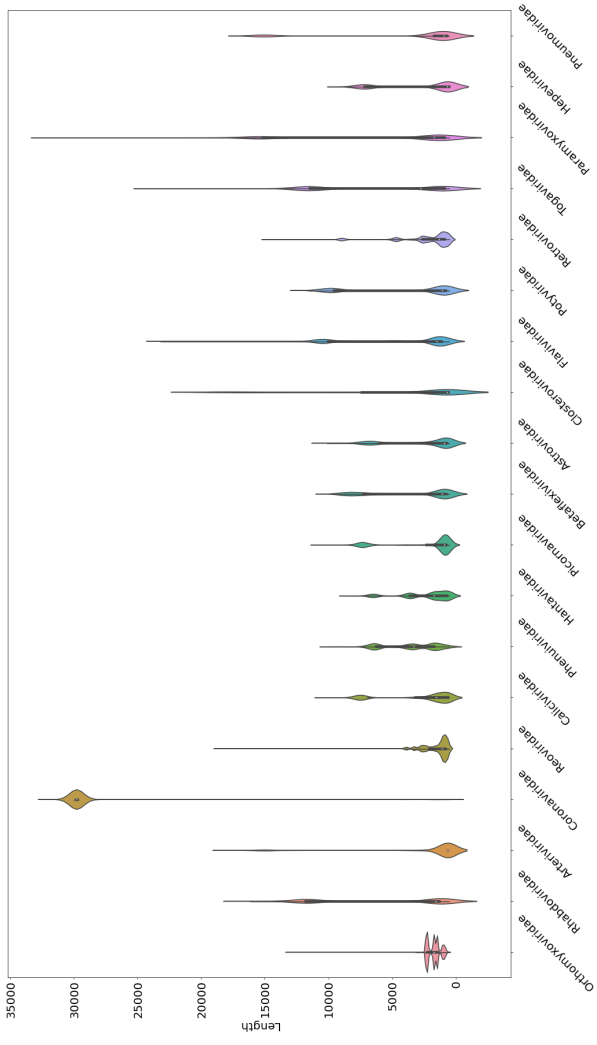


Figura 3-5.: Distribución de violín de los datos crudos de las 19 familias RNA seleccionadas.

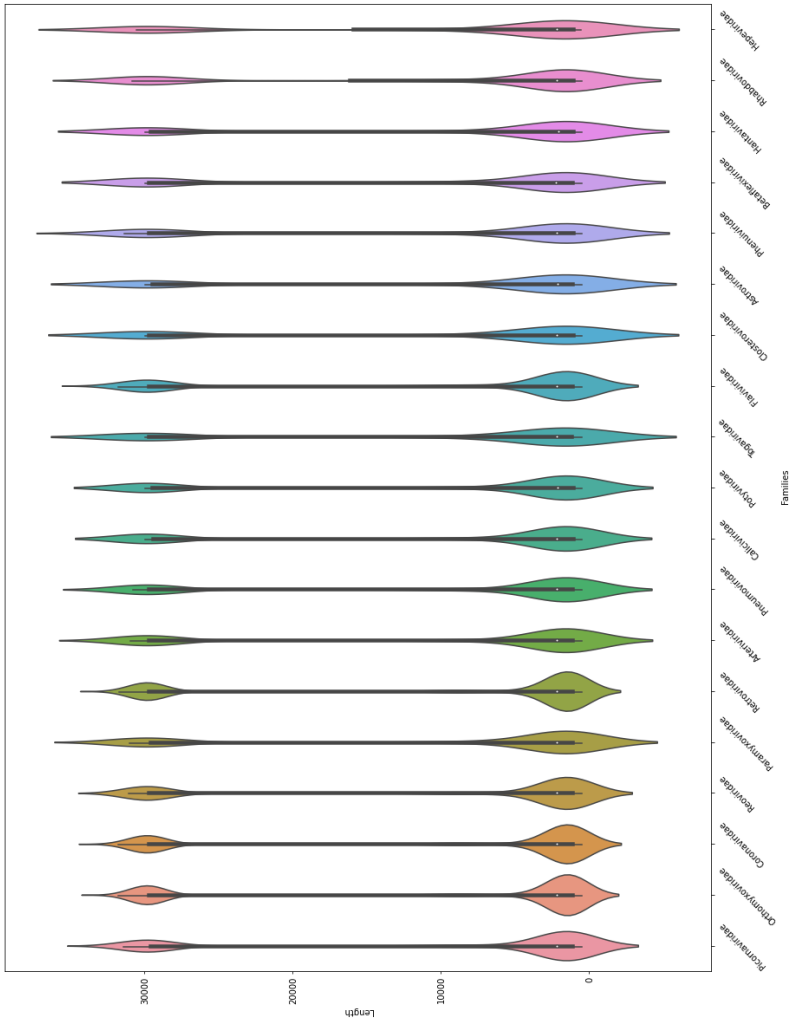


Figura 3-6.: Distribución de violín de cada una de las 19 familias después del filtrado.

En **3-6**, se graficó la distribución de la longitud de las 19 familias seleccionadas, donde se observa una distribución bimodal que se mantiene constante en todas las familias. De esta distribución puede intuirse que se debe a el tipo de información que puede extraerse y secuenciarse del material genético viral, encontrando una curva en longitudes pequeñas pertenecientes a fragmentos genómicos y secuenciamiento de genes y por otro lado una curva en longitudes mayores en donde está incluido el secuenciamiento de genomas completos de virus.

Posterior a la selección de las secuencias de estudio, se realizó la representación de los datos en las diferentes configuraciones propuestas. Se observó que aunque la representación en k-mers redujo la dimensionalidad de los datos significativamente, 64 dimensiones y 1344 dimensiones para 3 y 3, 4, y 5 $kmers$ respectivamente, persistió un volumen alto de información para ser computada por lo cual por medio de un análisis de componentes principales se determinó que el 97.5 % de la varianza es explicada en la primera dimensión para 3 $kmers$ **3-7**a, de la cual se mantuvieron 3 dimensiones y para 3,4 y 5 $kmers$ se explica en 28 dimensiones **3-7**b. Esta transformación de los datos, bajo la representación en el PCA se empleó en los posteriores análisis.

La representación espacial en 3 dimensiones del conjunto de datos provenientes de la representación de 3 k-mers y de 3,4,5 k-mers **3-8**, permite identificar que las diferencias en la configuración espacial de los datos extraídos con diferente magnitud de k-mers se deben a que las dos representaciones aportan información diferente. Sin embargo, en las dos gráficas se puede observar una similitud de la distribución en el espacio muestral, como también que entre las familias hay un sobrelapamiento y las fronteras son difusas. Debido a este resultados, se hizo necesario desarrollar los métodos de clasificación para cada conjunto de datos, con el fin de aprovechar la mayor cantidad de información posible.

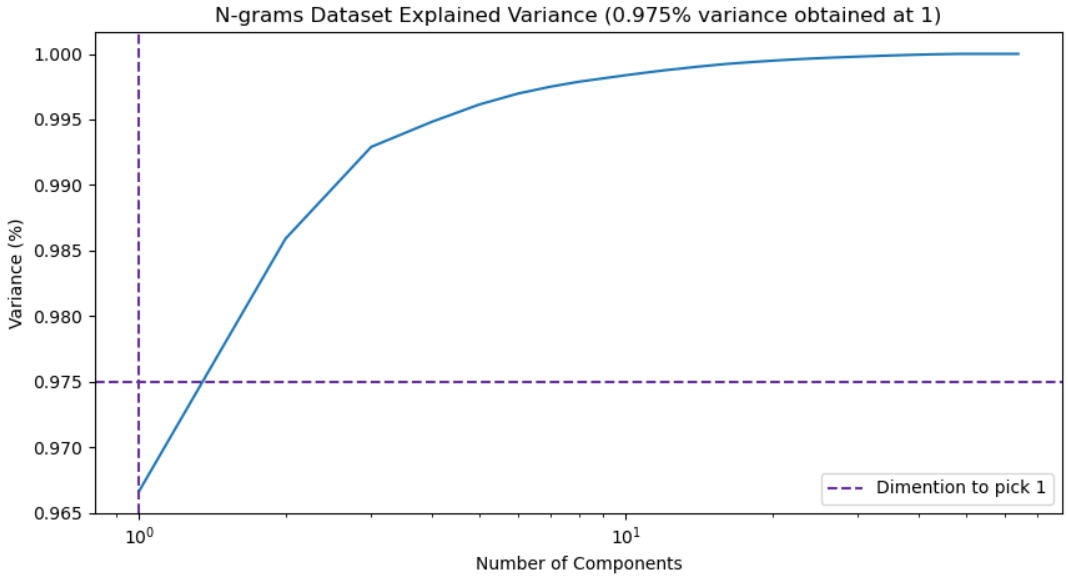
3.3.2. Evaluación de los modelos de clasificación

Árboles de decisión con aumento del gradiente (`Gradient boosting`

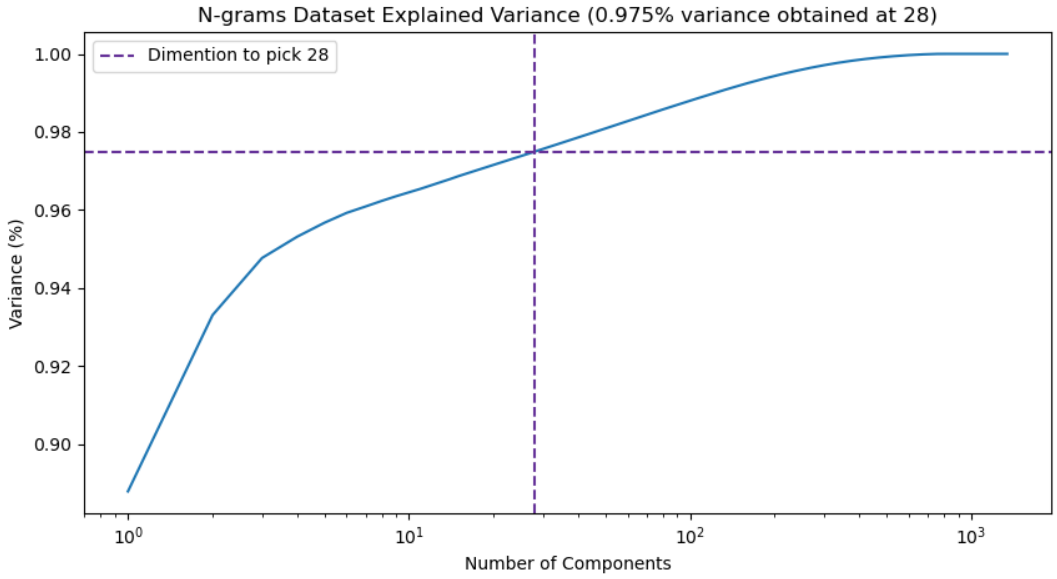
)

De acuerdo a la validación cruzada realizada para la construcción del modelo, se obtuvo un total de 270 combinaciones de los parámetros escogidos (**3-1**) que corresponden al número de iteraciones representados en el eje x de la gráfica **3-9**, también se observa para cada una de las épocas la pérdida logarítmica en la línea azul y en el área sombreada su desviación estándar, en donde se muestran tres tendencias claras que corresponden a las tres fracciones de muestreo de variables por árbol y en general la función de pérdida logarítmica se minimiza.

