UNIVERSITAT ROVIRA i VIRGILI

# ANÁLISIS DE MUTACIONES Y VARIABILIDAD GENÓMICA EN LOS VIRUS DEL ZIKA Y DENGUE MEDIANTE EL DESARROLLO DE UN ALGORITMO BIOINFORMÁTICO PARALELO PARA EL ALINEAMIENTO Y ANÁLISIS EFICIENTES DE SECUENCIAS

DIEGO ARCOS SAPENA

TRABAJO DE FIN DE GRADO DE BIOTECNOLOGÍA

**Tutor académico:** Dr. Santiago Garcia Vallvé, grado de Biotecnología, Departamento de Bioquímica y Biotecnología (santi.garcia-vallve@urv.cat).

**En cooperación con:** Cheminformatics and Nutrition Research Group.

**Supervisor:** Dr. Santiago Garcia Vallvé, grado de Biotecnología, Departamento de Bioquímica y Biotecnología (santi.garcia-vallve@urv.cat).

Facultad de Enología, Campus Sescelades, URV, Tarragona

06/06/2025

Fecha de convocatoria: junio de 2025

Yo, Diego Arcos Sapena, con DNI 20941305N, soy conocedor de la guía de prevención de plagio en la URV *Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants* (aprobada en julio de 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) y afirmo que este TFG no constituye ninguna de las conductas consideradas como plagio por la URV.

Tarragona, 06 de junio de 2025

## Agradecimientos

Quiero comenzar expresando mi gratitud, en primer lugar, a mi mejor amigo: yo mismo. A mi resiliencia, que me ha permitido mantenerme firme en los momentos más desafiantes. A mi ambición, esa fuerza que me impulsa a soñar en grande y a fijarme metas que me desafían constantemente. Y, sobre todo, a mi inconformismo, esa chispa que muchas veces se ha visto como un defecto, pero que hoy reconozco como una de mis mayores virtudes. Este inconformismo domado me permite cuestionar los límites establecidos y perseguir siempre la mejora constante. Hoy agradezco esta fortaleza interior que me guía hacia la excelencia y me acerca, paso a paso, a la mejor versión de mí mismo.

También, quiero agradecer a mis padres, hermano y familiares cercanos, por brindarme siempre las facilidades, el apoyo y el cariño necesarios para avanzar.

A mi tutor de TFG, el Dr. Santiago Garcia Vallvé, por abrirme las puertas al grupo de investigación Cheminformatics and Nutrition Research Group y por su confianza y orientación impecable a lo largo de este camino.

Por último, a la buena música, que me ha acompañado fielmente en todas las horas dedicadas a este proyecto, convirtiéndose en mi aliada silenciosa.

# ÍNDICE

# 1. DATOS DEL GRUPO DE INVESTIGACIÓN

El presente Trabajo de Fin de Grado se ha desarrollado en el grupo de investigación Cheminformatics and Nutrition Research Group de la Universitat Rovira i Virgili, con consolidada experiencia en diseño computacional de fármacos (cribados virtuales, docking proteína-ligando, farmacóforos) aplicados a productos naturales para identificar ingredientes bioactivos.

Desde la pandemia de COVID-19, el grupo reorientó significativamente sus esfuerzos al estudio del SARS-CoV-2, enfocándose en el desarrollo de inhibidores antivirales para la proteasa principal (M-pro) y en el análisis de las mutaciones del virus. Mediante cribado virtual, se identificaron cuatro compuestos (celecoxib, carprofen, sarafloxacina y perampanel) que demostraron actividad inhibitoria in vitro contra la M-pro, validando la aproximación. Adicionalmente, se analizaron las limitaciones de los métodos de docking para predecir nuevos inhibidores de esta proteasa.

En cuanto a la evolución viral, se desarrolló un método de aprendizaje automático (machine learning) para predecir mutaciones recurrentes en SARS-CoV-2 y se analizaron los puntos calientes (hotspots) y fríos (coldspots) de mutación en la M-pro. Los avances del grupo en SARS-CoV-2 se han publicado en revistas como International Journal of Molecular Sciences, International Journal of Infectious Diseases y Medicinal Research Reviews. Este TFG se enmarca en esta línea de investigación sobre virus ARN y análisis de mutaciones, aprovechando la experiencia y recursos del grupo.

## 2. RESUMEN Y PALABRAS CLAVE

Los virus del Zika (ZIKV) y del Dengue (DENV), Flavivirus de gran impacto en la salud pública global, presentan una alta variabilidad genómica que complica el desarrollo de contramedidas. El análisis eficiente de sus mutaciones es crucial para identificar dianas biotecnológicas. Este Trabajo de Fin de Grado se centró en analizar esta variabilidad mediante el desarrollo de un algoritmo bioinformático paralelo.

Para ello, se implementó una herramienta en C++ que integra OpenMP para la paralelización de tareas y la librería Parasail para la aceleración de alineamientos mediante instrucciones SIMD. Dicho algoritmo, optimizado adicionalmente con una estrategia de procesamiento segmentado por genes para reducir la complejidad computacional, fue validado con 1.093 genomas de ZIKV y 25.917 de DENV (GISAID). Los resultados demostraron notables aceleraciones (*speedup*) de hasta 69x para ZIKV y entre 43-65x para DENV, subrayando su idoneidad para el análisis a gran escala.

La aplicación de este algoritmo al análisis genómico de ZIKV y DENV reveló que ambos virus comparten una fuerte selección purificadora sobre mutaciones con cambio de sentido, así como patrones de sustitución nucleotídica dominantes consistentes con la acción de enzimas editoras de ARN del hospedador (APOBEC/ADARs). Además, se identificó una variabilidad heterogénea por gen: por un lado, las regiones UTR y ciertas proteínas de superficie (E y M en ZIKV; NS2A y 2K en DENV) mostraron mayor plasticidad, lo cual es relevante para el diseño de vacunas. Por otro lado, y en contraste, genes estructurales internos y enzimáticos (ancC, NS2B, NS3, NS4A y NS5 en ZIKV; NS1 y NS3 en DENV) exhibieron una mayor conservación funcional, confirmándose como prometedores objetivos para fármacos antivirales de amplio espectro.

En definitiva, la herramienta bioinformática desarrollada se presenta como una solución eficaz para el estudio de la evolución viral. Los hallazgos sobre la variabilidad y conservación genómica en ZIKV y DENV ofrecen información fundamental que contribuye al diseño de futuras intervenciones biotecnológicas contra estos patógenos, cumpliendo así los objetivos planteados.

**Palabras clave:** Virus del Zika, Virus del Dengue, Flavivirus, Variabilidad Genómica, Análisis de Mutaciones, Bioinformática, Alineamiento de Secuencias, Paralelización, OpenMP, SIMD, Parasail, Tasas de Mutación, Dianas Biotecnológicas, Evolución Viral.

# 3. INTRODUCCIÓN

## 3.1. Incidencia Global y Contexto Epidemiológico de Dengue y Zika

Los virus del dengue (DENV) y del zika (ZIKV) son arbovirus ARN pertenecientes a la familia Flaviviridae. Estos virus son transmitidos principalmente por mosquitos del género Aedes, en particular las especies *A. aegypti* y *A. albopictus* y ambos representan amenazas sanitarias globales significativas (Bhandari et al., 2023; Martin Reyes-Baque et al., 2020). Aunque presentan similitudes diagnósticas, DENV puede causar cuadros graves como dengue hemorrágico, mientras ZIKV, a menudo leve, es alarmante por sus complicaciones neurológicas y defectos congénitos (ej. microcefalia), como evidenció la epidemia en Brasil (2015-2016) declarada emergencia sanitaria internacional por la OMS (Martin Reyes-Baque et al., 2020).

Tal y como se observa en la Figura 1 y Figura 2, el impacto global de ambos virus es considerable, especialmente en regiones tropicales y subtropicales vulnerables (América Latina, Sudeste Asiático, África) (Brady et al., 2024; Liang & Dai, 2024). DENV afecta a decenas de millones anualmente; ZIKV, con menor incidencia, preocupa por sus secuelas severas y rápida expansión, problemas agravados por la densidad poblacional urbana y condiciones ambientales que favorecen al vector (Brady et al., 2024).

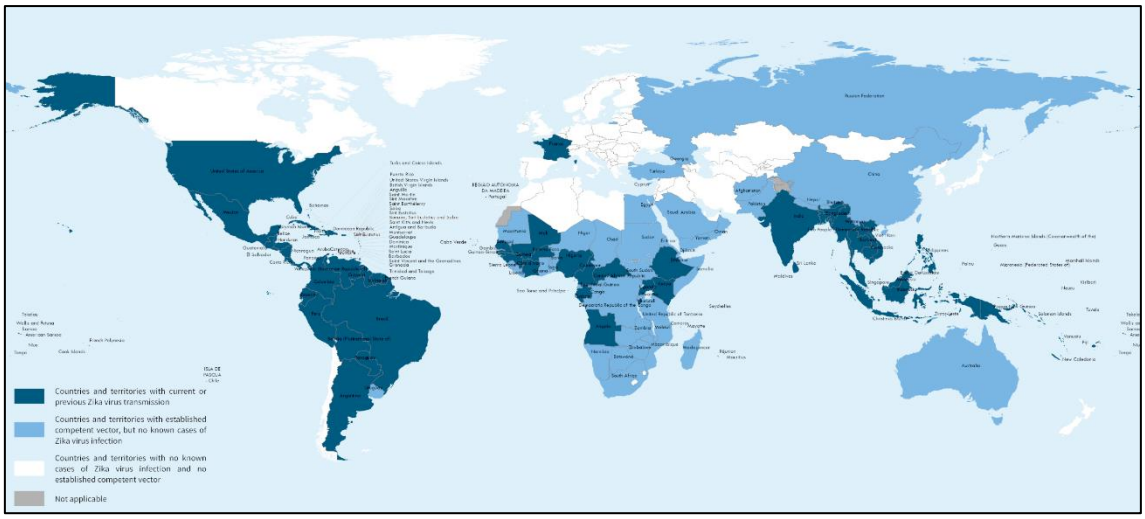


*Figura 1. Distribución global de la transmisión del virus del Zika y presencia de vectores competentes. Fuente: https://www.who.int/publications/m/item/zika-epidemiology-update-may-2024.*

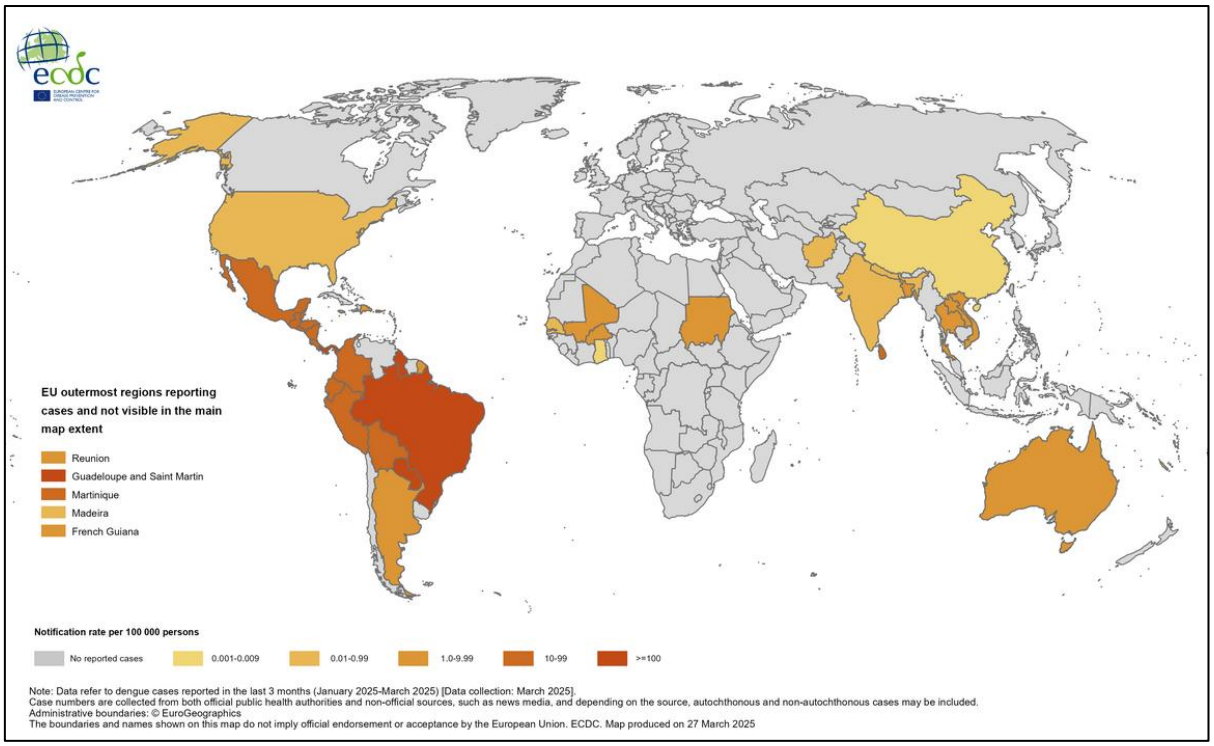


*Figura 2. Incidencia global de Dengue notificada (Primer Trimestre 2025). Fuente: https://www.ecdc.europa.eu/en/dengue/surveillance/dengue-virus-infections-travellers.*

El calentamiento global altera drásticamente los patrones climáticos (temperatura, precipitación, etc.), creando condiciones más favorables para la proliferación de Aedes y la transmisión viral (Feng et al., 2024). Este fenómeno acelera el ciclo del vector, reduce la incubación extrínseca del virus, puede incrementar los criaderos y expande las zonas de riesgo y las temporadas de actividad (Feng et al., 2024; Martin Reyes-Baque et al., 2020). Estas proyecciones configuran una "invasión climática" que exige incorporar el cambio climático en las políticas de salud pública. Factores como la urbanización acelerada, "islas de calor", movilidad humana y cambios en el uso del suelo también facilitan el establecimiento de nuevos focos epidemiológicos, retando a los sistemas de vigilancia (Brady et al., 2024; Martin Reyes-Baque et al., 2020).

Ante este panorama, urgen estrategias de intervención efectivas y multifactoriales, que incluyan control vectorial, mejores métodos diagnósticos (dada la similitud clínica y coinfección) y la consideración de determinantes sociales (Côrtes et al., 2023; Guanche Garcell et al., 2020; Liang & Dai, 2024; Morgan et al., 2021). La modelización geoespacial y análisis predictivos son herramientas avanzadas para identificar "áreas calientes" y proyectar escenarios (Brady et al., 2024; Liang & Dai, 2024), reconociendo

que la expansión viral es una convergencia de factores ambientales, vectoriales y socioeconómicos (Brady et al., 2024).

En resumen, DENV y ZIKV constituyen crecientes desafíos sanitarios globales. Su compleja epidemiología y expansión, influenciada por factores como el calentamiento global, subrayan la urgencia de un conocimiento molecular detallado de estos virus para fundamentar el desarrollo de nuevos fármacos y vacunas.

### 3.2. Características Virológicas y Organización Genómica de ZIKV y DENV

Los virus del Dengue y Zika, del género Flavivirus (familia Flaviviridae), comparten una organización genómica ARNmc+ de ~10.7-11 kb con otros flavivirus como el de la fiebre amarilla y Nilo Occidental (Contreras et al., 2021a; Démocratique Et Populaire,s. f.). Su genoma, representado esquemáticamente en la Figura 3, presenta un único marco de lectura abierto (ORF) flanqueado por regiones no traducidas (UTR) 5' y 3', cuyas estructuras secundarias y motivos conservados son cruciales para la replicación del ARN viral, su traducción eficiente y la interacción con factores del hospedador (BOULDJEDJE et al., 2019; Ye et al., 2016).

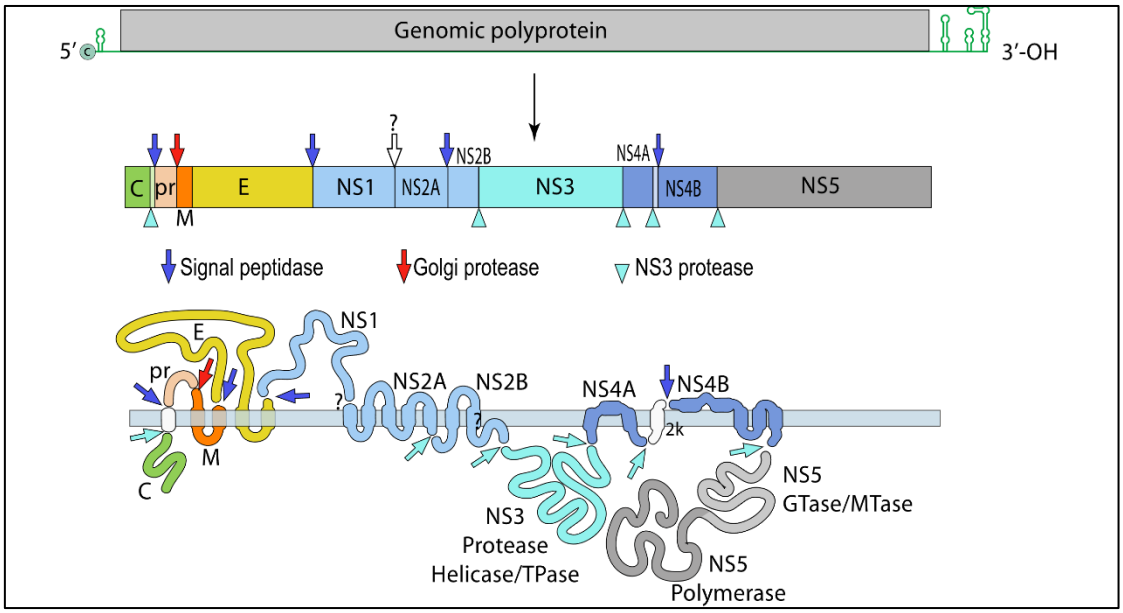


*Figura 3. Organización genómica y procesamiento de la poliproteína en ZIKV y DENV. Fuente: https://viralzone.expasy.org/24.*

El ORF se traduce en una poliproteína precursora de ~3400-3500 aminoácidos, procesada co- y postraduccionalmente por proteasas virales (principalmente NS3/NS2B) y del hospedador. Este clivaje secuencial genera tres proteínas estructurales (C, prM/M, E) y al menos siete no estructurales (NS1, NS2A, NS2B, NS3, NS4A, NS4B y NS5) (BOULDJEDJE et al., 2019; Dwivedi et al., 2017), todas esenciales para el ciclo viral.

Las proteínas estructurales (representadas en la Figura 4), que forman la partícula viral infecciosa (Tabla 1), se describen a continuación.

*Tabla 1. Proteínas estructurales del virus del Zika y Dengue.*

| Gen | Proteína | Función de la proteína |
|---|---|---|
| ancC | Proteína de la cápside anclada (ancC) | Forma la nucleocápside al unirse al ARN genómico viral |
| prM | Proteína precursora de membrana (prM) | Actúa como chaperona para el plegamiento de E. Durante la maduración viral, se produce una escisión, liberando la proteína M |
| E | Proteína de envoltura (E) | Media la entrada a células hospedadoras |

**Proteína de la cápside anclada (ancC):** La proteína de la cápside anclada es un componente estructural esencial del virión, caracterizada por su tamaño relativamente pequeño y su naturaleza básica, debido a la abundancia de aminoácidos cargados positivamente. Esta propiedad es crucial para su función principal: interactuar electrostáticamente con el ARN genómico viral, que está cargado negativamente, para empaquetarlo y formar la nucleocápside. Esta estructura proteica-nucleica protege el genoma viral de la degradación por nucleasas del hospedador y facilita su correcta organización dentro de la partícula viral. Durante el proceso de ensamblaje del virión, múltiples copias de la proteína ancC se autoensamblan alrededor del ARN genómico, formando el núcleo viral, que posteriormente será envuelto por la bicapa lipídica derivada del hospedador donde se anclan las proteínas de envoltura E y M(BOULDJEDJE et al., 2019).

**Proteína prM (prM):** Esta proteína es inicialmente sintetizada como el precursor prM, siendo crucial en las partículas virales inmaduras. En esta etapa, prM actúa como chaperona para el correcto plegamiento de la proteína de Envoltura (E) y, de forma crítica, previene su activación fusogénica prematura durante el ensamblaje y transporte del virión. Posteriormente, durante la maduración viral, la prM es escindida por la proteasa celular furina, generando la proteína M madura y el péptido pr. La proteína M se integra como un componente estructural de la envoltura del virión infeccioso, mientras que la disociación del péptido pr es necesaria para que la proteína E alcance su conformación plenamente funcional para la infección(BOULDJEDJE et al., 2019).

**Envoltura (E):** La glicoproteína de envoltura (E) es el principal antígeno expuesto en la superficie del virión y es esencial para iniciar la infección. En los viriones maduros, se organiza típicamente en homodímeros y su función primordial es mediar la entrada a la célula hospedadora. Este proceso incluye la unión a receptores celulares específicos y, tras la endocitosis, la fusión de la membrana viral con la membrana endosomal. Dicha fusión es desencadenada por el pH ácido del endosoma, que induce cambios conformacionales en E y expone un lazo de fusión. Dada su crucial función y localización, la proteína E es el principal inductor de anticuerpos neutralizantes, convirtiéndola en un objetivo prioritario para el desarrollo de vacunas. La variabilidad genética de esta proteína es un factor determinante en la diversidad antigénica y serotípica, especialmente relevante en virus como el DENV (BOULDJEDJE et al., 2019; Ekins et al., 2016; Higuera & Ramírez, 2019).

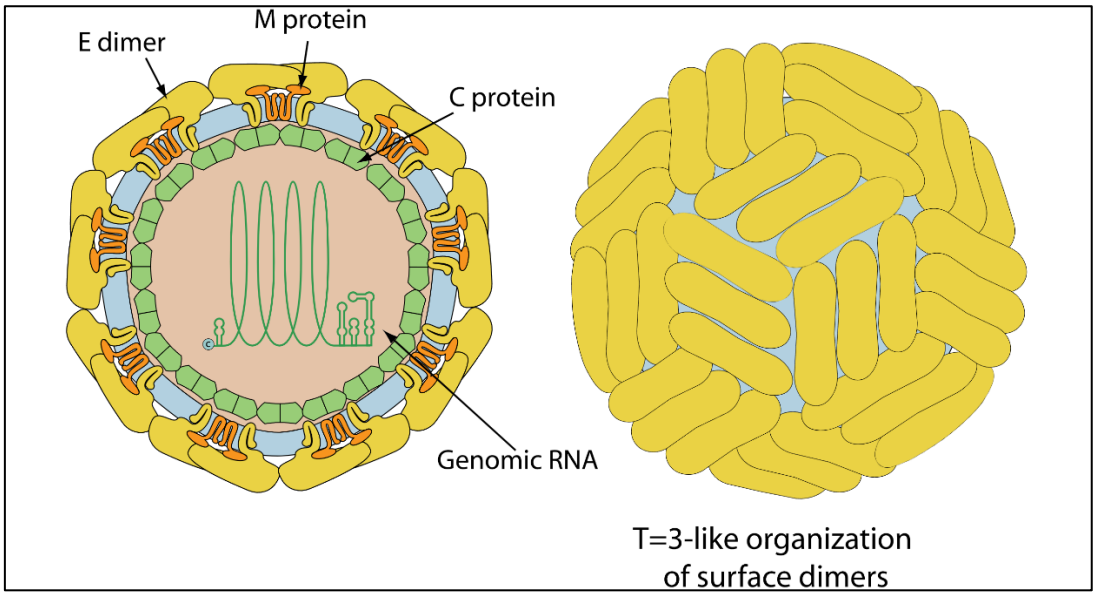


*Figura 4. Estructura del virión de un Flavivirus. Fuente: https://viralzone.expasy.org/24.*

Las proteínas no estructurales (Tabla 2) son cruciales para la replicación del genoma viral, el ensamblaje y la modulación de la célula hospedadora. Se describen a continuación.