Por otro lado se observa también otro comportamiento cíclico en cada una de las tres fases que esta determinado por la máxima profundidad permitida para cada árbol. Para el



(a) 3 k-mers



(b) 3, 4, y 5 k-mers

Figura 3-7.: Distribución de la métrica del promedio de la pérdida logarítmica y su desviación estándar para representar el 97.5 % de los datos.

final de la gráfica se encuentra una diferencia significativa entre el modelo de $3kmers$ y de $3, 4y5kmers$, en la gráfica **3-9**a llega al valor mínimo de la función en la época 190 aproximadamente y después incrementa, mientras para la gráfica **3-9**b tiene una minimización constante de la función.

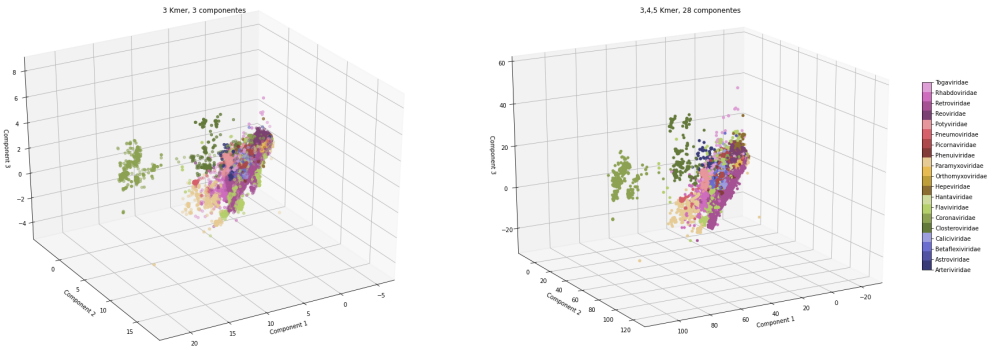
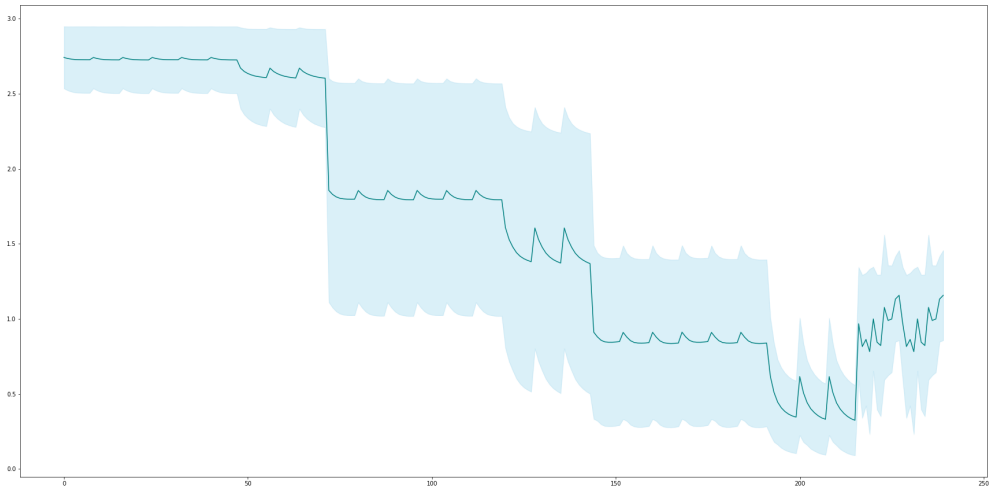


Figura 3-8.: Representación espacial 3 dimensional de los datos, a la izquierda con representación 3 *kmers* y a la derecha con representación 3,4,5 *kmers* posterior a la reducción de dimensionalidad.

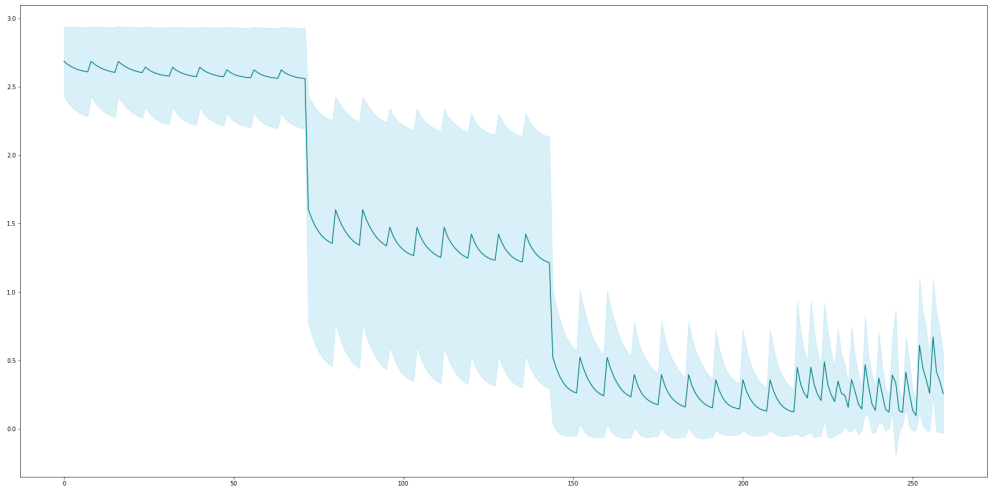
Por otro lado, en la gráfica **3-10** se observa la pérdida logarítmica del modelo de gradiente potenciado para $3kmers$ para el intervalo de 197 a 215 épocas, las cajas rojas representan el comportamiento para el conjunto de entrenamiento y las cajas azúles el comportamiento para el conjunto de prueba la región donde la función se encuentra con los valores mínimos, de los cuales se seleccionaron 2 de los modelos evaluados para la clasificación, el 247 y el 251: el primero por ser el que menor varianza tenía lo que traduce una buena precisión y el segundo por tener la mejor exactitud, teniendo en cuenta la búsqueda del mejor balance y que no siempre puede ser un solo modelo el que cumple con el balance de las dos métricas.

Para el modelo de $3, 4y5kmers$ **3-11**, el comportamiento entre épocas y entre el conjunto de datos de entrenamiento y el de prueba muestran una mayor diferencia por lo que se hizo necesario graficar un mayor número de épocas, las cuales van desde 199 a 251 épocas, de las cuales la época 247 aunque las dos cajas no se sobrelapan, presentan la desviación más pequeña y los puntos atípicos más cercanos al rango intercuartil, por lo cual es el conjunto de parámetros con mejores métricas.

De acuerdo con la matriz de confusión **3-12** para los datos de prueba, representa las 19 familias en el eje Y con las clases verdaderas de cada secuencia y en el eje X las clases que fueron predichas por el modelo, mostrando la relación de familias predichas versus familias verdaderas para cada una de las clases, teniendo en términos generales un buen desempeño del modelo con tres familias con una precisión mayor al 98 %, Coronaviridae, Orthomyxoviridae y Retroviridae, dos familias más con porcentajes superiores al 90 %, Reoviridae y Arterivi-



(a) 3 k-mers



(b) 3, 4, y 5 k-mers

Figura 3-9.: Distribución de la métrica del promedio de la pérdida logarítmica y su desviación estándar para representar el 99 % de los datos, en el eje x se representan las épocas o iteraciones del modelo y en el eje y la pérdida logarítmica.

ridae. La familia con el porcentaje de predicción más bajo fue Astroviridae con el 28 %, en la cual hubo una mayor predicción errónea sobre las familias Picornaviridae y Reoviridae, sin embargo, estas familias obtuvieron porcentajes de clasificación altos y no hubo prediccio-

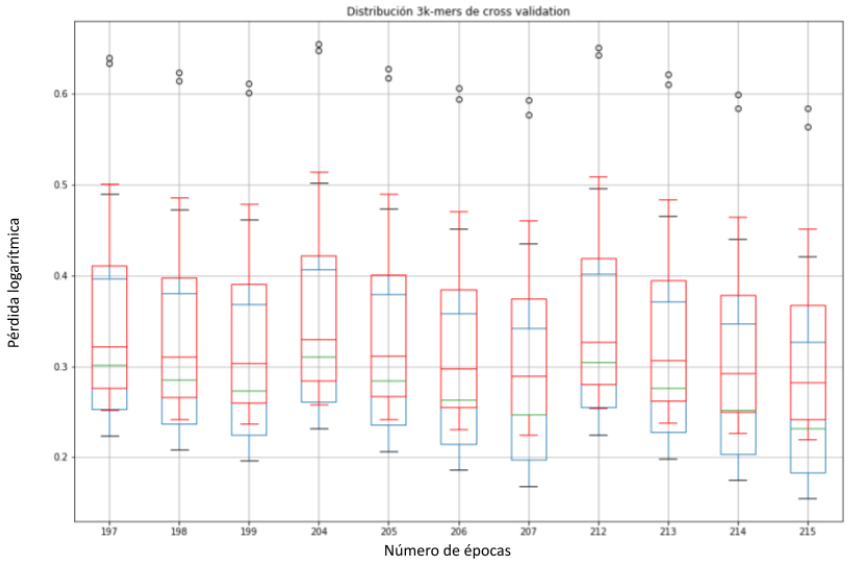


Figura 3-10.: Comportamiento de los modelos de gradiente potenciado con menor promedio de pérdida logarítmica en la validación cruzada para 3 k-mers.

nes erróneas altas en ninguna familia en especifico, otro caso interesante es el de la familia Hantaviridae que obtuvo un 40 % de clasificación correcto y 31 % de las secuencias fueron clasificadas en Orthomyxoviridae, en la cual ocurre lo mismo que en el caso anterior.

Las tres familias con el mayor número de secuencias, Orthomyxoviridae, Retroviridae y Coronaviridae, con un número de secuencias mayor a 100.000, obtuvieron las mejores predicciones por encima del 98 %, mientras que las familias con menor número de datos fueron las familias con menores predicciones.

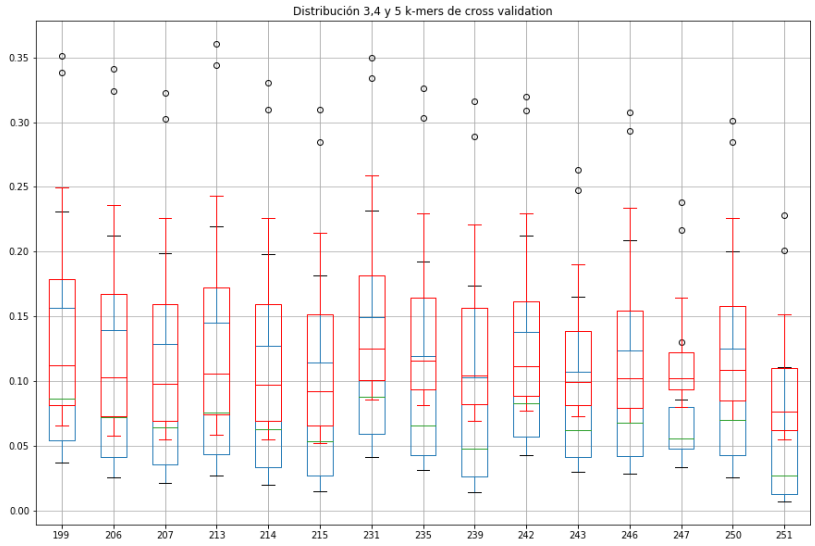


Figura 3-11.: Comportamiento de los modelos de gradiente potenciado con menor promedio de pérdida logarítmica en la validación cruzada para 3, 4 y 5 k-mers.