*Tabla 2. Proteínas no estructurales del virus del Zika y Dengue.*

| Gen | Proteína | Función de la proteína |
|---|---|---|
| NS1 | Proteína NS1 | Implicada en la replicación temprana del ARN viral |
| NS2A | Proteína NS2A | Implicada en el ensamblaje viral |
| NS2B | Proteína NS2B | Cofactor esencial para la actividad proteasa de NS3 |
| NS3 | Proteína NS3 | Proteína multifuncional |
| NS4A | Proteína NS4A | Inducen el reordenamiento de membranas del retículo endoplásmico para formar los complejos de replicación |
| NS4B | Proteína NS4B | Ídem NS4A |
| NS5 | Proteína NS5 | Es esencial para la replicación y modificación del ARN viral, gracias a sus actividades de ARN polimerasa dependiente de ARN (RdRp) y metiltransferasa (MTasa). |

**NS1:** Esta glicoproteína desempeña un papel crucial en las etapas tempranas de la replicación del ARN viral, participando en la formación de los complejos de replicación. Además de su función intracelular, la NS1 es secretada por las células infectadas en forma de un hexámero soluble, el cual puede circular en el torrente sanguíneo del hospedador. Esta forma secretada es un importante antígeno y se ha implicado en la modulación de la respuesta inmune, por ejemplo, interfiriendo con la cascada del complemento o la señalización del interferón, y puede contribuir a la patogénesis viral, como la disfunción endotelial observada en infecciones severas.(BOULDJEDJE et al., 2019; Higuera & Ramírez, 2019).

**NS2A:** La NS2A es una proteína de membrana hidrofóbica que participa en múltiples aspectos del ciclo viral. Es fundamental para el correcto ensamblaje de las nuevas partículas virales, probablemente facilitando la interacción entre la nucleocápside y las proteínas de envoltura en las membranas del retículo endoplásmico. Además de su rol en

el ensamblaje, se ha demostrado que NS2A contribuye a la evasión de la respuesta inmune innata del hospedador, interfiriendo con la vía de señalización del interferón tipo I, lo que ayuda al virus a establecer una infección productiva.

**NS2B:** Esta pequeña proteína transmembrana es un componente indispensable del complejo proteolítico viral. Actúa como un cofactor esencial para la actividad de la serina proteasa NS3, anclando el dominio proteasa de NS3 (NS3pro) a las membranas celulares y orientándolo correctamente para la escisión de la poliproteína viral.(Dwivedi et al., 2017).

**NS3:** La NS3 es una proteína altamente multifuncional y una de las enzimas clave del virus. Su región N-terminal (NS3pro), en complejo con su cofactor NS2B, forma una serina proteasa responsable del procesamiento de la poliproteína viral en múltiples sitios específicos, liberando las proteínas individuales necesarias para el ciclo viral. La región C-terminal de NS3 (NS3hel) alberga varias actividades enzimáticas distintas: funciona como una ARN helicasa, que desenrolla estructuras secundarias del ARN viral durante la replicación; una NTPasa, que proporciona la energía para esta actividad; y una RTPasa (ARN trifosfatasa), implicada en la formación del capuchón (cap) del ARN viral (Dwivedi et al., 2017; Wahaab et al., 2021; Ye et al., 2016).

**NS4A y NS4B:** Estas dos son proteínas hidrofóbicas pequeñas, ancladas a membranas, que juegan roles importantes en la manipulación de las membranas celulares del hospedador y en la evasión inmune. La NS4A, junto con otras proteínas NS, induce el reordenamiento de las membranas del retículo endoplásmico para formar los orgánulos de replicación viral, creando un entorno protegido para la síntesis del ARN. La NS4B también participa en la formación de estos complejos de replicación y, adicionalmente, ambas proteínas han sido implicadas en el antagonismo de la respuesta inmune innata, particularmente interfiriendo con la señalización del interferón para favorecer la persistencia de la infección.

**NS5:** La proteína más grande y conservada. Su N-terminal metiltransferasa (MTasa) realiza el capping del ARN viral (clave para estabilidad, traducción y evasión inmune); su C-terminal es la ARN polimerasa dependiente de ARN (RdRp) que sintetiza el genoma (Van Den Elsen et al., 2021; Ye et al., 2016). La acción coordinada de las proteínas NS en complejos anclados a membranas es indispensable para la propagación viral (Dwivedi et al., 2017; Van Den Elsen et al., 2021).

En cuanto a diversidad viral, DENV presenta cuatro serotipos distintos (DENV1 a 4, ocasionalmente un quinto), con variabilidad genética en la proteína E como base de su distinción antigénica; esto complica el desarrollo de vacunas y se relaciona con inmunopatología como la ADE en infecciones secundarias heterólogas (Dwivedi et al., 2017; Ekins et al., 2016). ZIKV, con linajes genéticos africano y asiático (este último responsable de la reciente expansión epidémica y complicaciones neurológicas) (Jablunovsky et al., 2024), presenta cambios genéticos que podrían haber afectado su replicación, tropismo o evasión inmune (Zhu et al., 2016).

Una comprensión detallada de esta virología y organización genómica es fundamental para el análisis de mutaciones. Variaciones genéticas en proteínas o regiones reguladoras pueden impactar la virulencia, replicación, transmisibilidad, respuesta a antivirales y evasión inmune (Zhu et al., 2016). El estudio exhaustivo de estas mutaciones (frecuencia, localización, consecuencias fenotípicas) es crucial para la vigilancia epidemiológica y el desarrollo racional de vacunas y terapias antivirales efectivas (Bhutkar et al., 2022; Gupta et al., 2016).

### 3.3. Relevancia del Análisis Genómico para la Salud Pública y Dianas Terapéuticas de Zika y Dengue

El dengue y Zika representan amenazas sanitarias globales. Su elevada tasa de mutación exige un análisis genómico indispensable para el seguimiento de variantes emergentes, comprender su dinámica evolutiva (patogenicidad, evasión inmune) y ajustar intervenciones terapéuticas y profilácticas, cuya eficacia puede verse impactada por cambios en secuencias específicas (Dutta & Langenburg, 2023; Vig, 2024). Este análisis es también fundamental para el desarrollo de vacunas y fármacos, al ayudar a identificar regiones conservadas o variables para diseñar antígenos estables, dianas terapéuticas menos propensas a la resistencia, y guiar la selección de epítopos inmunogénicos y el diseño de antivirales (Lorenza Trabalzini & Pini Filippo Dragoni SUPERVISOR Maurizio Zazzi, 2019; Zhao et al., 2021).

Entre las proteínas virales clave, la proteína cápside (C), esencial para la encapsulación y ensamblaje del virión, es una diana vacunal prometedora; mutaciones en C pueden afectar el empaquetamiento y la respuesta inmune del huésped (Antonios & Daelemans Jury members Piet Maes Paul Proost Kevin Arien Jan Munch, 2021; Dutta & Langenburg, 2023). Las enzimas de replicación, como la ARN polimerasa NS5 (con actividad RdRp y

metiltransferasa) y el complejo NS3/NS2B (proteasa/helicasa que procesa la poliproteína viral), son objetivos primordiales para antivirales, ya que su inactivación detiene la replicación o la liberación de proteínas funcionales. Es crucial destacar que mutaciones en estas enzimas pueden inducir resistencia, lo que subraya la importancia del seguimiento genómico continuo para el rediseño de inhibidores (Lorenza Trabalzini & Pini Filippo Dragoni SUPERVISOR Maurizio Zazzi, 2019; Mittal et al., 2022).

Las proteínas de envoltura E y M son centrales en la entrada viral. La proteína E media la unión a receptores celulares y la fusión de membranas, siendo el principal blanco de anticuerpos neutralizantes, mientras que la proteína M participa en la maduración y estabilización del virión (Roy et al., 2024). Las comparaciones estructurales, por ejemplo, con la proteína spike de SARS-CoV-2, pueden inspirar estrategias vacunales basadas en la identificación de epítopos conservados con potencial para intervenciones inmunológicas transversales (Tampere, 2021; Wollner & Richner, 2021).

Actualmente, no existen vacunas universalmente recomendadas y plenamente eficaces para dengue y Zika; Dengvaxia, por ejemplo, tiene un uso restringido en individuos seronegativos debido al riesgo de enfermedad potenciada, lo que refleja la complejidad de la respuesta inmune (Dutta & Langenburg, 2023). En cuanto a los fármacos, diversos candidatos inhibidores de NS5 o NS2B-NS3 han mostrado eficacia clínica variable y han enfrentado problemas de aparición de resistencia viral (Diani et al., 2023; Lorenza Trabalzini & Pini Filippo Dragoni SUPERVISOR Maurizio Zazzi, 2019). Por ello, el desarrollo de nuevos antivirales se enfoca en compuestos que interactúen con sitios moleculares críticos, utilizando modelado molecular para predecir interacciones y minimizar la aparición de resistencia (Kumar et al., 2022).

La considerable diversidad genética de los flavivirus exige un enfoque integral que combine la vigilancia genómica constante de variantes emergentes con estudios funcionales que evalúen el impacto de las mutaciones identificadas. Este seguimiento continuo es vital para anticipar posibles cambios en la infectividad del virus y en su sensibilidad frente a las terapias disponibles, permitiendo así la adaptación y optimización de las intervenciones de salud pública (Diani et al., 2023; Zhao et al., 2021).

El análisis genómico y la comprensión de las mutaciones en proteínas virales clave como C, NS5, NS3/NS2B, E y M son, por tanto, fundamentales para enfrentar el desafío que suponen el DENV y ZIKV.

### 3.4. Mutabilidad Viral y Patrones de Mutación de Zika y Dengue

Las estrategias evolutivas de los virus ARN como Dengue y Zika difieren notablemente de otros, como el SARS-CoV-2. DENV y ZIKV se caracterizan por una alta mutabilidad, principalmente porque su ARN polimerasa dependiente de ARN (RdRp) carece de actividad exonucleasa correctora de errores (*proofreading*) (Contreras et al., 2021b). Esto resulta en tasas de error elevadas, estimadas en ~$1x10^{-4}$ mutaciones por genoma y replicación, lo que conduce a la rápida formación de poblaciones virales heterogéneas (cuasiespecies) y una continua adaptación evolutiva (Contreras et al., 2021b; Pérez, 2019). Esta baja fidelidad replicativa es un factor que limita el tamaño de su genoma a unas 10-11 kilobases (kb), para así evitar la acumulación excesiva de mutaciones deletéreas (Pérez, 2019).

En contraste, SARS-CoV-2, con un genoma de ARN mucho más extenso (~30 kb), emplea un mecanismo de corrección de errores mediado por la actividad exonucleasa de su proteína no estructural NSP14 (Cordo, 2020; Ramírez Corona, 2023). Esta capacidad de *proofreading* reduce significativamente su tasa de mutación en comparación con DENV y ZIKV, permitiendo mantener la integridad de un genoma mayor (Villanueva Romero, 2023). Se observa así una relación inversa entre la presencia de mecanismos correctores y la viabilidad de genomas de mayor tamaño (Ramírez Corona, 2023).

Estas diferencias en las tasas de mutación intrínsecas son el punto de partida para la acción de mecanismos evolutivos más amplios. Así, la mutación, incluyendo inserciones y deleciones, junto con la recombinación, son dos mecanismos importantes que generan la variabilidad genómica en las variantes virales. La mayoría de las mutaciones virales tiende a ser neutra o ligeramente deletérea, ya que cambios que no alteran drásticamente las funciones proteicas o la estructura del virus tienen poca repercusión sobre la aptitud viral (Dennehy, 2017). Mutaciones altamente perjudiciales –por ejemplo, aquellas que impiden la capacidad del virus para invadir a su huésped– raramente se fijan en la población, dada la fuerte acción de la selección purificadora que elimina variantes con efectos deletéreos (Lauring, 2020).

Sin embargo, en presencia de presiones selectivas, tales como la respuesta inmune del huésped, la aplicación de vacunas o el uso de fármacos antivirales, ciertos cambios mutacionales que confieran ventajas adaptativas pueden ser seleccionados positivamente. Estas mutaciones favorecen características como el incremento de la virulencia, la

infectividad, la transmisibilidad, una mayor capacidad de replicación o la evasión del sistema inmune. Al incrementar la aptitud del virus en contextos específicos, estos cambios pueden aumentar rápidamente en frecuencia y propagarse a lo largo de la población viral (Grubaugh, 2016).

Es importante destacar que la alta frecuencia de ciertas mutaciones en la población viral no implica necesariamente que la mutación tenga un efecto beneficioso directo. En algunos casos, el efecto fundador puede explicar este fenómeno: una mutación que aparezca tempranamente en el curso de una epidemia puede propagarse a todos los descendientes si se encuentra en el genotipo original que dio origen a la epidemia. De manera similar, una mutación puede llegar a alta frecuencia no por su propio beneficio, sino por encontrarse en el mismo genoma que otra mutación ventajosa, un fenómeno conocido como efecto de arrastre o *hitchhiking*. Estos procesos son particularmente evidentes en virus como DENV y ZIKV, que experimentan altas tasas de generación de variantes debido a su elevado número de replicación y la dinámica rápida de su ciclo de vida.

Esta diferencia en la fidelidad replicativa tiene profundas implicaciones. La alta mutabilidad de DENV y ZIKV facilita la rápida generación de variantes con ventajas adaptativas, permitiéndoles evadir la respuesta inmune y adaptarse a diversos entornos y vectores (Contreras García, 2020; Contreras et al., 2021b). Ejemplos específicos de mutaciones con impacto fenotípico incluyen la sustitución S139N en la proteína prM de ZIKV, relacionada con mayor infectividad y asociación con microcefalia (Giné, 2020), o mutaciones en su proteína NS1 que afectan la patogenicidad (González Almanza, 2018). En DENV, la mutación N390D en el dominio III de la proteína E se ha asociado con un aumento de la virulencia en ciertos contextos (Contreras et al., 2021b). La identificación de estas mutaciones específicas y sus consecuencias es fundamental para profundizar en la comprensión de la patogénesis viral (González Almanza, 2018).

Adicionalmente a los errores intrínsecos de la polimerasa viral, se ha planteado la hipótesis de que enzimas del huésped, como las deaminasas de citidina de la familia APOBEC y las deaminasas de adenosina que actúan sobre ARN (ADARs), contribuyen al perfil mutacional. Estas enzimas de la inmunidad innata pueden inducir patrones específicos de mutaciones (predominantemente transiciones C→U por APOBEC, y A→I –leída como G– por ADARs). Aunque se han observado patrones consistentes con la acción de APOBEC/ADARs en SARS-CoV-2, también se sugiere su influencia en la

variabilidad de flavivirus (Ortega Pérez et al., 2022). Esta edición del ARN por el huésped puede generar diversidad funcional en las proteínas virales, afectando el reconocimiento inmune y la interacción con receptores, lo que constituye un arma de doble filo al poder favorecer la evolución viral o generar formas no viables (Contreras et al., 2021b; Giné, 2020; Ortega Pérez et al., 2022).

La comprensión detallada de estos mecanismos de mutabilidad (tanto los errores inherentes a la RdRp de los flavivirus como la posible edición por enzimas del huésped) es, por tanto, esencial para contextualizar la evolución y adaptación de DENV y ZIKV.

### 3.5. Necesidad de Enfoques Computacionales de Alto Rendimiento en el Análisis Genómico

La genómica moderna, con la secuenciación masiva (NGS y tercera generación), genera volúmenes de datos (GBs a TBs) que representan un desafío computacional sin precedentes. Para enfermedades emergentes como Dengue y Zika, la extracción rápida y precisa de esta información es crucial para la investigación y las decisiones en salud pública. El procesamiento secuencial tradicional es inviable, haciendo indispensable la computación de alto rendimiento (HPC) y la paralelización (Terán Amores, 2023; Varela Tabares, 2019).

La paralelización, dividiendo tareas para ejecución simultánea en diversos entornos (múltiples núcleos, GPUs, clústeres), es clave para el análisis eficiente de estos vastos *datasets* genómicos (Naiouf, 2022). Aplicada en todo el pipeline bioinformático (preprocesamiento, QC, alineamiento, identificación de variantes) (Koile, 2022), su principal ventaja es la drástica reducción de tiempos de ejecución (ver Figura 5), permitiendo respuestas urgentes en contextos clínicos y epidemiológicos, como brotes virales.

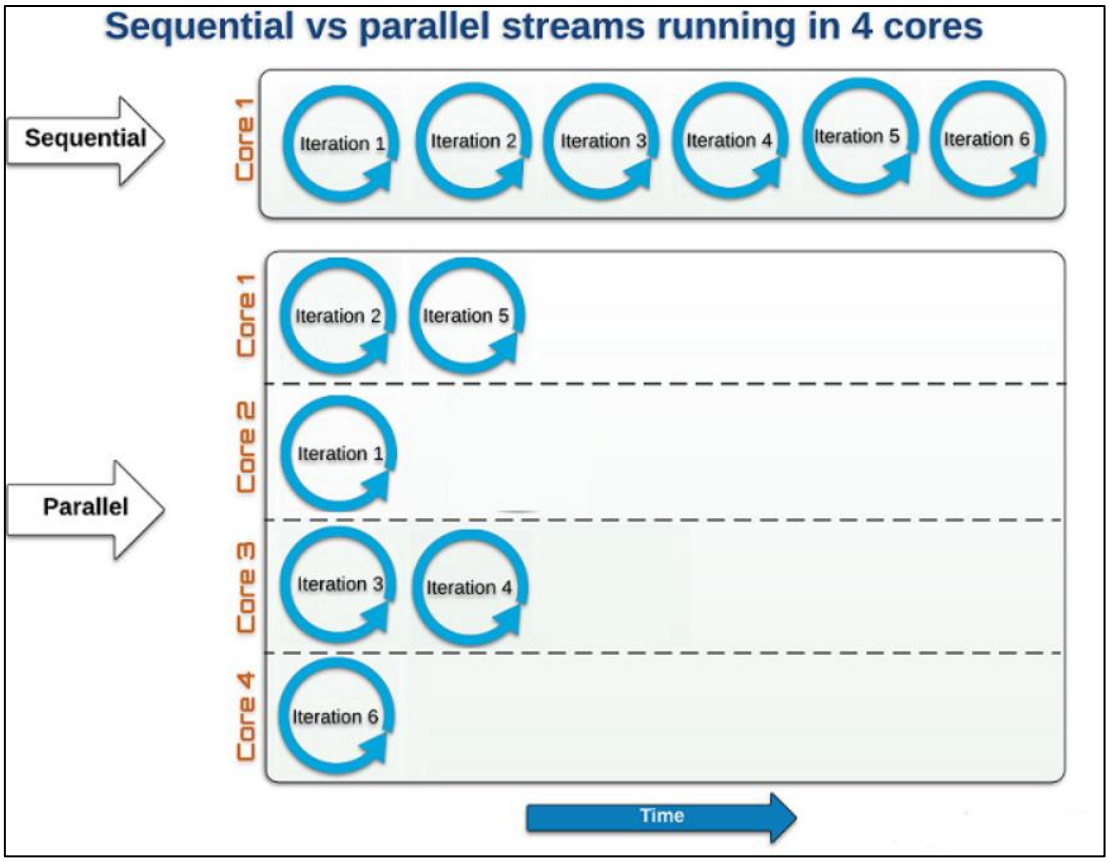


*Figura 5. Comparación de tiempo entre ejecutar 6 tareas en secuencial y ejecutar estas 6 tareas en paralelo con 4 núcleos. Fuente: https://www.logicbig.com/tutorials/core-java-tutorial/java-util-stream/sequential-vs-parallel.html.*

Además de la velocidad, el procesamiento paralelo y HPC mejoran la calidad y precisión, permitiendo controles de calidad exhaustivos, ajustes de parámetros e identificación precisa de variantes genéticas relevantes –crítico en virus como Dengue y Zika por el impacto de mutaciones sutiles–. También facilitan la reproducibilidad y validación de los análisis, cruciales en investigación biomédica (Suárez González & Romero Béjar, 2024).

HPC y paralelización son, además, habilitadores clave para la integración de análisis complejos, como los estudios multiómicos (genómica, transcriptómica, proteómica). Estas capacidades son esenciales para procesar datos heterogéneos y masivos, y lograr una comprensión holística de la biología viral y su interacción con el hospedador (Suárez González & Romero Béjar, 2024).

Para virus como Dengue y Zika, con alta mutabilidad y diversidad, el análisis genómico rápido y a gran escala es crucial, permitiendo vigilancia casi en tiempo real, identificación de variantes emergentes impactantes y adaptación de estrategias de salud pública.

Es precisamente en respuesta a esta imperiosa necesidad de herramientas eficientes que el presente Trabajo de Fin de Grado adquiere su relevancia, al haberse centrado en el desarrollo y aplicación de una solución computacional de alto rendimiento destinada al análisis genómico masivo de estos patógenos.

# 4. HIPÓTESIS Y OBJETIVOS

El presente Trabajo de Fin de Grado tiene como hipótesis que el análisis masivo de mutaciones y la variabilidad genómica en los virus del Zika y del Dengue permite la identificación de patrones clave para el diseño de nuevas estrategias biotecnológicas, como el desarrollo de vacunas y compuestos antivirales más efectivos. Con ello, el objetivo principal de este estudio es **analizar las mutaciones y la variabilidad genómica en ambos virus para comprender su dinámica evolutiva e identificar posibles dianas para el desarrollo de fármacos y vacunas**.

Para lograr este fin, se establecen los siguientes objetivos secundarios:

- Diseñar e implementar un algoritmo computacional paralelo capaz de hacer alineamiento por pares y analizar grandes volúmenes de secuencias genómicas, haciendo posible la detección y cuantificación de mutaciones en periodos de tiempo prácticos.
- Validar la eficiencia y el rendimiento del algoritmo desarrollado en el procesamiento de grandes volúmenes de datos de secuencias virales.
- Aplicar el algoritmo desarrollado a *datasets* de secuencias genómicas del virus del Dengue y del virus del Zika.
- Caracterizar y comparar los perfiles de mutación, las tasas de variación y los patrones de variabilidad genómica en ambos virus.
- Visualizar y discutir los hallazgos obtenidos en un contexto biotecnológico.

# 5. MATERIALES Y METODOLOGÍA

## 5.1. Obtención y Selección de Datos Genómicos

Las secuencias genómicas y los segmentos codificantes completos de DENV y ZIKV utilizados en este estudio se obtuvieron de la base de datos pública: GISAID (Global Initiative on Sharing All Influenza Data). Se han utilizado un total de 1093 genomas de ZIKV y 25.917 de DENV.

Para asegurar la relevancia y calidad de los datos, se establecieron fechas de corte para la descarga de las secuencias: para ZIKV, se incluyeron todas las disponibles hasta el 26 de mayo de 2025, y para DENV, hasta el 30 de abril de 2025. Solo se utilizaron secuencias de virus que han infectado a hospedadores humanos. Adicionalmente, se aplicaron parámetros de calidad para filtrar las secuencias seleccionadas. Entre estos, se utilizó *high coverage*, el cual filtra aquellas entradas genómicas con un máximo del 5% de nucleótidos ambiguos (Ns). Asimismo, el análisis se centró fundamentalmente en secuencias que representan "Secuencias Codificantes Completas" (Complete Coding Sequences, Cpl. CDS), es decir, aquellas que abarcan la totalidad de las regiones del genoma que se traducen directamente en proteínas virales, lo cual es esencial para enfocar el estudio de mutaciones en zonas de impacto funcional directo.

Para el alineamiento y la posterior identificación de variantes, se emplearon genomas de referencia específicos y bien caracterizados. En el caso del virus del Dengue, se utilizó como referencia una secuencia del serotipo 2 (DENV2), correspondiente a la cepa 16681 (Número de Acceso en NCBI: NC_001474.2). Esta cepa fue aislada en Tailandia en 1964 y su genoma completo fue descrito por Kinney et al., 1997. Para el virus Zika, se tomó como referencia la cepa MR 766 (Número de Acceso en NCBI: NC_012532.1, derivada de AY632535), cuya secuenciación completa y caracterización genómica fue publicada por Kuno & Chang, 2007. El uso de estos genomas de referencia estandarizados es crucial para garantizar la consistencia y comparabilidad de los resultados del análisis de mutaciones.