Máquina de soporte vectorial

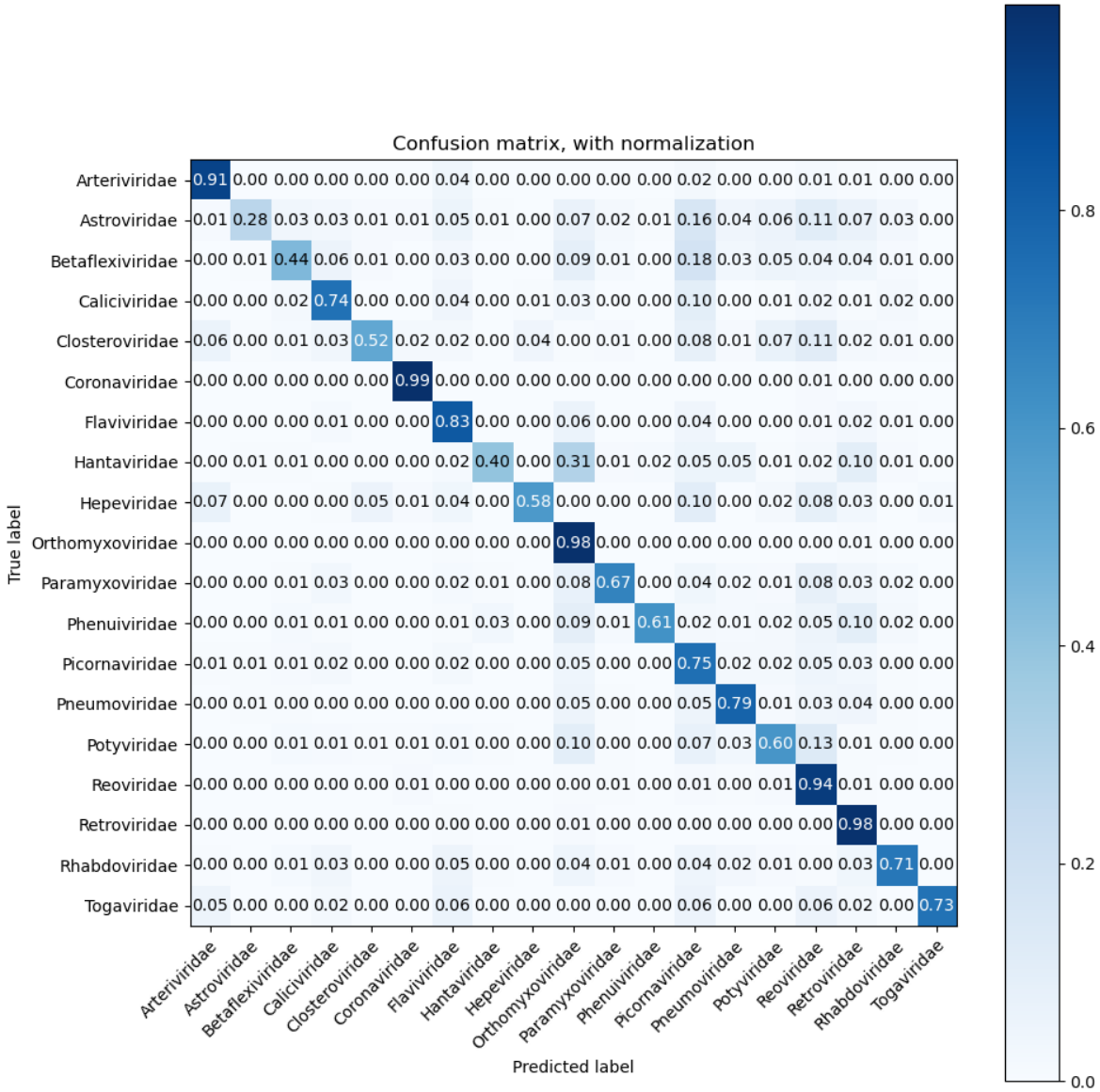


Figura 3-12.: Matriz de confusión para el conjunto de datos de prueba del modelo de gradiente potenciado para 3 k-mers

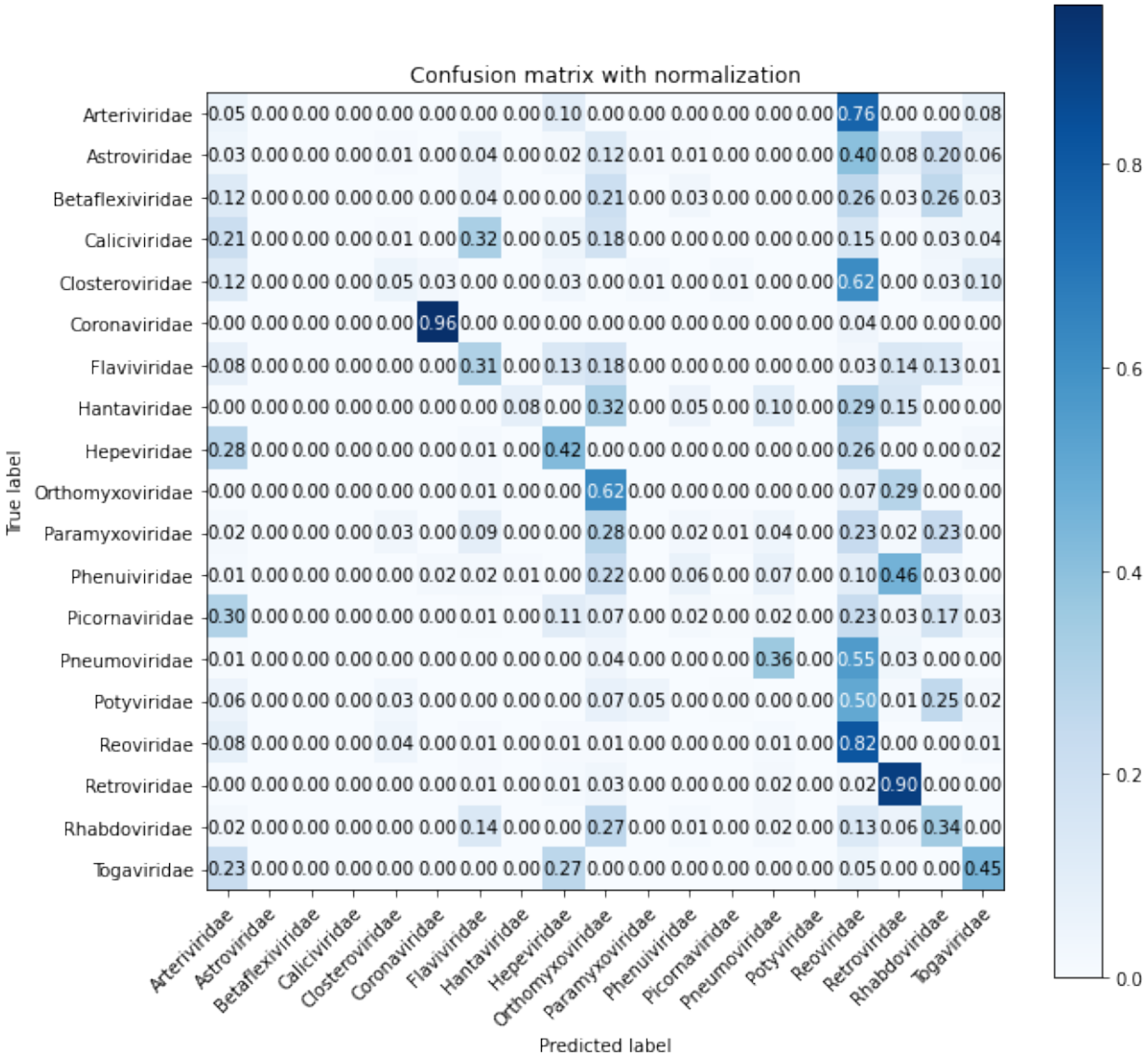


Figura 3-13.: Matriz de confusión del modelo **SVM** del entrenamiento, para 3 *kmers*

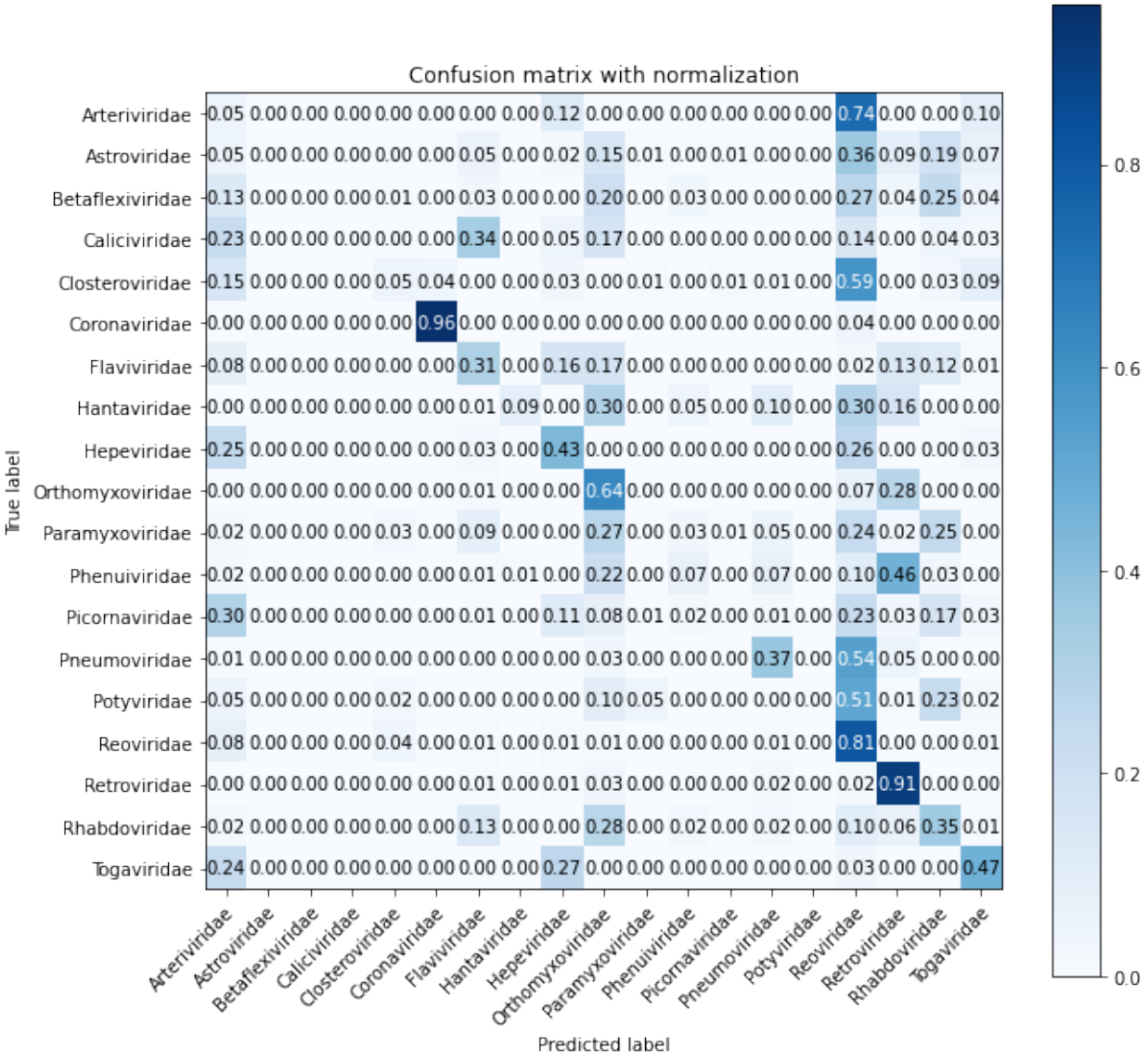


Figura 3-14.: Matriz de confusión del modelo **SVM** para los datos de prueba, para 3 k-mers

Aunque la máquina de soporte vectorial no dio los resultados esperados, teniendo en cuenta las dos matrices de confusión (entrenamiento y prueba), se puede ver que para las dos matrices se obtiene un resultado similar, lo que indica que no hay un sobreajuste del modelo. De haber sobre ajuste así se observaría una matriz de confusión de entrenamiento con porcentajes de predicción muy buenos y una matriz de del conjunto de prueba con resultados muy bajos, el cual no es el caso.