### 5.2. Desarrollo del Algoritmo Bioinformático para Análisis de Mutaciones

Para abordar el análisis de mutaciones en los genomas de los virus Dengue y Zika, se desarrolló un algoritmo bioinformático robusto. La elección del lenguaje de programación C++ fue fundamental debido a su alto rendimiento y eficiencia en la gestión de memoria y procesamiento, características cruciales para manejar los grandes volúmenes de datos genómicos y las operaciones computacionalmente intensivas. Con el fin de maximizar la velocidad de los alineamientos de secuencias, una de las tareas más costosas, se integró la librería Parasail (Daily, 2016). Esta biblioteca, implementada en C, está específicamente diseñada para acelerar alineamientos pareados (como Smith-Waterman o Needleman-Wunsch) mediante el uso avanzado de instrucciones SIMD (Single Instruction, Multiple Data) disponibles en los procesadores modernos, lo que permite realizar múltiples operaciones de comparación de forma paralela a un nivel de instrucción muy bajo. Adicionalmente, para la gestión de la concurrencia y la paralelización de tareas a un nivel superior, se empleó la API OpenMP (Open Multi-Processing). OpenMP facilita la creación de programas que pueden ejecutarse en múltiples hilos (*threads*) de manera eficiente en arquitecturas de procesadores multinúcleo, distribuyendo la carga de trabajo y reduciendo el tiempo total de ejecución.

El algoritmo desarrollado sigue un flujo de trabajo estructurado para procesar cada secuencia viral de consulta contra el genoma de referencia correspondiente (DENV2 o ZIKV) y realizar un análisis detallado de las variaciones genéticas. Los pasos más importantes del programa se pueden esquematizar de la siguiente manera:

1. **Carga y Preparación de Datos:**
    - Lectura del archivo FASTA del genoma de referencia.
    - Lectura del archivo multi-FASTA con las secuencias virales de consulta.
    - Procesamiento de un archivo de configuración TSV que define la anotación de los genes virales (nombre, coordenadas, tipo: CDS/UTR).
2. **Sondeo Inicial para Desfase (Offset):**
    - Para cada secuencia de consulta, asignada a un hilo de ejecución (gestionado por OpenMP), se realiza un alineamiento local (Smith-Waterman, utilizando Parasail) de una porción inicial de la consulta contra una porción inicial de la referencia para estimar un desfase (offset) global. Este paso es crucial para mejorar la precisión de los alineamientos genómicos subsecuentes.

3. **Procesamiento Gen por Gen (Paralelizado por Consulta):**
    - Cada secuencia de consulta es asignada a un hilo de ejecución independiente (gestionado por OpenMP).
    - Dentro de cada hilo, se itera sobre cada gen anotado:
        - Se extraen los segmentos correspondientes del gen de referencia y de la secuencia de consulta (considerando el offset).
        - Se realiza un alineamiento semi-global (utilizando Parasail) entre el segmento del gen de referencia y el de la consulta. Este tipo de alineamiento es una variante del alineamiento global que no penaliza los huecos (gaps) al principio o al final de una o ambas secuencias, lo que resulta útil cuando se comparan secuencias de longitudes diferentes o cuando se sospecha que una secuencia puede ser un fragmento de otra más larga.
4. **Análisis de Mutaciones y Anotación:**
    - Se calculan métricas del alineamiento (cobertura, identidad).
    - Se identifican y clasifican las mutaciones nucleotídicas (sustituciones, inserciones, deleciones).
    - Para regiones CDS, se traducen las secuencias y se determinan mutaciones aminoacídicas (mutaciones con cambio de sentido, sin sentido, sinónimas) y posibles *frameshifts*.
    - Se cuantifican mutaciones en UTRs y se registran bases anómalas e inserciones intergénicas.
5. **Generación de Reporte:**
    - Los resultados detallados para cada gen de cada consulta se escriben en un archivo de salida en formato TSV.

La eficiencia de este programa y su capacidad para analizar grandes conjuntos de datos genómicos en tiempos reducidos se basa en un enfoque de optimización triple. Primero, la división del trabajo en hilos mediante OpenMP, donde múltiples secuencias problema se procesan en paralelo y de manera independiente en cada núcleo del procesador, tanto en el primer alineamiento para el cálculo del desfase, como en el segundo, donde se procesan las secuencias gen por gen. Tal y como se observa en la Figura 6, esto explota el paralelismo a nivel de tarea.

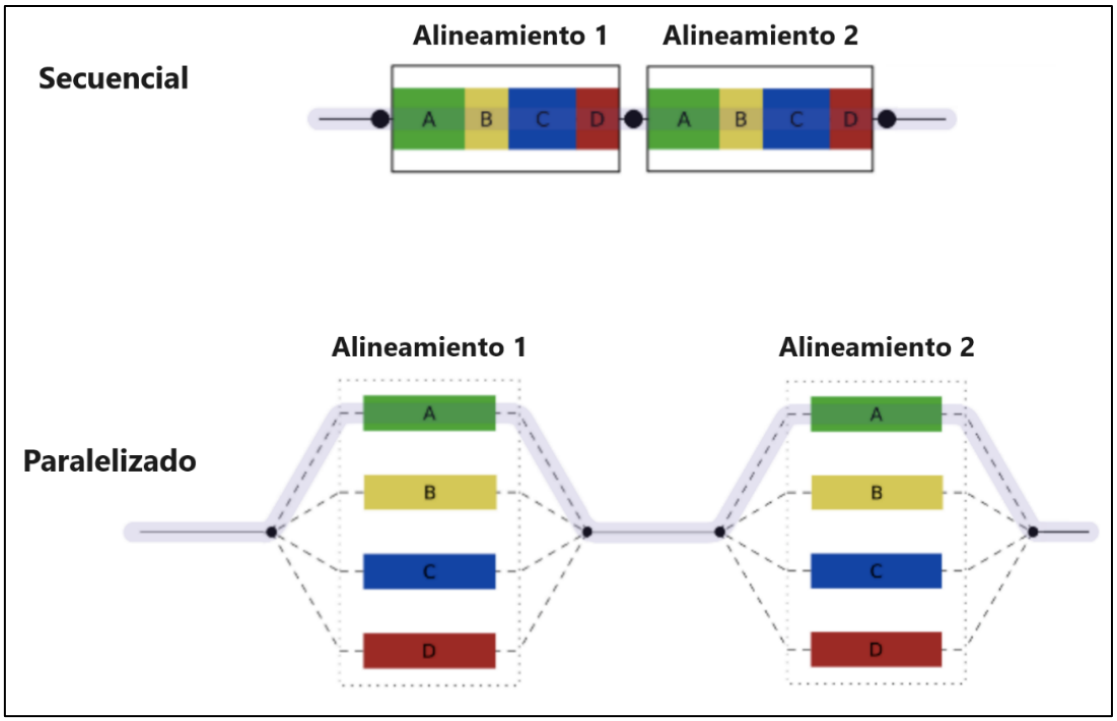


*Figura 6. Comparación entre realizar los 2 alineamientos en secuencial y en paralelo con la librería OpenMP. Paralelización a nivel de tarea.*

Segundo, el uso de la librería Parasail permite aprovechar las instrucciones SIMD para acelerar los cálculos intensivos de los alineamientos de secuencias, logrando un paralelismo a nivel de instrucción, representado en la Figura 7.

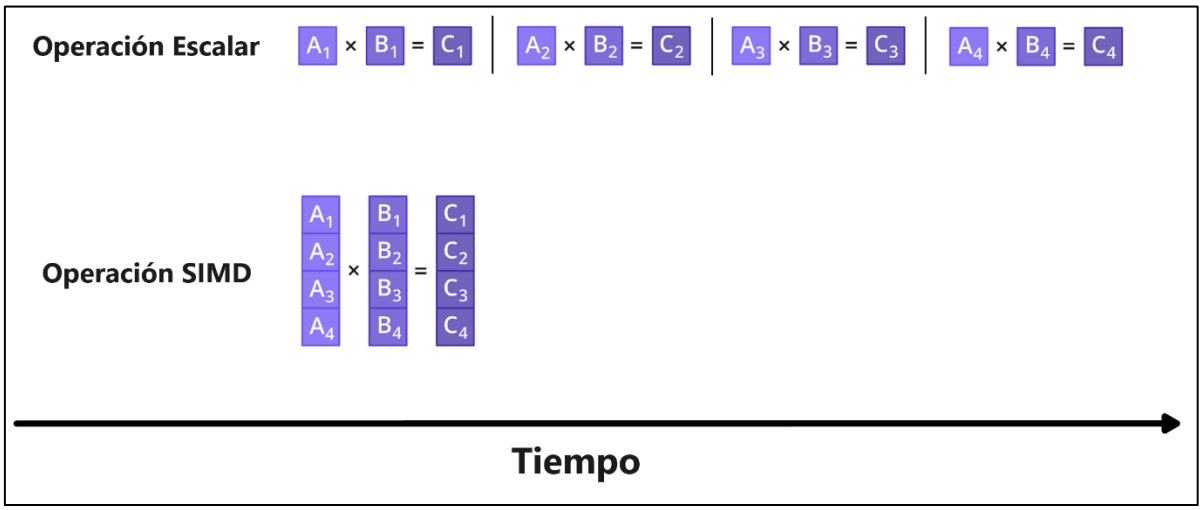


*Figura 7. Comparación entre ejecutar 4 instrucciones de CPU para realizar 4 cálculos (operación escalar) y ejecutar una sola instrucción de CPU para realizar los mismos 4 cálculos (operación SIMD). Paralelización a nivel de instrucción.*

Tercero, y por último, una estrategia crucial que consiste en dividir los genomas en segmentos más pequeños correspondientes a genes individuales, realizando alineamientos específicos para cada uno. Esto reduce significativamente la complejidad, el coste computacional y, por tanto, el tiempo de ejecución. Las razones son las siguientes:

- Impacto en la complejidad algorítmica: El algoritmo Smith-Waterman, utilizado en esta herramienta bioinformática a través de Parasail, tienen una complejidad O(m·n). Por tanto, alinear genomas completos de 10.000 bases implicaría 100 millones de operaciones. Sin embargo, al dividir esos genomas en 10 genes de 1.000 bases y alinear cada par, cada alineamiento sería de 1 millón de operaciones. Así pues, la suma sería un total de 10 millones de operaciones, una reducción drástica comparado con el alineamiento completo inicial. Más adelante se presenta la demostración matemática que valida esta afirmación.
- Menor consumo de memoria: La matriz de programación dinámica de estos algoritmos requiere memoria proporcional a O(m·n). Al trabajar con segmentos génicos más cortos, la demanda de memoria disminuye drásticamente, haciendo el proceso más eficiente computacionalmente.

**Complejidad de un alineamiento completo (genoma entero):**

$$C_{\text{total}} = O(L_1 \cdot L_2)$$

donde $L_1$ y $L_2$ son las longitudes de los dos genomas que se van a alinear.

Si $L_1 = L_2 = L$:

$$C_{\text{total}} = O(L^2)$$

**Complejidad de alineamientos con genomas troceados (por genes):**

Consideramos $N$ segmentos de longitud $l$ cada uno, tal que $L = N \cdot l \Rightarrow l = \frac{L}{N}$

$$C_{\text{segmento}} = O(l^2)$$

complejidad de un alineamiento individual.

$$C'_{\text{total}} = N \cdot C_{\text{segmento}}$$

suma de las complejidades de todos los segmentos.

$$C'_{\text{total}} = N \cdot O(l^2)$$

$$C'_{\text{total}} = N \cdot O\left(\left(\frac{L}{N}\right)^2\right)$$

$$C'_{\text{total}} = N \cdot O\left(\frac{L^2}{N^2}\right)$$

$$C'_{\text{total}} = O\left(\frac{L^2}{N}\right)$$

**Reducción de complejidad:**

$$\frac{C_{\text{total}}}{C'_{\text{total}}} = \frac{O(L^2)}{O\left(\frac{L^2}{N}\right)} = N$$

Esta reducción de la complejidad algorítmica por un factor de N (el número de segmentos o genes) es una validación matemática fundamental de la estrategia de división del genoma. Significa que, al fragmentar el problema original de alineamiento de genomas completos en N subproblemas más pequeños, el costo computacional total se vuelve N veces menor. Esto transforma la tarea de un análisis potencialmente intratable para genomas de gran tamaño en una operación eficiente y factible, lo que es crucial para el procesamiento rápido de vastos volúmenes de datos genómicos en contextos de salud pública y biotecnología.

Esta combinación de estrategias de paralelización y optimización algorítmica consigue reducir drásticamente el tiempo de ejecución total, aprovechando al máximo las capacidades del hardware moderno actual para la investigación genómica.

### 5.3. Herramientas Adicionales de Procesamiento y Visualización de Datos

Además del algoritmo principal desarrollado en C++ para el análisis de mutaciones, se utilizaron herramientas y lenguajes de programación complementarios para el procesamiento subsecuente de los datos generados y para la visualización de los resultados.

Se empleó el lenguaje de programación Python para la creación de diversos scripts personalizados. Estos scripts fueron diseñados para interactuar con los archivos de salida en formato TSV generados por el programa de alineamiento y análisis de mutaciones descrito en el apartado anterior. Las funcionalidades implementadas mediante Python incluyen:

- Parseo y filtrado de datos: Lectura de los archivos TSV para extraer información específica, aplicar filtros adicionales según criterios definidos (por ejemplo, seleccionar solo ciertos tipos de mutaciones o genes con una cobertura mínima), y reorganizar los datos para análisis posteriores.
- Cálculos estadísticos y agregación: Realización de cálculos estadísticos descriptivos sobre los datos de mutaciones, como frecuencias de mutaciones, promedios, o agregación de resultados por gen o por tipo de mutación. También cálculos sobre el rendimiento de la herramienta bioinformática.
- Generación de gráficos y tablas: Creación de representaciones visuales de los datos, tanto del análisis genómico como del análisis de rendimiento de la propia herramienta. Para la generación de gráficos, se utilizaron librerías de Python populares como Matplotlib y Seaborn.

# 6. RESULTADOS Y DISCUSIÓN

## 6.1. Herramienta Paralela para el Alineamiento y Análisis de Secuencias

Para abordar el análisis de mutaciones en los genomas de los virus Dengue y Zika, se ha desarrollado un algoritmo bioinformático utilizando C++, lenguaje elegido por su alto rendimiento y eficiencia en la gestión de operaciones computacionalmente intensivas. Esta herramienta maximiza la velocidad de procesamiento mediante la integración de la biblioteca Parasail, que acelera los alineamientos de secuencias a través de instrucciones SIMD, y la API OpenMP, que permite la paralelización de tareas distribuyendo el análisis de cada secuencia viral de consulta entre múltiples núcleos de procesador. Siguiendo un flujo de trabajo que incluye un sondeo inicial para el cálculo de desfase y un procesamiento gen por gen, el sistema realiza alineamientos específicos y un posterior análisis detallado para identificar y caracterizar las variaciones genéticas y mutaciones.

Para evaluar la eficacia de las optimizaciones implementadas, se ha llevado a cabo un análisis de rendimiento comparando los tiempos de ejecución en diferentes configuraciones: ejecución puramente secuencial, ejecución paralela utilizando 28 núcleos, ejecución secuencial con optimizaciones SIMD, y la combinación de paralelismo y SIMD. Estas pruebas se realizaron sobre conjuntos de datos correspondientes al genoma del virus Zika y cuatro serotipos del virus del Dengue (DENV1, DENV2, DENV3 y DENV4), lo que permite observar el comportamiento de la herramienta ante diferentes tamaños de genoma y volúmenes de datos. Los resultados se centran en la reducción de los tiempos absolutos de ejecución y en el cálculo del *speedup* o ganancia de velocidad relativa.

El primer conjunto de resultados, visualizado en la Figura 8 **¡Error! No se encuentra el origen de la referencia.**, ilustra los tiempos de ejecución brutos para cada conjunto de datos y modo de operación. Se observa de manera consistente que la versión "Secuencial" (barra azul) es la que consume mayor tiempo en todos los casos, destacando especialmente en los conjuntos de datos del serotipo DENV1 y DENV2. Tiene sentido que así sea, pues son los casos en los que más secuencias problema se han tratado (ver Tabla 3). La introducción del paralelismo ("Paralelo", barra naranja) logra una reducción drástica de estos tiempos, por ejemplo, para DENV1 el tiempo disminuye de 2557 segundos a 203 segundos. De forma similar, la optimización SIMD en un solo núcleo ("Secuencial + SIMD", barra verde) también ofrece una mejora sustancial respecto a la

versión secuencial base; para ZIKV, reduce el tiempo de 191 segundos a 62 segundos, y para DENV1, de 2557 segundos a 608 segundos. Notablemente, la combinación de ambas estrategias ("Paralelo + SIMD", barra roja) resulta en los tiempos de ejecución más bajos de forma sistemática. Para ZIKV, el tiempo se reduce a tan solo 3 segundos, y para el caso más demandante, el de DENV1, el tiempo final es de 53 segundos, demostrando la eficacia de aplicar conjuntamente el paralelismo a nivel de hilos y a nivel de instrucción de CPU.

*Tabla 3. Número de secuencias problema analizadas en cada caso.*

| **Genoma Analizado** | **Número de Secuencias Problema** |
|---|---|
| ZIKV | 1093 |
| DENV1 | 10976 |
| DENV2 | 9905 |
| DENV3 | 3596 |
| DENV4 | 1440 |

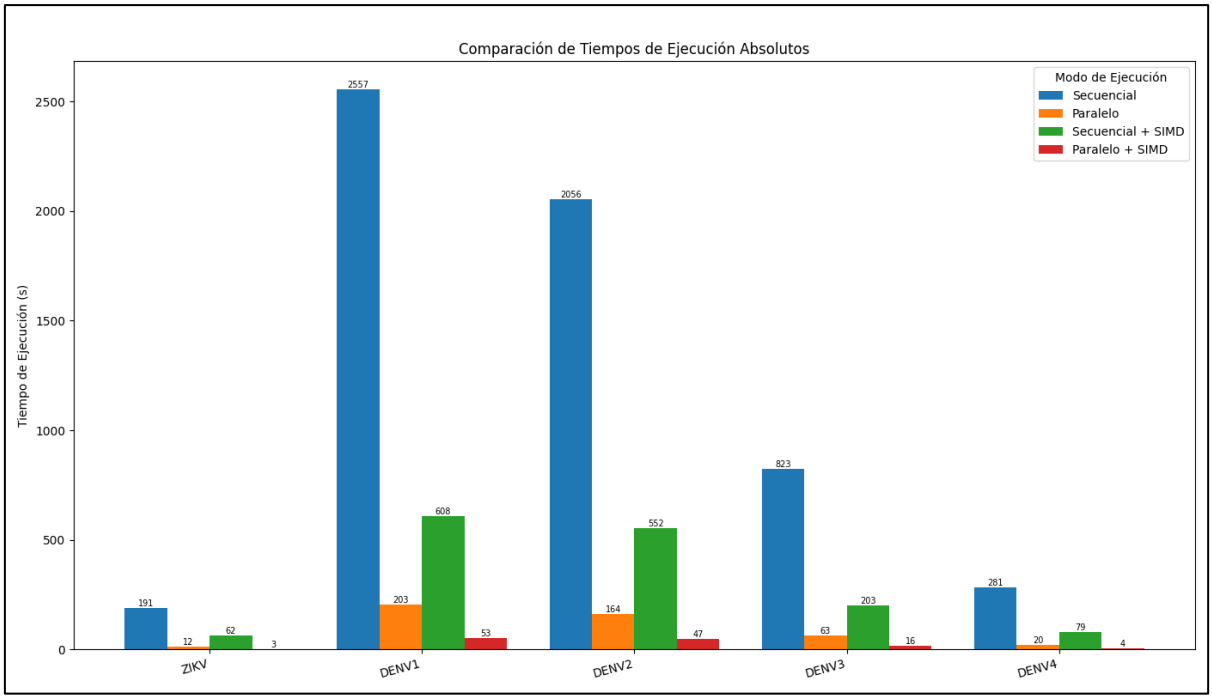


*Figura 8. Comparación de tiempos de ejecución absolutos con diferentes niveles de optimización.*

La Figura 9 cuantifica estas mejoras en términos de ganancia de velocidad relativa con respecto a la ejecución secuencial original. La "Paralelización Pura" (barras azules) muestra un *speedup* significativo, alcanzando valores entre 13x y 16x en los diferentes *datasets*. Esto indica un buen aprovechamiento de los 28 núcleos disponibles. Por su parte, el "SIMD en Secuencial" (barras naranjas) ofrece un *speedup* más modesto, generalmente entre 3x y 4x, lo cual es esperable ya que SIMD acelera principalmente la fase de alineamiento dentro de un único hilo de ejecución. El impacto más notable se observa en el "Total (Paralelo+SIMD)" (barras verdes), donde se alcanzan *speedups* muy elevados: 69x para ZIKV, 48x para DENV1, 43x para DENV2, 53x para DENV3 y 65x para DENV 4. Estos valores, que superan el número de los 28 núcleos lógicos, se explican porque este *speedup* total acumula tanto la ganancia del paralelismo multi-núcleo como la ganancia del paralelismo a nivel de instrucción (SIMD), ambas comparadas contra la línea base más lenta (secuencial sin SIMD).

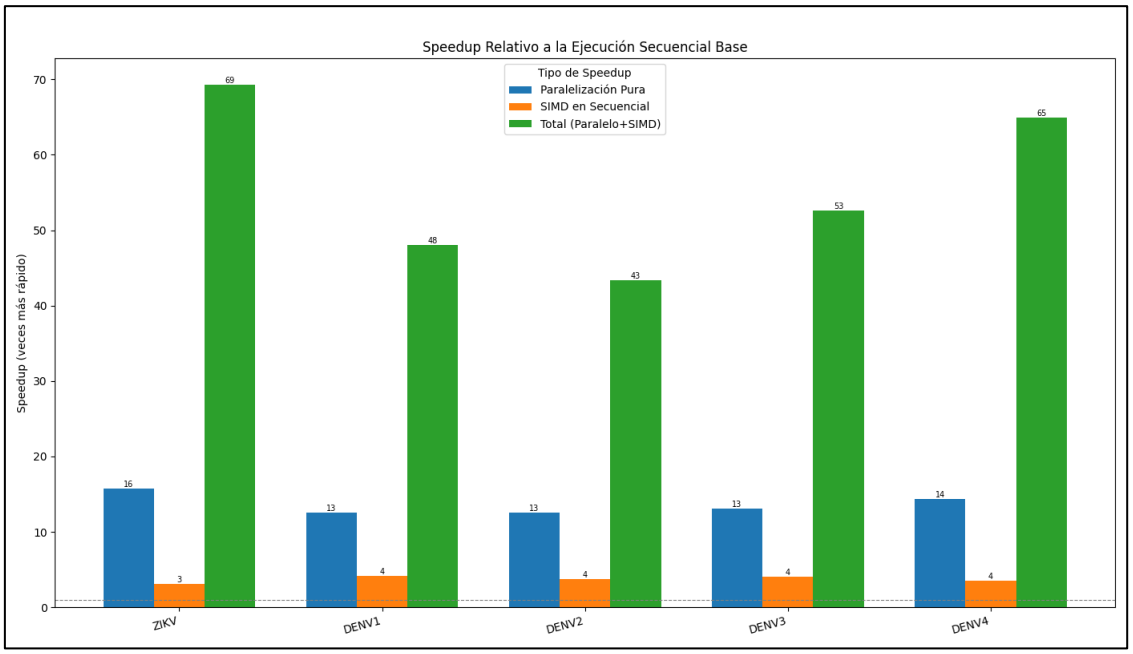


*Figura 9. Speedup comparativo de las estrategias de optimización. La línea discontinua representa un speedup de 1, que es el propio de la versión sin ningún tipo de optimización.*

Los resultados obtenidos demuestran de forma contundente la efectividad de las estrategias de optimización implementadas en la herramienta desarrollada. La paralelización mediante OpenMP, distribuyendo el procesamiento de las secuencias problema entre los 28 núcleos, es el factor que proporciona la mayor reducción inicial en

los tiempos de ejecución. El *speedup* obtenido por la paralelización pura (entre 13x y 16x) es considerable y refleja una buena escalabilidad del componente paralelizable del algoritmo, que es el procesamiento independiente de cada secuencia problema.

La incorporación de SIMD a través de la biblioteca Parasail para los alineamientos aporta una capa adicional de paralelización y, por tanto, de aceleración. Como se observa en el *speedup* del modo "SIMD en Secuencial", esta técnica puede cuadruplicar la velocidad de la fase de alineamiento en un solo núcleo. Este notable incremento de velocidad se debe a que las instrucciones SIMD permiten a un único núcleo de CPU procesar múltiples elementos de datos (conformando un vector de datos) de forma simultánea mediante una sola instrucción. Dicha paralelización resulta especialmente ventajosa en los cálculos repetitivos y matriciales característicos de los algoritmos de alineamiento.