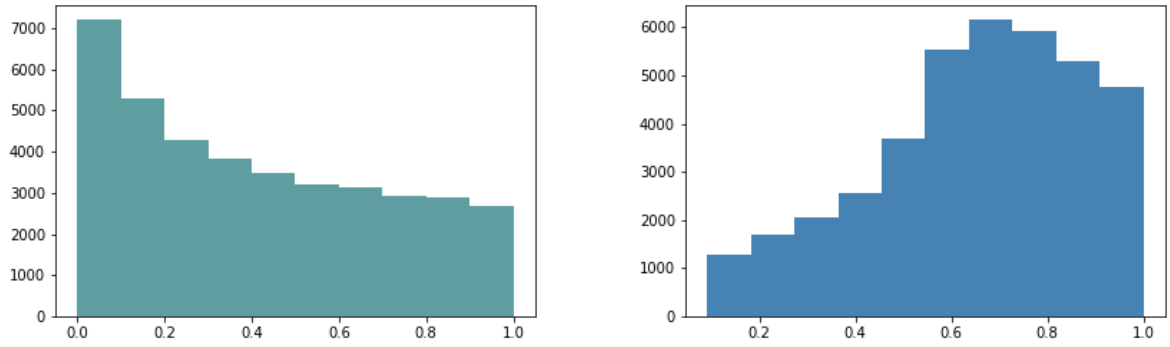
Debido a que precisamente para este modelo se realizó una búsqueda exhaustiva de los parámetros que mejor se ajustaban al modelo y que estos resultados son los mejores que se encontraron después de 5000 iteraciones, cada una con 10 particiones de los datos aleatorias,

puede ser descartada una falla en la selección de parámetros en el modelo, de acuerdo con la organización espacial en el gráfico en 3D y teniendo en cuenta como funciona el modelo, a través de la división del conjunto de datos, por división de los hiperplanos en el espacio, clases que no tienen unas fronteras fuertemente delimitadas o clases altamente dispersas tendrán bajas posibilidades de ser clasificadas en un entramado de elementos como en este caso.

Para el caso de Orthomyxoviridae, familia que con el modelo de (`Gradient boosting`) había tenido una clasificación muy alta, en este caso con 64 % de probabilidad de predicción correcta fortalece la idea de que aunque una clase tenga una gran cantidad de datos, si estos están muy dispersos y se sobrelapan en mayor proporción con otras clases no van a poder ser separados con una alta precisión por este algoritmo. Por otro lado, estos resultados permiten reconocer familias que de acuerdo a la frecuencia de codones en la secuencia, están compartiendo un espacio cercano.

La organización espacial de los datos que se puede visualizar en las gráficas de tres dimensiones **3-8**, muestran que la distribución de las clases tiene fronteras muy difusa, cerca del 90 % del entramado de puntos se encuentra concentrado en una región muy pequeña, el modelo de soporte vectorial se construyo para medir las distancias entre los puntos mediante la distancia Euclidiana, de lo cual surgen dos posibles explicaciones a los resultados que se obtuvieron, la primera, que debido a la alta dimensionalidad de los datos, 64 dimensiones para 3 *kmers* y 1344 dimensiones para 3, 4 y 5 *kmer*, al medir la distancia Euclidiana los resultados se ven afectados por la maldición de la dimensionalidad, por otro lado, los supuestos de este algoritmo incluyen una distribución normal de los datos, en el caso de que el conjunto de datos no tenga esta distribución tambien se verán afectados los resultados, finalmente, tener fronteras tan difusas hace compleja la división de las clases por medio de hiperplanos en esta dimensionalidad sin afectar el nivel de tolerancia.

De acuerdo con los datos arrojados por los dos modelos, existen diferentes hipótesis que pueden describir este comportamiento, primero que todo y el más evidente, la representación de una familia afecta directamente el desempeño del modelo al predecir, lo que quiere decir que la muestra es insuficiente para predecir con alta fidelidad las secuencias de esta familia, segundo, la base de datos puede tener conflictos entre familias que hayan sido erróneamente clasificadas y que basados en ese supuesto hemos entrenado los modelos con etiquetas que estaban erróneas, situación que ocurre a menudo en especial para este grupo de genomas, cada año la `ICTV` [58] reporta nuevas familias y cambios en la clasificación de familias que pertenecían a esta nueva familia y se había catalogado como otra, lo que ocurre con muchos de los registros de Orthomyxoviridae, que en la representación espacial 3D de los datos también se observaba que esta familia era de las mas abundantes y dispersa, tercero, realmente esta equivocación en la predicción, esta realmente demostrando que las fronteras entre fa-



(a) modelo xgboost (b) modelo SVM

Figura 3-15.: Histograma de la razón entre las dos mayores probabilidades de predicción para los dos modelos, en el eje X los valores la razón entre las dos mayores probabilidades y en el eje y la frecuencia de estos eventos.

milias víricas son difusas o se están descubriendo patrones que muestran elemento que se comparten entre estas.

3.3.3. Clasificación de los datos experimentales

La clasificación de familias en los metacontigs ensamblados de las muestras metavirómicas experimentales, para el modelo de árboles de decisión se realizó mediante el método `predict`, el cual retorna una matriz de probabilidades de tamaño $(n-secuencias, n-clases)$, en donde se encuentra la probabilidad de predicción de una secuencia para cada familia y para el modelo de soporte vectorial, que por el método estándar de predicción retorna la clase y no las probabilidades. Se necesitó el uso del método de predicción de probabilidades de `linearSVC`, con el método `predict_ proba` de la clase `calibratedclassifierCV`. El uso de las matrices de probabilidad de las predicciones permite tener una perspectiva más amplia del comportamiento del modelo respecto a nuestros datos, como también reconocer las familias que se encuentran en conflicto y que presentan problemas de predicción.

En las gráficas **3-15**, se observa el conteo de la razón entre las dos mayores probabilidades de predicción para cada modelo, donde se espera que la distribución decaiga rápidamente en un modelo con una precisión alta, en este caso, en el modelo de xgboost aunque la probabilidad sigue esta tendencia, cerca del 14 % de las secuencias tienen una razón mayor a 0,8, para el caso del modelo SVM la distribución esta corrida hacia los valores más altos, lo que está relacionado con una disminución de la precisión de los modelos.

Librería	Predicciones	Predicciones significativas	Librería	Predicciones	Predicciones significativas
SRR6019484	11768	8468	Primera	180	113
SRR6019485	8010	5879	Segunda	83	63
SRR6019486	2573	1836	Tercera	15334	10633
SRR6019487	178	143			
SRR6019488	815	563			

Tabla 3-3.: Número de predicciones para los 8 ensamblajes metavirómicos con el modelo SVM

De acuerdo con esto y teniendo en cuenta que al tener 19 clases, el valor de completa incertidumbre es 0,05, para cada modelo, se calculó la razón entre las dos familias con mayor probabilidad para cada secuencia y se filtraron las que tuvieron una razón mayor o igual a 0,8, para el modelo de árboles de decisión 5571 secuencias tienen dos familias con probabilidades muy cercanas y para el modelo de soporte vectorial 11243 secuencias de 38941 evaluadas tuvieron una razón mayor a 0,8

En las tablas **3-3** y **3-4**, están representadas las predicciones totales para cada ensamblaje, como también el número de predicciones significativas, que hace referencia a las secuencias que obtuvieron una predicción de familia con una probabilidad que tiene una diferencia alta respecto a el resto de familias, que se ha descrito anteriormente con el termino: razón.

Librería	Predicciones	Predicciones significativas	Librería	Predicciones	Predicciones significativas
SRR6019484	11768	10158	primera	180	149
SRR6019485	8010	6833	segunda	83	73
SRR6019486	2573	2234	tercera	15334	13084
SRR6019487	178	151			
SRR6019488	815	688			

Tabla 3-4.: Número de predicciones para los 8 ensamblajes metavirómicos con el modelo xgboost.

Análisis de la clasificación del metaviroma de Droshopila suzukii.

De la clasificación del metaviroma de *Drosophila suzukii* se obtuvo aproximadamente el 80 % de las predicciones significativas para el modelo de árboles de decisión y de cerca del 70 % para el modelo de soporte vectorial, teniendo en cuenta que para el primero las probabilidades fueron mayores, de estas predicciones se puede observar en la gráfica **3-16**, en la columna derecha las gráficas que muestran el número de predicciones por familia, en donde `SRR6019485`

tiene el mayor número de familias predichas, adicionalmente en la columna izquierda por medio de mapas de calor se muestran el número de secuencias que tienen predicciones de familias con una probabilidad cercana.

Se encontraron predicciones para las familias Orthomyxoviridae, Reoviridae, Arteriviridae, Retroviridae, Coronaviridae, Faviviridae y Picornaviridae de las cuales existen reportes de presencia de la familia Reoviridae, Picornaviridae y Flaviviridae las cuales concuerdan con lo hallado en [70] y en [1], tambien se encontro que Arteriviridae y Coronaviridae tambien han sido reportadas en [71].

Respecto a los virus reportados como nuevos en la publicación de referencia de estos datos [1], son caracterizados como pertenecientes a las familias Lueoviridae, Solenoviridae, Nodaviridae y a clasificaciones aun no resueltas para virus de cadena sencilla RNA en sentido positivo según el ICTV y por tanto son reportadas como *unclassified ssRNA positive-strand viruses* en el NCBI [69].

Análisis de la clasificación del metaviroma de Culex sp.

En la clasificación del metagenoma de *Culex sp.* se encontraron las familias Orthomyxoviridae, Reoviridae, Retroviridae y Arteriviridae, con mayor frecuencia en las tres librerías las dos primeras familias, que a menudo tambien presentaron clasificaciones con conflictos entre estas dos familias. Se conoce que las familias Togaviridae, Flaviviridae, Bunnyaviridae y Reoviridae continenen la mayoría de Arbovirus que causan enfermedades en animales, incluidos los humanos [10], como tambien se ha reportado la presencia de virus de las familias Orthomyxoviridae y Reoviridae en Artrópodos de la familia Culicidae y en garrapatas [72], en lagos en Australia tambien se ha encontrado la presencia de Reoviridae en *Culex sp.* [73], en cuanto a *Aedes sp.* se ha encontrado que puede ser infectado por Reoviridae (Orbivirus) [74], teniendo la capacidad de infectar tambien a *Culex sp.* [75].

Las familias que fueron encontradas en los datos experimentales coinciden con las familias reportadas en diversos estudios en su mayoría moleculares, lo que permite tener una certeza de que estás familias se encuentran dentro del rango conocido por infectar estos tipos de Artrópodos, sin embargo, tambien existen diferentes familias que han sido reportadas y han sido caracterizadas por métodos moleculares que no se encuentran en nuestros hallazgos, esto puede deberse a que en la base de datos GoldStandard que curamos, únicamente tuvimos referencia de 19 familias, para las familias que no están incluidas existe una limitante en este tipo de modelo de clasificación.

3.4. Conclusiones

- Familias que obtienen un bajo desempeño en la clasificación de los modelos no necesariamente están involucradas con un mal desempeño general del modelo, por el contrario, puede deberse a causas multifactoriales, como que las características individuales de dicha familia no están siendo cuantificadas dentro de la extracción de características, que la diversidad de la familia es muy amplia o que la clasificación de las secuencias en la base de datos es errónea, por lo cual, por su naturaleza no debe atribuirse la responsabilidad única al modelo, mas bien, se propone una evaluación de la familia desde un enfoque diferente, en donde se analicen sus particularidades.

- La validación cruzada parámetros para el entrenamiento **3-1** para los dos conjuntos de datos y la posterior evaluación de los modelos de *gradient boost* muestran de acuerdo a la gráfica de 3 kmers **3-9**a y mas evidente para 3, 4, 5 kmers **3-9**b, que el cambio en la tasa de aprendizaje (*learning rate*) genera el cambio más significativo en el desempeño del modelo de acuerdo a las métricas de la función de pérdida (*(*log loss)) y el error mínimo (*m error*), tambien se observa que localmente la proporción de subconjuntos de columnas seleccionadas favorece a la minimización de las dos funciones, encontrando que en cuanto al número de subconjuntos de filas y la profundidad máxima no tienen un cambio significativo en el desempeño del modelo.

- Por otro lado, las familias Orthomyxoviridae y Reoviridae, que tienen un alto número de representación, también tuvieron un alto porcentaje de secuencias que se predijeron intercambiando estas dos familias, caso que se repitió en todas las muestras, siendo también las familias más frecuentes en las muestras.

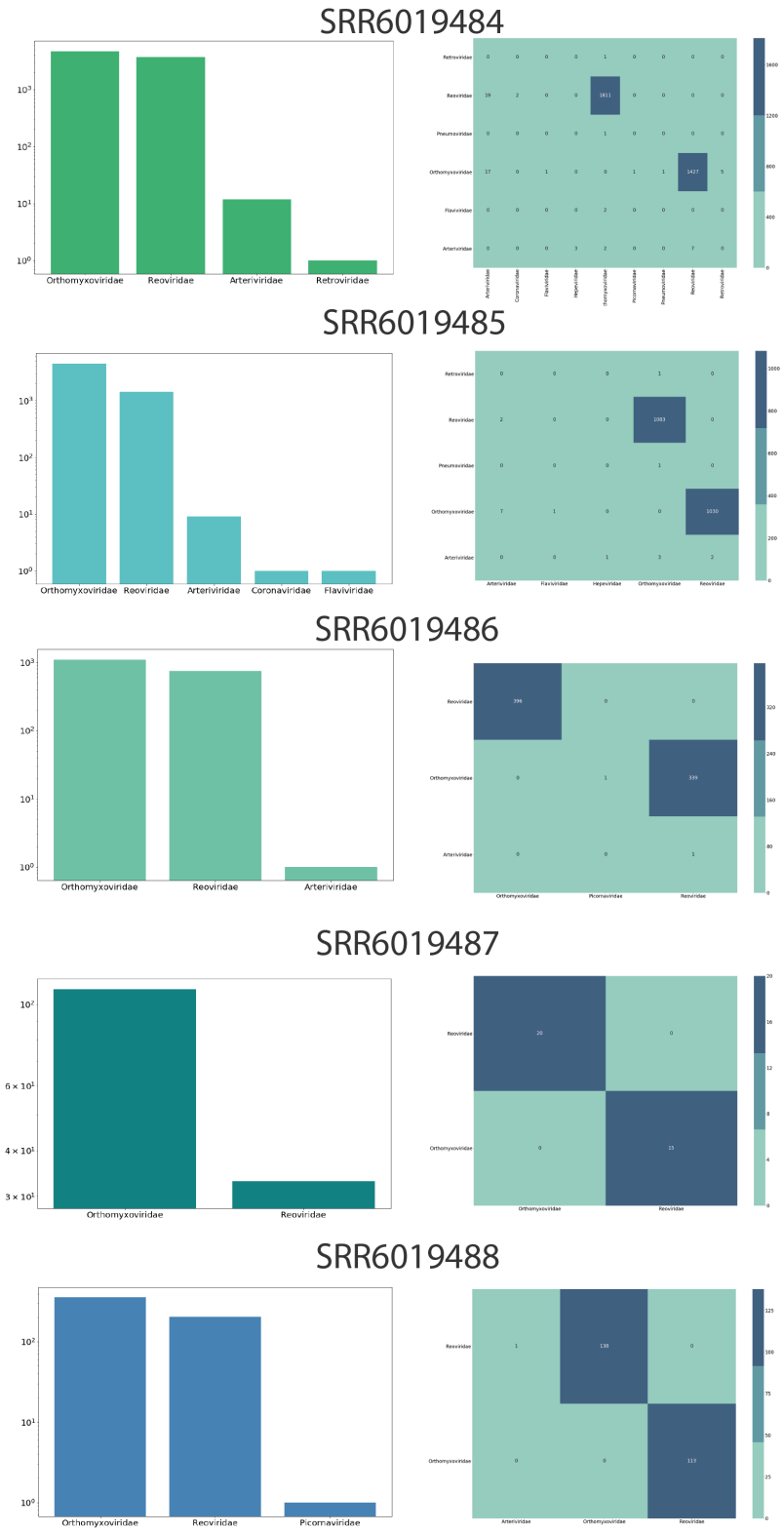


Figura 3-16.: A la izquierda, las gráficas para cada corrida con la frecuencia de las familias predichas con una razón baja y a la derecha, los mapas de calor de las familias con una razón alta para los datos de *Drosohpila suzukii*.

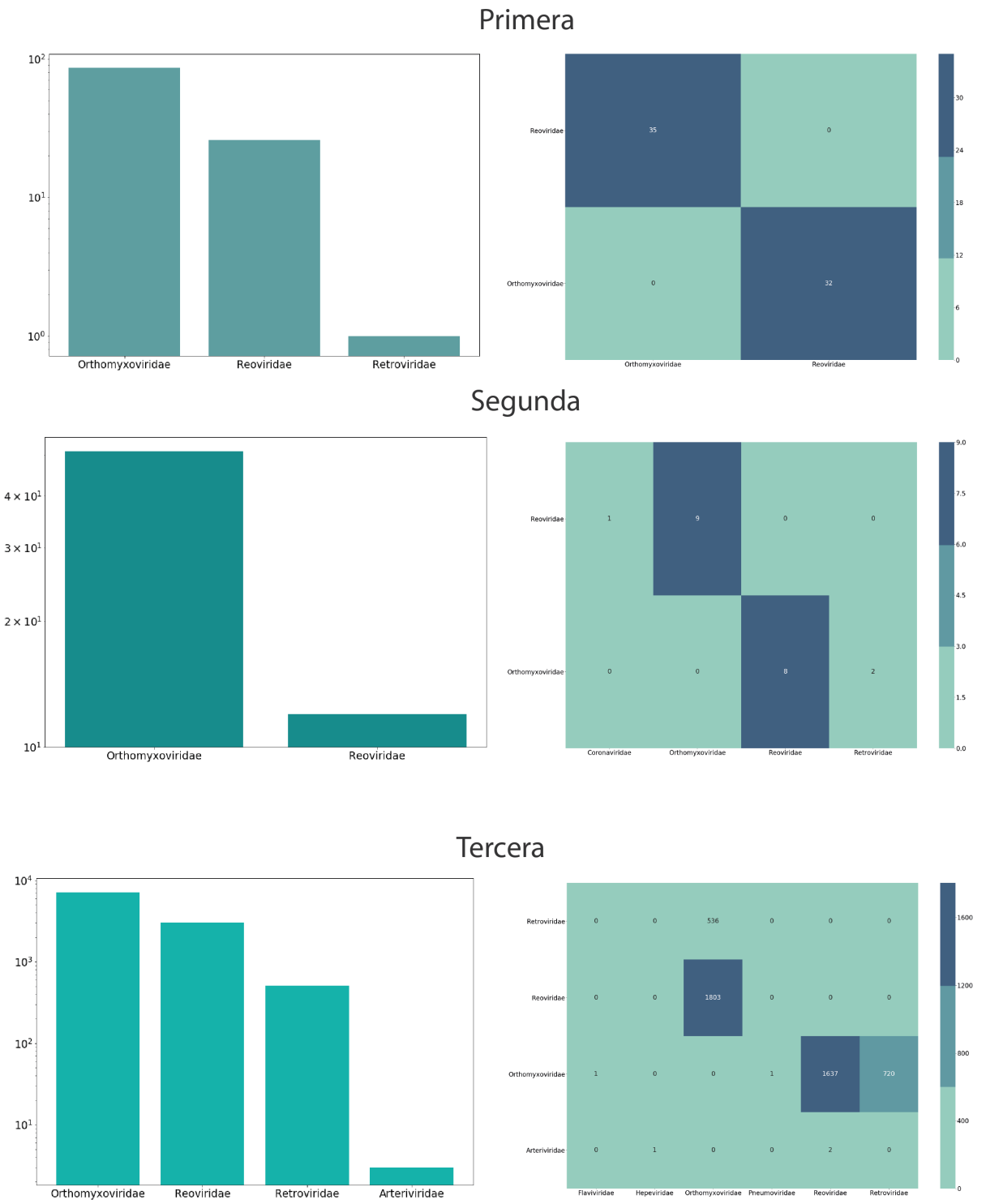


Figura 3-17.: A la izquierda, los histogramas para cada corrida con la frecuencia de las familias predichas con una razón baja y a la derecha, los mapas de calor de las familias con una razón alta para los datos experimentales.

4. Conclusiones

- El flujo metodológico para la limpieza y ensamblaje de muestras metavirómicas, fue desarrollado con herramientas bioinformáticas que se adaptan a las complejidades y limitaciones de los datos metavirómicos, el cual se diseño con protocolos estrictos para eliminar elementos contaminados del ensamblaje, por otra parte, se logró hacer un análisis adicional del tipo de contaminantes que fueron encontrados en la muestra, permitiendo tener un panorama general del contexto en el que fue recolectada la muestra.

  Por tanto, se obtuvo un flujo metodológico que se adapta a el tipo de muestras virómicas, obteniendo la mayor cantidad de información y de calidad del ensamblaje, teniendo además una perspectiva del entorno de la muestra.

- Obtener una probabilidad como resultado, a diferencia de una familia únicamente, permite reconocer que las clasificaciones tienen una incertidumbre, es importante recalcar esto, ya que en biologíaa muchas veces olvidamos que estas predicciones dependen de una gran cantidad de diferentes variables, como los fenómenos e interacciones en la naturaleza, las limitaciones de los métodos de secuenciamiento o el direccionamiento de los muestreos que no permiten que sea una muestra aleatoria, los modelos están determinados por la historia de los datos que no podemos borrar, por eso es importante señalar que los resultados obtenidos son candidatos con un nivel de incertidumbre variable y que en algunos casos en que esta crece serán más de un posible candidato.

  De acuerdo a esto, se clasificaron las muestras metavirómicas y se logro determinar que familias se encontraban dentro de las muestras con un intervalo de confianza que da una mayor validez al modelo.