La sinergia entre OpenMP y SIMD es evidente en los resultados del modo "Paralelo + SIMD". El *speedup* total, que supera en muchos casos las 60 veces, subraya que ambas técnicas de paralelismo (a nivel de hilo y a nivel de instrucción) son en gran medida ortogonales y sus beneficios se componen. Mientras OpenMP distribuye las tareas grandes e independientes (análisis por secuencia query), SIMD acelera las sub-tareas computacionalmente intensivas (alineamiento) dentro de cada hilo, aprovechando al máximo el hardware del sistema. Esta combinación es crucial para alcanzar tiempos de ejecución que permiten el análisis de grandes volúmenes de secuencias genómicas en plazos factibles. Las ligeras variaciones en la magnitud del *speedup* total entre los diferentes conjuntos de datos pueden atribuirse a las características intrínsecas de cada genoma (longitud, complejidad, divergencia respecto a la referencia), que pueden influir en la eficiencia relativa de la fase de alineamiento (donde SIMD destaca) versus la fase de análisis de mutaciones posterior.

Adicionalmente a estas optimizaciones basadas en el hardware, la estrategia algorítmica de procesar el genoma por genes individuales, en lugar de intentar un alineamiento global de genomas completos, constituye una tercera capa de optimización fundamental. Trocear el problema en alineamientos más pequeños y localizados para cada gen reduce la complejidad computacional y los requerimientos de memoria. Esta aproximación no solo acelera cada tarea de alineamiento individual, sino que también facilita un análisis más enfocado al interés biotecnológico eventual de los resultados (por genes) y la gestión eficiente de los datos, contribuyendo significativamente a la reducción global del tiempo de ejecución en comparación con un enfoque de alineamiento genómico integral.

### 6.2. Análisis de Mutaciones en Genomas del Virus del Zika

El análisis genómico del virus del Zika (ZIKV) es crucial para comprender su evolución, patogenicidad y capacidad de adaptación. Utilizando la herramienta bioinformática desarrollada, se procesaron un total de 1093 secuencias genómicas completas del ZIKV obtenidas de GISAID, permitiendo una caracterización detallada de los patrones mutacionales a lo largo de su genoma. Los resultados obtenidos no solo proporcionan una visión profunda de la variabilidad genética del virus, sino que también identifican mutaciones y regiones genómicas clave que podrían influir en su comportamiento biológico y servir como potenciales dianas para futuras intervenciones biotecnológicas. A continuación, se presentan y discuten los hallazgos más relevantes.

Para iniciar el análisis de mutaciones del virus del Zika, se consideró fundamental contextualizar la disponibilidad de los datos genómicos utilizados. La Figura 10 ilustra la distribución temporal de estas secuencias, ofreciendo una visión cronológica de la recopilación de datos, lo cual es esencial para comprender los periodos de mayor actividad de secuenciación y la representatividad de las variantes virales a lo largo del tiempo.

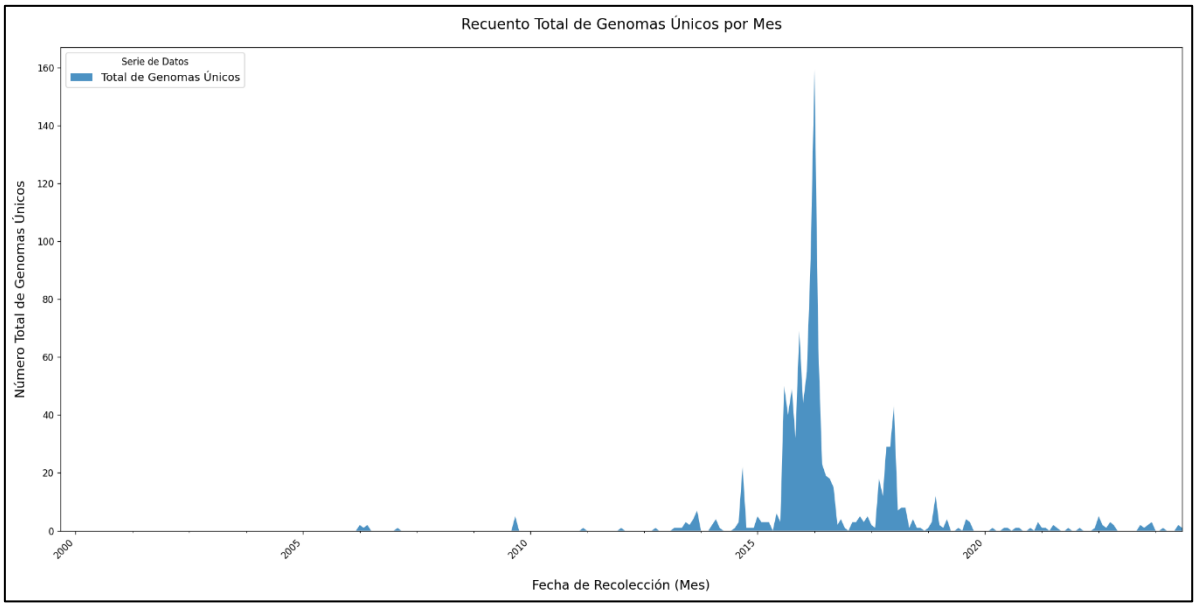


*Figura 10. Distribución temporal de las secuencias completas del genoma del virus del Zika (ZIKV) depositadas en GISAID hasta el 26 de mayo de 2025.*

La Figura 10 revela una dinámica de secuenciación del genoma de ZIKV que refleja los picos epidemiológicos y los esfuerzos de vigilancia. Se observa un marcado incremento en el número de secuencias depositadas entre 2015 y 2016, lo que coincide directamente con la epidemia global de Zika en Brasil y su subsiguiente declaración como emergencia de salud pública internacional por la OMS. Este volumen de datos durante el brote subraya los intensos esfuerzos de caracterización viral y muestreo. Posteriormente, el número de secuencias depositadas disminuye significativamente, lo cual es un patrón esperable una vez que la fase aguda de una epidemia remite y los recursos de secuenciación se reorientan. Comprender esta distribución temporal es crucial para la interpretación de los patrones mutacionales hallados, ya que la mayoría de las mutaciones analizadas en este estudio provienen de un periodo de alta presión evolutiva y expansión viral, y la falta de datos en otros momentos no necesariamente indica una ausencia de mutaciones, sino una menor actividad de secuenciación.

Una vez establecida la distribución temporal de las secuencias y el panorama general de las mutaciones de un solo nucleótido, el siguiente paso crítico en el análisis mutacional es comprender la consecuencia funcional de estas variantes genéticas a nivel de proteína. La Tabla 4 detalla la frecuencia y la prevalencia de los distintos tipos de mutaciones genómicas identificadas en el genoma del ZIKV, categorizadas según su impacto en la secuencia de aminoácidos de la poliproteína viral, así como la categoría de SNVs co-ocurrentes.

*Tabla 4. Frecuencia total y prevalencia media de los tipos de mutaciones genómicas y sus consecuencias funcionales en los genomas analizados de ZIKV.*

| **Tipo de Mutación** | **Recuento** | **Número medio de genomas (%)** |
|---|---|---|
| Con cambio de sentido | 1553 | 2 (0,18 %) |
| Sin sentido | 10 | 1030 (94 %) |
| Sinónimas | 3648 | 5 (0,46 %) |
| Not alone | 1045 | 6 (0,55 %) |

La Tabla 4 detalla la diversidad mutacional funcional del ZIKV. Las mutaciones sinónimas (3648) son las más frecuentes, lo cual es esperado al no alterar la secuencia de aminoácidos y acumularse con menor presión selectiva. Las mutaciones con cambio de sentido (1553), que cambian el aminoácido, son el segundo tipo y son relevantes por su

potencial impacto en la función viral, aunque su baja prevalencia media (0.18%) sugiere que muchas son neutras. Las mutaciones sin sentido (10) son las menos frecuentes en total debido a su severo efecto deletéreo al crear codones de parada prematuros. A pesar de esto, su alta prevalencia media (94%) en los genomas donde aparecen es llamativa. Esto podría deberse a que están situadas muy al final de un gen, lo que minimizaría el impacto en la proteína (resultando solo ligeramente más corta) y permitiría su persistencia en ciertos aislados sin una rápida eliminación por selección. Finalmente, la categoría "not alone" (1045) agrupa SNVs que co-ocurren en el mismo codón, mostrando un patrón significativo de variabilidad genética que puede tener complejas implicaciones en la evolución y funcionalidad del virus.

Tras caracterizar la cantidad y el tipo de mutaciones identificadas, es esencial comprender la variabilidad en la carga mutacional entre los distintos genomas de ZIKV. La Figura 11 presenta un *boxplot* que ilustra la distribución del número total de mutaciones detectadas por cada genoma viral individual. Este análisis proporciona una visión estadística de la heterogeneidad mutacional dentro de la población de ZIKV estudiada, complementando los conteos agregados previos.

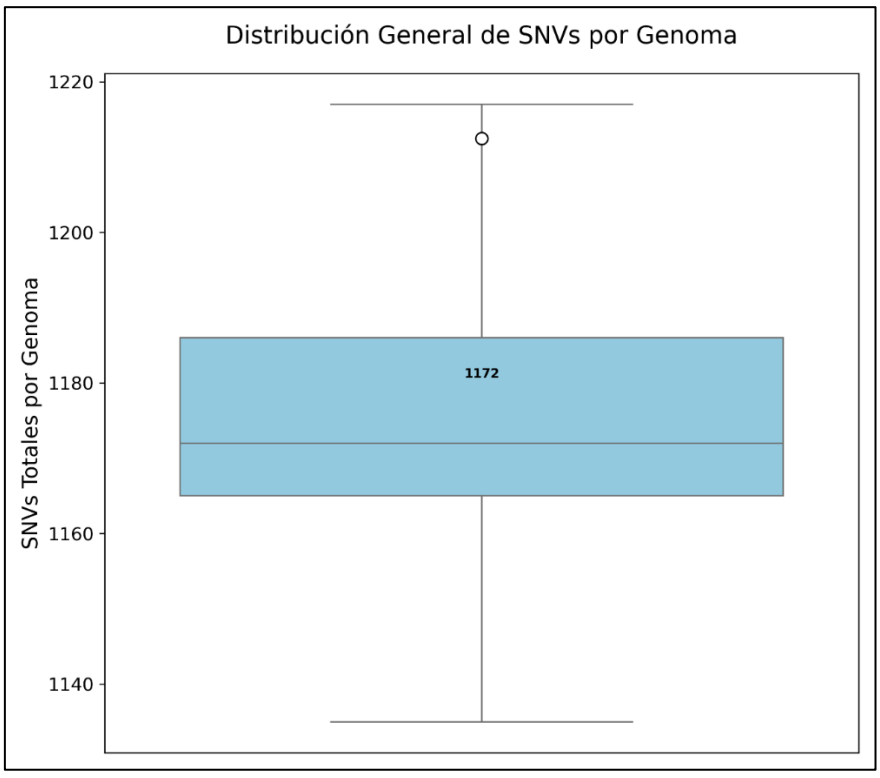


*Figura 11. Distribución del número total de mutaciones por genoma en las secuencias del virus del Zika analizadas.*

La Figura 11 visualiza la dispersión del número total de mutaciones por genoma de ZIKV. El *boxplot* muestra que la mediana de mutaciones por genoma, respecto a la secuencia de referencia, es 1172, con la mayoría de los genomas agrupados entre las 1165 y 1186 mutaciones (representando el 50% central de los datos). La media de mutaciones por genoma, representada por el punto hueco, es de 1212, situándose muy cerca de la mediana. Esta proximidad entre la media y la mediana sugiere que la distribución de mutaciones no presenta un sesgo significativo. Esta variabilidad en la carga mutacional entre genomas individuales es un reflejo de la alta tasa de mutación intrínseca de los virus ARN, que conduce a la formación de poblaciones cuasiespecies.