- Aunque los modelos de aprendizaje de máquina están limitados al volumen de los datos y al desbalance de sus clases. Teniendo en cuenta que la búsqueda por homología no puede considerarse en el área de la virómica, la implementación de estos modelos permite tener un acercamiento más acertado a la clasificación de secuencias.

- Aunque este trabajo muestra una pequeña parte de todo el potencial que tienen las máquinas de aprendizaje en el área de la clasificación virologica constituye un punto de partida para el desarrollo de modelos con requerimientos computacionales mas demandantes, pues el análisis y uso de esta cantidad de información, los cálculos que deben realizarse y las iteraciones para que converjan los modelos tienen una demanda computacional muy alta.

4.1. Recomendaciones

- La revisión exhaustiva de las bases de datos de secuencias genómicas, es primordial ya que estas contienen secuencias duplicadas clasificadas en diferentes familias, duplicaciones dentro de las familias o secuencias con un nivel alto de incertidumbre en las bases nucleótidas, lo que termina siendo determinante en un proceso de entrenamiento.

- Debido a que constantemente se están descubriendo nuevas familias víricas y se está reevaluando la clasificación de muchas otras, se hace necesario el análisis de métodos no supervisados que den señales de como se están agrupando este entramado de información genómica.

A. Anexo: Control de calidad

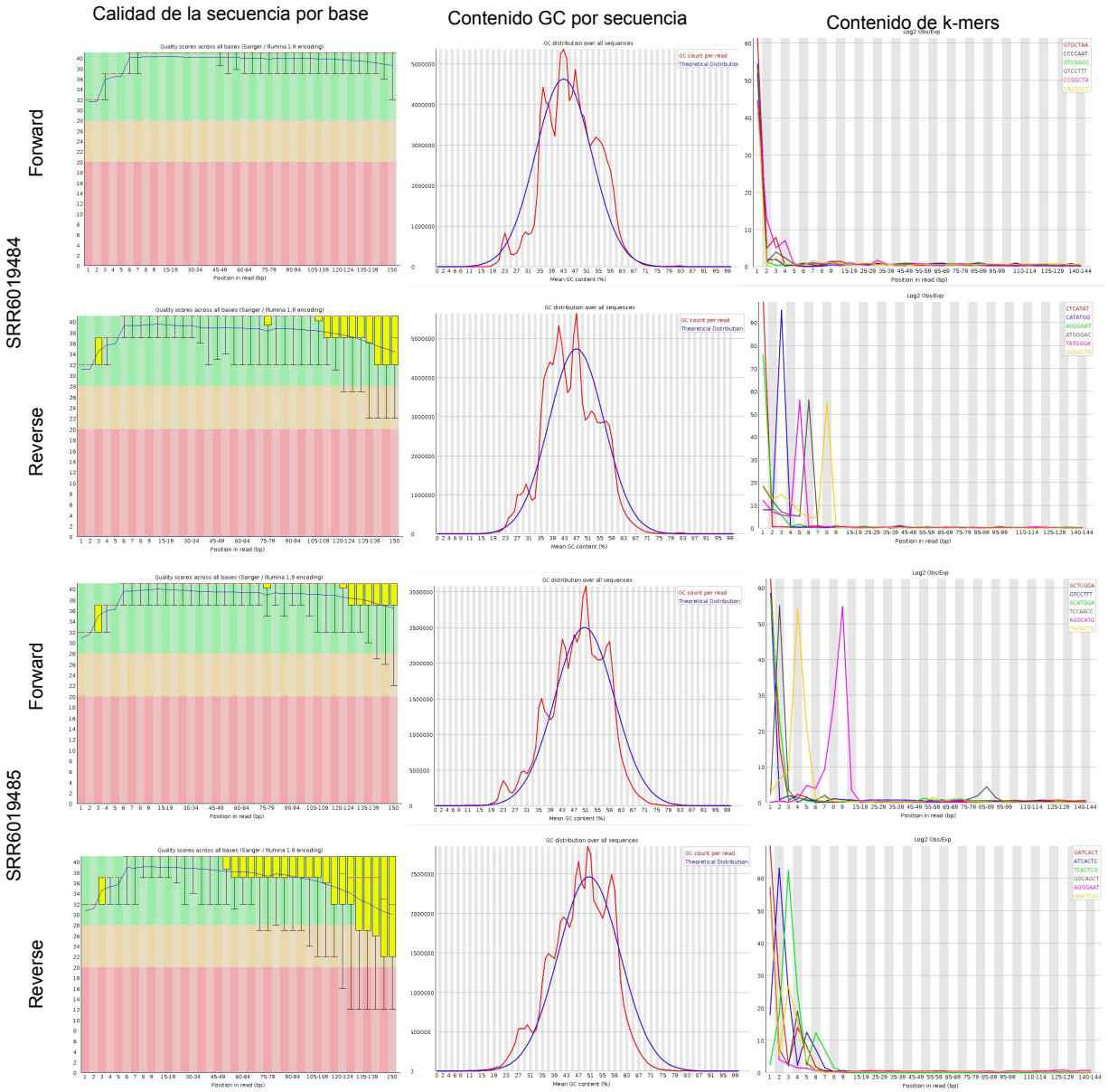


Figura A-1.: Control de calidad librerías SRR6019484, SRR6019485

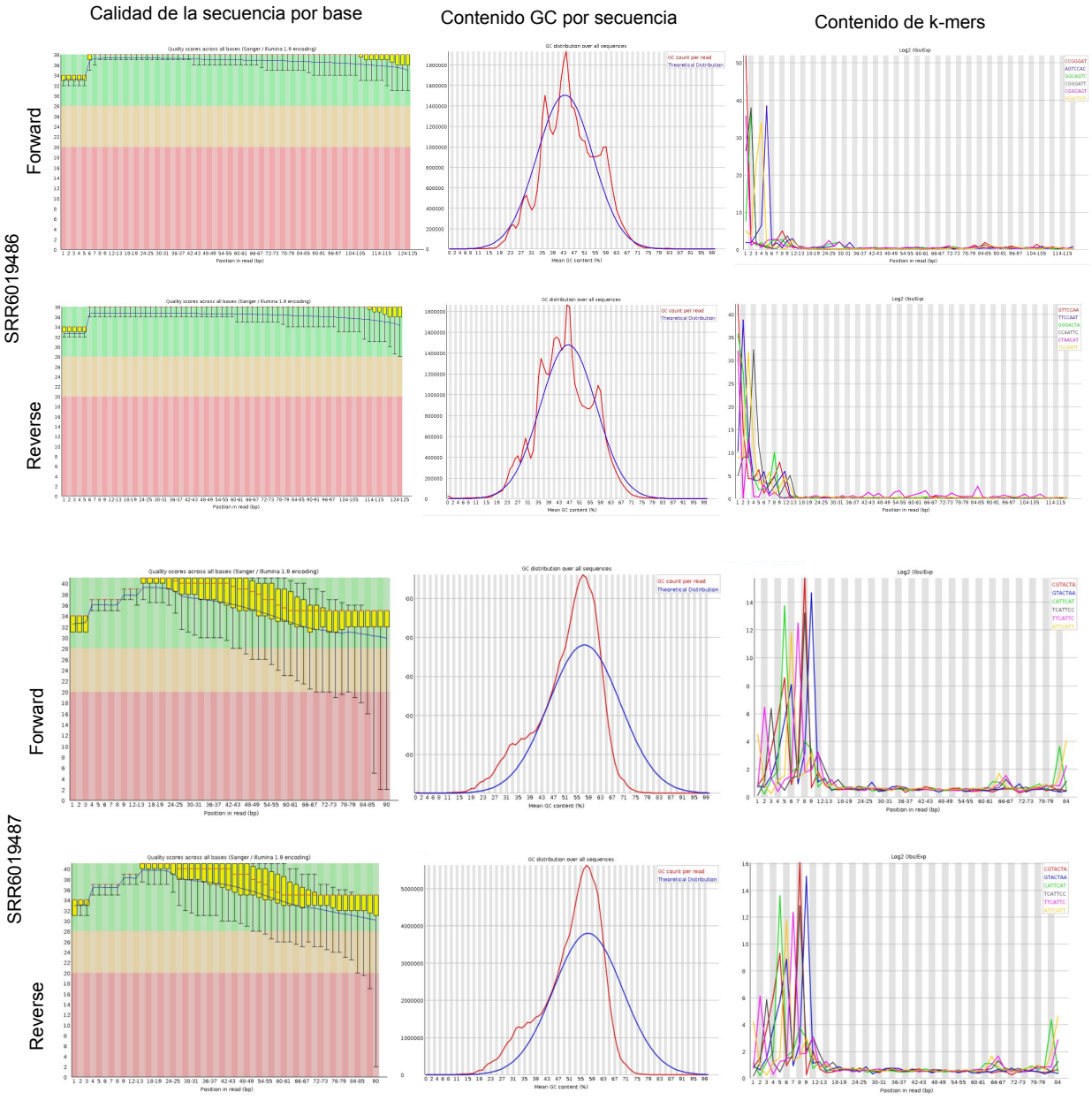


Figura A-2.: Control de calidad librerías SRR6019486, SRR6019487

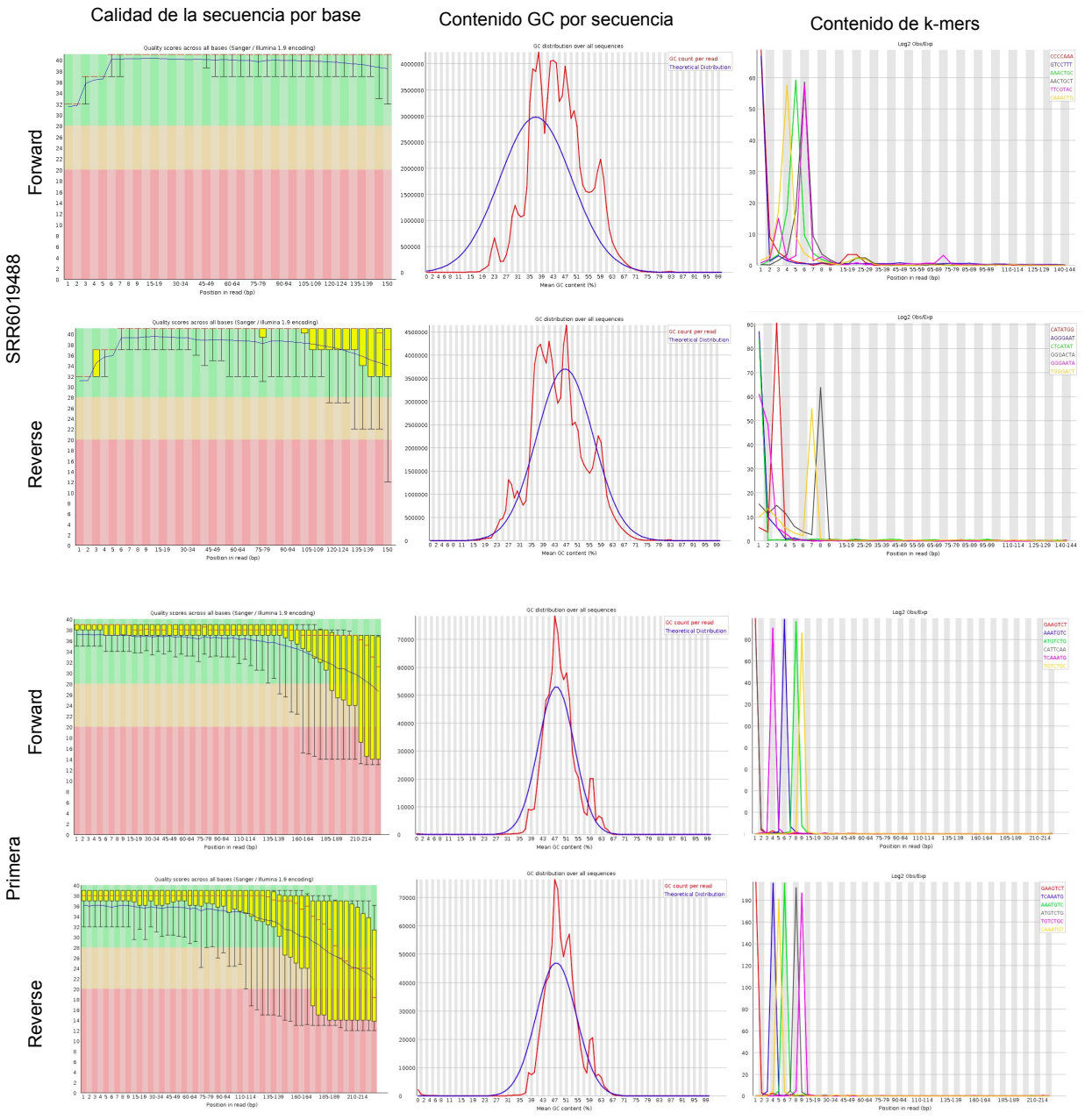


Figura A-3.: Control de calidad librerías SRR60194888, Primera

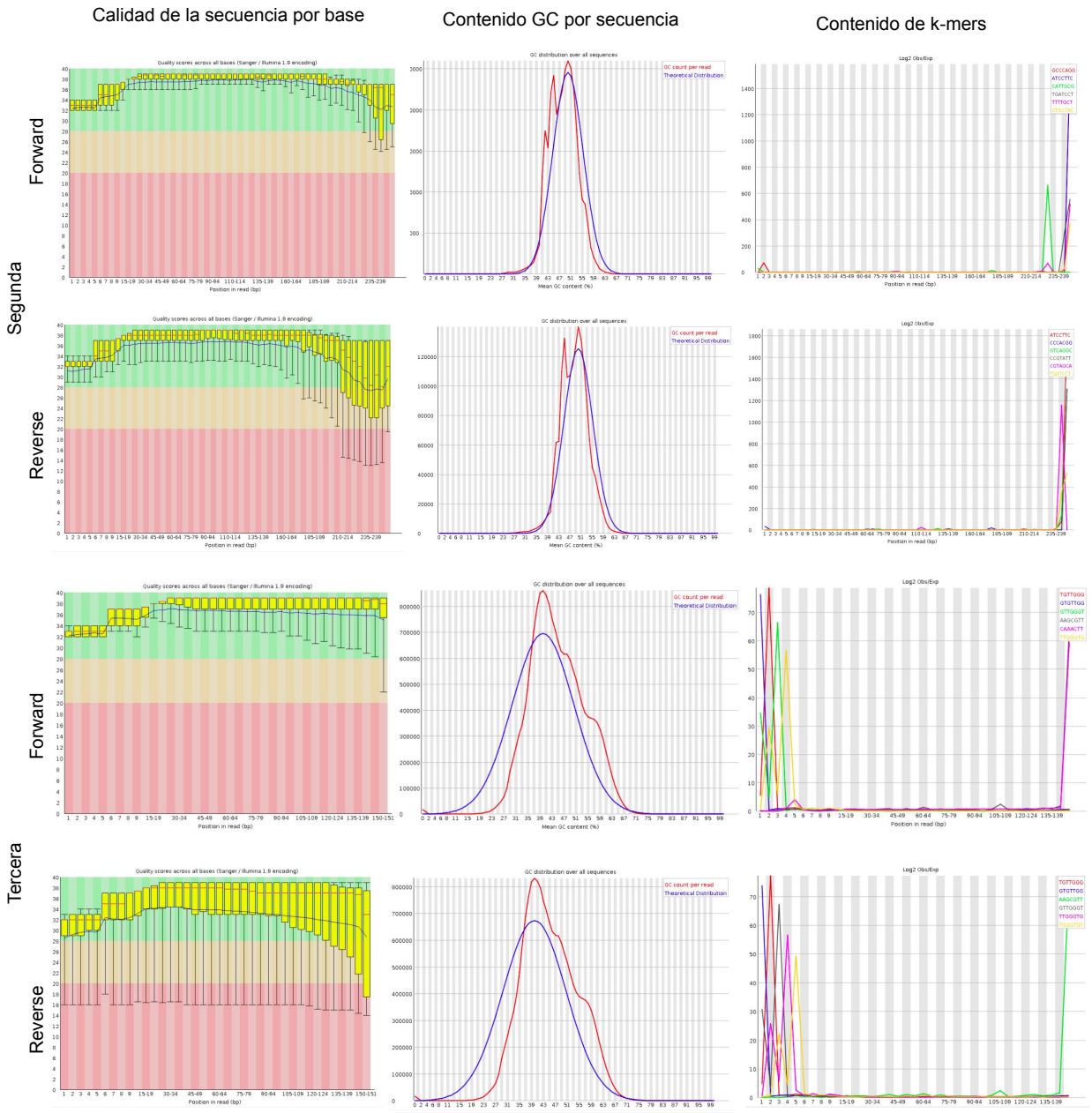


Figura A-4.: Control de calidad librerías Segunda, Tercera

Bibliografía

[1] Nathan C Medd, Simon Fellous, Fergal M Waldron, Anne Xuéreb, Madoka Nakai, Jerry V Cross, and Darren J Obbard. The virome of Drosophila suzukii, an invasive pest of soft fruit. *Virus Evolution*, 4(1), 03 2018. vey009.

[2] Alice. Lustig and Arnold j. Levine. one hundred years of virology. *Journal of virology*, 66(8):4629–4631, 08 1992.

[3] Boriana Marintcheva. Chapter 1 - introduction to viral structure, diversity and biology-parts of this chapter were originally published in marintcheva b. a box of paradoxes: the fascinating world of viruses. In Boriana Marintcheva, editor, *Harnessing the Power of Viruses*, pages 1 – 26. Academic Press, 2018.

[4] Manja Marz, Niko Beerenwinkel, Christian Drosten, Markus Fricke, Dmitrij Frishman, Ivo L. Hofacker, Dieter Hoffmann, Martin Middendorf, Thomas Rattei, Peter F. Stadler, and Armin Töpfer. Challenges in RNA virus bioinformatics. *Bioinformatics*, 30(13):1793–1799, 03 2014.

[5] Curtis A. Suttle. Marine viruses – major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812, 2007.

[6] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 03 2006.

[7] Yuri I. Wolf, Darius Kazlauskas, Jaime Iranzo, Adriana LucÍa-Sanz, Jens H. Kuhn, Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin. Origins and evolution of the global rna virome. *bioRxiv*, 2018.

[8] Men-Bao Qian and Xiao-Nong Zhou. Global burden on neglected tropical diseases. *The Lancet Infectious Diseases*, 16(10):1113–1114, Oct 2016.

[9] Peter Daszak Simon A. Levin Oscar Patterson-Lomba Charles Perrings George Poste Sherry Towers Carlos Castillo-Chavez, Roy Curtiss. Beyond ebola: lessons to mitigate future pandemics. 3(7):PE354–E355, July 2015.

[10] Guodong Liang, Xiaoyan Gao, and Ernest A Gould. Factors responsible for the emergence of arboviruses; strategies, challenges and limitations for their control. *Emerging Microbes & Infections*, 4(1):1–5, 2015. PMID: 26038768.

[11] Goro Kuno and Gwong-Jen J. Chang. Biological transmission of arboviruses: Reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends. *Clinical Microbiology Reviews*, 18(4):608–637, 2005.

[12] Cameron P. Simmons, Jeremy J. Farrar, Nguyen van Vinh Chau, and Bridget Wills. Dengue. *New England Journal of Medicine*, 366(15):1423–1432, 2012. PMID: 22494122.

[13] S C Weaver. Evolutionary influences in arboviral disease. *Current topics in microbiology and immunology*, 299:285–314, 2006.

[14] Crosskey RW. *Old tools and new taxonomic problems in blood-sucking insects in: Biosystematics of Haematophagous Insects.* Clarendon Press, Oxford, UK, 1988.

[15] Peter Simmonds, Paul Becher, Jens Bukh, Ernest A. Gould, Gregor Meyers, Tom Monath, Scott Muerhoff, Alexander Pletnev, Rebecca Rico-Hesse, Donald B. Smith, Jack T. Stapleton, and ICTV Report Consortium. Ictv virus taxonomy profile: Flaviviridae. *Journal of General Virology*, 98(1):2–3, 2017.

[16] Rubing Chen, Suchetana Mukhopadhyay, Andres Merits, Bethany Bolling, Farooq Nasar, Lark L. Coffey, Ann Powers, Scott C. Weaver, and ICTV Report Consortium. Ictv virus taxonomy profile: Togaviridae. *Journal of General Virology*, 99(6):761–762, 2018.

[17] International Committee on Taxonomy of Viruses (ICTV). Ictv 9th report (2011) bunyaviridae (html), bunya: from bunyamwera, place in uganda, where type virus was isolated. `https://talk.ictvonline.org/ictv-reports/ictv_9th_report/negative-sense-rna-viruses-2011/w/negrna_viruses/205/bunyaviridae`. Accessed: 2020-04-14.

[18] Peter J. Walker, Kim R. Blasdell, Charles H. Calisher, Ralf G. Dietzgen, Hideki Kondo, Gael Kurath, Ben Longdon, David M. Stone, Robert B. Tesh, Noël Tordo, Nikos Vasilakis, Anna E. Whitfield, and ICTV Report Consortium. Ictv virus taxonomy profile: Rhabdoviridae. *Journal of General Virology*, 99(4):447–448, 2018.

[19] Viral zone. Reoviridae report. `https://viralzone.expasy.org/104?outline=all_by_species`. Accessed: 2020-04-14.

[20] Ana Valeria Bussetti, Gustavo Palacios, Amelia Travassos da Rosa, Nazir Savji, Komal Jain, Hilda Guzman, Stephen Hutchison, Vsevolod L. Popov, Robert B. Tesh, and W. Ian Lipkin. Genomic and antigenic characterization of jos virus. *Journal of General Virology*, 93(2):293–298, 2012.

[21] Covadonga Alonso, Manuel Borca, Linda Dixon, Yolanda Revilla, Fernando Rodriguez, Jose M. Escribano, and ICTV Report Consortium. Ictv virus taxonomy profile: Asfarviridae. *Journal of General Virology*, 99(5):613–614, 2018.

[22] Alexander T Ciota and Laura D Kramer. Insights into arbovirus evolution and adaptation from experimental studies. *Viruses*, 2(12):2594–2617, 12 2010.

[23] Dilip K. Nag, Matthew Brecher, and Laura D. Kramer. Dna forms of arboviral rna genomes are generated following infection in mosquito cell cultures. *Virology*, 498:164 – 171, 2016.

[24] Cordaux R Gilbert C. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Current Opinion in Virology*, 25:16–22, Aug 2017.

[25] Rebecca Rose, Bede Constantinides, Avraam Tapinos, David L Robertson, and Mattia Prosperi. Challenges in the analysis of viral metagenomes. *Virus Evolution*, 2(2), 08 2016. vew022.

[26] Maureen A. O'Malley. *Metagenome*, pages 1283–1283. Springer New York, New York, NY, 2013.

[27] Evans Atoni, Yujuan Wang, Samuel Karungu, Cecilia Waruhiu, Ali Zohaib, Vincent Obanda, Bernard Agwanda, Morris Mutua, Han Xia, and Zhiming Yuan. Metagenomic virome analysis of culex mosquitoes from kenya and china. 10(1):30, Jan 2018. 29329230[pmid].

[28] Pengpeng Xiao, Chenghui Li, Ying Zhang, Jicheng Han, Xiaofang Guo, Lv Xie, Mingyao Tian, Yiquan Li, Maopeng Wang, Hao Liu, Jingqiang Ren, Hongning Zhou, Huijun Lu, and Ningyi Jin. Metagenomic sequencing from mosquitoes in china reveals a variety of insect and human viruses. *Frontiers in Cellular and Infection Microbiology*, 8:364, 2018.

[29] Mohammadreza Sadeghi, Eda Altan, Xutao Deng, Christopher M. Barker, Ying Fang, Lark L. Coffey, and Eric Delwart. Virome of ¿12 thousand culex mosquitoes from throughout california. *Virology*, 523:74 – 88, 2018.

[30] W B Whitman, D C Coleman, and W J Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583, 06 1998.

[31] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, Jun 2012. 22546560[pmid].

[32] Simon Roux, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. Virsorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, May 2015.

[33] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3–3, Feb 2012.

[34] Saskia L Smits, Rogier Bodewes, Aritz Ruiz-González, Wolfgang Baumgärtner, Marion P Koopmans, Albert D M E Osterhaus, and Anita C Schürch. Recovering full-length viral genomes from metagenomes. *Frontiers in microbiology*, 6:1069–1069, 10 2015.

[35] Xingyu Liao, Min Li, You Zou, Fang-Xiang Wu, Yi-Pan, and Jianxin Wang. Current challenges and solutions of de novo assembly. *Quantitative Biology*, 7(2):90–109, Jun 2019.

[36] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991, 2011.

[37] Francesca Di Giallonardo, Armin Töpfer, Melanie Rey, Sandhya Prabhakaran, Yannick Duport, Christine Leemann, Stefan Schmutz, Nottania K. Campbell, Beda Joos, Maria Rita Lecca, Andrea Patrignani, Martin Däumer, Christian Beisel, Peter Rusert, Alexandra Trkola, Huldrych F. Günthard, Volker Roth, Niko Beerenwinkel, and Karin J. Metzner. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic acids research*, 42(14):e115–e115, Aug 2014.

[38] Nadim J. Ajami, Matthew C. Wong, Matthew C. Ross, Richard E. Lloyd, and Joseph F. Petrosino. Maximal viral information recovery from sequence data using virmap. *Nature Communications*, 9(1):3205, 2018.

[39] Gregory; Smyth Padhraic Fayyad, Usama; Piatetsky-Shapiro. From data mining to knowledge discovery in databases. *AI Magazine*, pages 35–54, Fall 1996.

[40] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData mining*, 10:35–35, 12 2017.

[41] Qian Zhou, Xiaoquan Su, and Kang Ning. Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports*, 4(1):6957, Nov 2014.

[42] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, 07 2009.

[43] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–W37, 07 2011.

[44] Qian Zhou, Xiaoquan Su, Gongchao Jing, Songlin Chen, and Kang Ning. Rna-qc-chain: comprehensive and fast quality control for rna-seq data. *BMC Genomics*, 19(1):144, 2018.

[45] Sejal Modha, Joseph Hughes, Giovanni Bianco, Heather M Ferguson, Barbara Helm, Lily Tong, Gavin S Wilkie, Alain Kohl, and Esther Schnettler. Metaviromics reveals unknown viral diversity in the biting midge culicoides impunctatus. *Viruses*, 11(9):865, 09 2019.

[46] Coverage depth recommendations. `https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html`. Acceso: 2021-12-14.

[47] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 01 2015.

[48] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012.

[49] Lohse M. Usadel B. Bolger, A. M. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 2014.

[50] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, Mar 2012.

[51] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 11 2012.

[52] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–745, Jan 2016.

[53] B. J. Matthews, O. Dudchenko, S. B. Kingan, S. Koren, I. Antoshechkin, J. E. Crawford, W. J. Glassford, M. Herre, S. N. Redmond, N. H. Rose, G. D. Weedall, Y. Wu, S. S. Batra, C. A. Brito-Sierra, S. D. Buckingham, C. L. Campbell, S. Chan, E. Cox, B. R. Evans, T. Fansiri, I. Filipović, A. Fontaine, A. Gloria-Soria, R. Hall, V. S. Joardar, A. K. Jones, R. G. G. Kay, V. K. Kodali, J. Lee, G. J. Lycett, S. N. Mitchell, J. Muehling,

M. R. Murphy, A. D. Omer, F. A. Partridge, P. Peluso, A. P. Aiden, V. Ramasamy, G. Rašić, S. Roy, K. Saavedra-Rodriguez, S. Sharan, A. Sharma, M. L. Smith, J. Turner, A. M. Weakley, Z. Zhao, O. S. Akbari, W. C. Black, H. Cao, A. C. Darby, C. A. Hill, J. S. Johnston, T. D. Murphy, A. S. Raikhel, D. B. Sattelle, I. V. Sharakhov, B. J. White, L. Zhao, E. L. Aiden, R. S. Mann, L. Lambrechts, J. R. Powell, M. V. Sharakhova, Z. Tu, H. M. Robertson, C. S. McBride, A. R. Hastie, J. Korlach, D. E. Neafsey, A. M. Phillippy, and L. B. Vosshall. Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature*, 563(7732):501–507, 11 2018.

[54] G. N. Artemov, A. N. Peery, X. Jiang, Z. Tu, V. N. Stegniy, M. V. Sharakhova, and I. V. Sharakhov. The Physical Genome Mapping of Anopheles albimanus Corrected Scaffold Misassemblies and Identified Interarm Rearrangements in Genus Anopheles. *G3 (Bethesda)*, 7(1):155–164, 01 2017.

[55] V. Dritsou, P. Topalis, N. Windbichler, A. Simoni, A. Hall, D. Lawson, M. Hinsley, D. Hughes, V. Napolioni, F. Crucianelli, E. Deligianni, G. Gasperi, L. M. Gomulski, G. Savini, M. Manni, F. Scolari, A. R. Malacrida, B. Arcà, J. M. Ribeiro, F. Lombardo, G. Saccone, M. Salvemini, R. Moretti, G. Aprea, M. Calvitti, M. Picciolini, P. A. Papathanos, R. Spaccapelo, G. Favia, A. Crisanti, and C. Louis. A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus. *Pathog Glob Health*, 109(5):207–220, Jul 2015.

[56] Alexander Bowe, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Succinct de bruijn graphs. In Ben Raphael and Jijun Tang, editors, *Algorithms in Bioinformatics*, pages 225–235, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[57] Moreno Zolfo, Federica Pinto, Francesco Asnicar, Paolo Manghi, Adrian Tett, Frederic D. Bushman, and Nicola Segata. Detecting contamination in viromes using viromeqc. *Nature Biotechnology*, 37(12):1408–1412, Dec 2019.

[58] Peter J. Walker, Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Evelien M. Adriaenssens, Donald M. Dempsey, Bas E. Dutilh, Balázs Harrach, Robert L. Harrison, R. Curtis Hendrickson, Sandra Junglen, Nick J. Knowles, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, Max Nibert, Richard J. Orton, Luisa Rubino, Sead Sabanadzovic, Peter Simmonds, Donald B. Smith, Arvind Varsani, Francisco Murilo Zerbini, and Andrew J. Davison. Changes to virus taxonomy and the statutes ratified by the international committee on taxonomy of viruses (2020). *Archives of Virology*, 165(11):2737–2748, Nov 2020.

[59] Arun Jagota. *Data analysis and classification for Bioinformatics*. Bioinformatics by the Bay, 2000.

[60] N.G Nguyen, V.A Tran, D.L Ngo, D. Phan, F.R Lumbaranja, and M.R et al. Faizal. Dna sequence classification by convolutional neural network. *Biomedical Science and Engineering*, 2016.

[61] Pierre Baldi, Søren Brunak, and Francis Bach. *Bioinformatics: the machine learning approach*. MIT press, 2001.

[62] Yanqing Zhang and Jagath C Rajapakse. *Machine learning in bioinformatics*, volume 4. John Wiley & Sons, 2009.

[63] Vladimir Vapnik. *The Nature of Statistical Learning*, volume 1. Springer-verlag, 1995.

[64] Chih-Jen Lin Chih-Chung Chang. *LIBSVM: A Library for Support Vector Machines*. Department of Computer Science, National Taiwan University, Taipei, Taiwan, Initial version: 2001 Last updated: January 20, 2021.

[65] Christopher JC Burges, Bernhard Scholkopf, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press Cambridge, MA, USA:, 1999.

[66] El Baúl Del Programador. creating-trees-dependency-graphs-svms-in-tikz. `https://elbauldelprogramador.com/en/creating-trees-dependency-graphs-svms-in-tikz/`, November 2019.

[67] Xgboost. `https://www.nvidia.com/en-us/glossary/data-science/xgboost/`. Acceso: 2021-12-14.

[68] Joaquín Amat Rodrigo. Gradient Boosting con Python octubre 2020. `https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html`. Accessed: 2021-02-18.

[69] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova. NCBI viral genomes resource. *Nucleic Acids Res*, 43(Database issue):D571–577, Jan 2015.

[70] Claire L. Webster, Ben Longdon, Samuel H. Lewis, and Darren J. Obbard. Twenty-five new viruses associated with the drosophilidae (diptera). *Evolutionary bioinformatics online*, 12(Suppl 2):13–25, Jun 2016. 27375356[pmid].

[71] Florian Zirkel, Andreas Kurth, Phenix-Lan Quan, Thomas Briese, Heinz Ellerbrok, Georg Pauli, Fabian H. Leendertz, W. Ian Lipkin, John Ziebuhr, Christian Drosten, Sandra Junglen, and Michael J. Buchmeier. An insect nidovirus emerging from a primary tropical rainforest. *mBio*, 2(3):e00077–11, 2011.

[72] Gustavo Fermin. Host range, host–virus interactions, and virus transmission. *Viruses*, pages 101–134, 2018. PMC7173471[pmcid].

[73] Chris Cowled, Gustavo Palacios, Lorna Melville, Richard Weir, Susan Walsh, Steven Davis, Aneta Gubala, W. Ian Lipkin, Thomas Briese, and David Boyle. Genetic and epidemiological characterization of stretch lagoon orbivirus, a novel orbivirus isolated from culex and aedes mosquitoes in northern australia. *The Journal of general virology*, 90(Pt 6):1433–1439, Jun 2009. 19282430[pmid].

[74] Hurt SL Dunphy BM Smith RC Bartholomay LC-Blitvich BJ. Tangudu CS, Charles J. Skunk river virus, a novel orbivirus isolated from aedes trivittatus in the united states. *The Journal of general virology*, 100(2):295–300, Feb 2019.

[75] Belhouchet M Aldrovandi N Tao S Chen B-Liang G Tesh RB de Micco P de Lamballerie X. Attoui H, Jaafar FM. a new orbivirus species isolated from culex tritaeniorhynchus mosquitoes in china. *The Journal in General Virology*, 86(12):3409–3417, Dec 2005.