Más allá de la cuantificación general y la variabilidad de las mutaciones, es fundamental desglosar los patrones de sustitución nucleotídica específicos, ya que estos pueden revelar los mecanismos subyacentes de mutagénesis. La Figura 12 ilustra la distribución porcentual de los diferentes tipos de transiciones y transversiones identificadas en los genomas del virus del Zika, proporcionando una visión detallada de las "firmas" mutacionales predominantes en el virus.

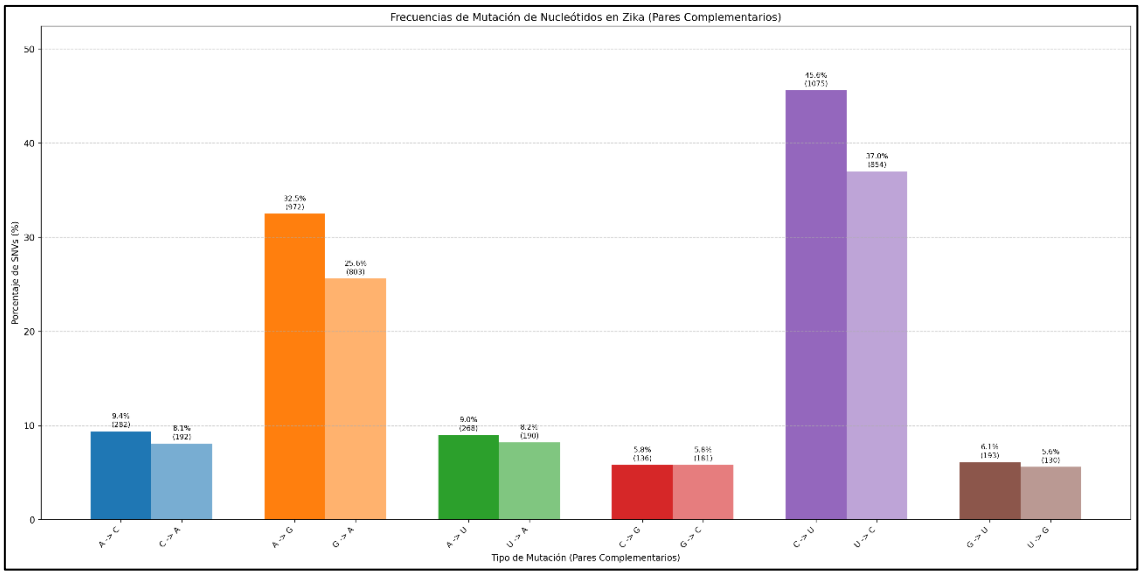


*Figura 12. Frecuencias de mutación de nucleótidos, respecto al número total de veces que aparece cada nucleótido inicial en el genoma de referencia, en el virus del Zika.*

La Figura 12 revela de manera contundente los patrones dominantes de sustitución nucleotídica en el genoma del ZIKV. Como se observa, las transiciones son más comunes que las transversiones. Específicamente, las transiciones C>U (45,6 %) y U>C (37,0 %), junto con A>G (32,5 %) y G>A (25,6 %), constituyen los picos más prominentes. Las

ocho sustituciones restantes presentan frecuencias considerablemente menores, generalmente por debajo del 5% cada una.

Estos picos en C>U, U>C; A>G, G>A reafirman de manera significativa la hipótesis de la participación de enzimas de edición del ARN del hospedador en la mutagénesis viral, un fenómeno que se ha comprobado extensamente en el SARS-CoV-2 (Di Giorgio et al., 2020) y que parece replicarse en ZIKV. Por un lado, las desaminasas de citidina de la familia APOBEC actúan desaminando citosinas (C) a uracilos (U). Si APOBEC actúa directamente sobre la cadena de ARN viral de sentido positivo (+) del ZIKV, el resultado directo es una mutación C>U. Sin embargo, los Flavivirus replican a través de una cadena intermedia de ARN de sentido negativo (-). Si APOBEC desamina una C en esta cadena negativa, esa C se convierte en U; cuando esta cadena negativa se usa como molde para sintetizar una nueva cadena de ARN positiva, el uracilo (U) se emparejará con una adenina (A), lo que resultará en un cambio de G a A en la nueva cadena de ARN positiva. De este modo, APOBEC genera tanto mutaciones C>U como G>A.

Por otro lado, las desaminasas de adenosina ADARs actúan convirtiendo adenosinas (A) en inosinas (I) en regiones de ARN de doble hebra. Dado que la inosina intruduce una citosina (C) cuando actúa como cadena molde, la acción de ADAR sobre una A en la cadena positiva resultará en una mutación A>G. Análogamente, si ADAR actúa sobre una A en la cadena de sentido negativo (-) (dentro de un dsRNA), esta A se convierte en I. Al copiarse a la cadena positiva, esta I se emparejará con una C, de modo que la mutación se manifestará como un cambio de U a C en la cadena positiva. Así, ADAR es responsable de los picos A>G y U>C.

La observación de estos pares específicos de picos es una fuerte evidencia de que estas enzimas del huésped actúan sobre ambas cadenas de ARN (positiva y negativa o intermediarios de doble hebra) durante la replicación viral.

Las demás transiciones (A>C, C>G, G>C, T>A, T>G, U>A, U>G) y todas las transversiones se observan en valores muy similares y significativamente más bajos entre sí. Estos patrones de mutación, que carecen de la especificidad y la magnitud de los picos antes mencionados, tienen sentido precisamente porque están sujetos a una pura aleatoriedad, sin una dirección mutagénica preferencial.

Tras el análisis detallado de los tipos de sustituciones nucleotídicas y la fuerte evidencia de la acción de enzimas de edición del ARN del hospedador, resulta pertinente investigar

cómo se distribuye la carga mutacional a lo largo de los diferentes genes del ZIKV. Esta perspectiva puede revelar si ciertas regiones genómicas son más propensas a la variación o si diferentes tipos de mutaciones se acumulan preferentemente en genes específicos. La Figura 13 presenta las tasas de mutación normalizadas por cada 100 nucleótidos para cada gen del ZIKV, desglosadas por el tipo de consecuencia funcional de la mutación (sinónima, con cambio de sentido, sin sentido, inserción, deleción y mutaciones en UTR).

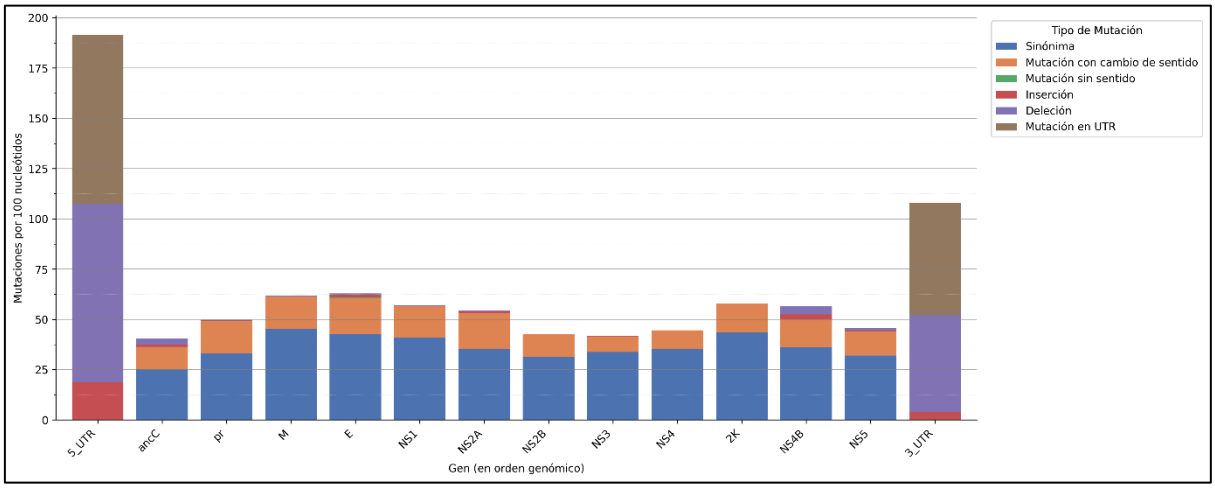


*Figura 13. Tasas de mutación por gen en el genoma del virus del Zika.*

El análisis de la distribución de mutaciones por gen en el ZIKV (Figura 13) revela patrones de variabilidad heterogéneos a lo largo del genoma, destacando regiones con tasas mutacionales extremas. Las regiones no codificantes 5'UTR y 3'UTR presentan las tasas de eventos de mutación únicos más elevadas (191,51 y 107,94 mutaciones por 100 nucleótidos, respectivamente), compuestas casi en su totalidad por sustituciones (Mutación en UTR) y deleciones. Esta alta plasticidad en las UTRs refleja una menor presión de selección purificadora en comparación con las regiones codificantes.

Entre los genes codificantes, E y M destacan por presentar las tasas mutacionales más altas (62,93 y 61,78 mutaciones/100 nt, respectivamente). Esta elevada variabilidad, con una notable contribución de mutaciones con cambio de sentido, es esperable para proteínas de superficie que están en constante interacción con el sistema inmune del hospedador. La presión selectiva para evadir la respuesta inmune podría impulsar la fijación de cambios aminoacídicos en estas proteínas. Este hallazgo se alinea con la discusión previa sobre la proteína E como principal inductor de anticuerpos y su variabilidad como determinante de la diversidad del ZIKV.

En el extremo opuesto, varios genes que codifican proteínas estructurales internas o enzimas virales críticas exhiben tasas de mutación considerablemente menores, lo que sugiere una fuerte conservación funcional. Entre ellos se encuentran ancC (cápside, 40.44 mutaciones/100 nt), NS2B (cofactor de la proteasa, 42.56 mutaciones/100 nt), NS3 (proteasa/helicasa, 41.92 mutaciones/100 nt), NS4 (implicado en complejos de replicación, 44.62 mutaciones/100 nt) y NS5 (polimerasa/metiltransferasa, 45.69 mutaciones/100 nt). La proteína C, por ejemplo, requiere interacciones precisas para el empaquetamiento del ARN y el autoensamblaje de la nucleocápside. De forma similar, las proteínas NS2B, NS3, NS4A y NS5 desempeñan funciones enzimáticas o de andamiaje vitales (procesamiento de la poliproteína, replicación, modificación del ARN y formación de complejos de replicación) que dependen de estructuras tridimensionales y sitios activos altamente conservados. Los cambios aminoacídicos disruptivos en estas proteínas probablemente son deletéreos y eliminados por selección purificadora, lo que explica su menor variabilidad en contraste con las proteínas de superficie más expuestas a la presión inmune.

Respecto a los tipos de mutación dentro de los genes codificantes, la predominancia de mutaciones sinónimas, seguidas por las de cambio de sentido, y finalmente por las mutaciones sin sentido, inserciones y deleciones (estas últimas con muy baja frecuencia) es un patrón común en la evolución viral. Las mutaciones sinónimas, al no cambiar el aminoácido, suelen tener un impacto fenotípico menor y se acumulan más fácilmente. Las mutaciones con cambio de sentido, que sí alteran el aminoácido, son filtradas más rigurosamente por la selección, al igual que las mutaciones sin sentido, inserciones y deleciones, ya que estas últimas suelen tener consecuencias aún más drásticas sobre la estructura y función de las proteínas, llevando a menudo a productos no funcionales o truncados que son fuertemente seleccionados en contra.

### 6.3. Análisis de Mutaciones en Genomas del Virus del Dengue

Continuando con el análisis, esta sección se centra en el virus del Dengue. Utilizando la misma herramienta bioinformática, se procesaron un total de 25.917 secuencias genómicas completas de DENV obtenidas de GISAID, distribuidas entre sus cuatro serotipos: 10.976 secuencias de DENV1, 9.905 de DENV2, 3.596 de DENV3 y 1.440 de DENV4. El objetivo fue caracterizar sus perfiles mutacionales, identificar variantes de interés y comparar los hallazgos con los observados en el ZIKV, buscando así patrones de variabilidad genética específicos y compartidos.

Para comenzar este análisis del DENV, es fundamental entender la disponibilidad y la distribución temporal de los datos genómicos con los que se ha trabajado. La Figura 14 ilustra el recuento de genomas únicos de Dengue depositados en GISAID, desglosados por mes de recolección y por serotipo, lo que proporciona un contexto crucial sobre la actividad de secuenciación y la prevalencia de los diferentes serotipos a lo largo del tiempo.

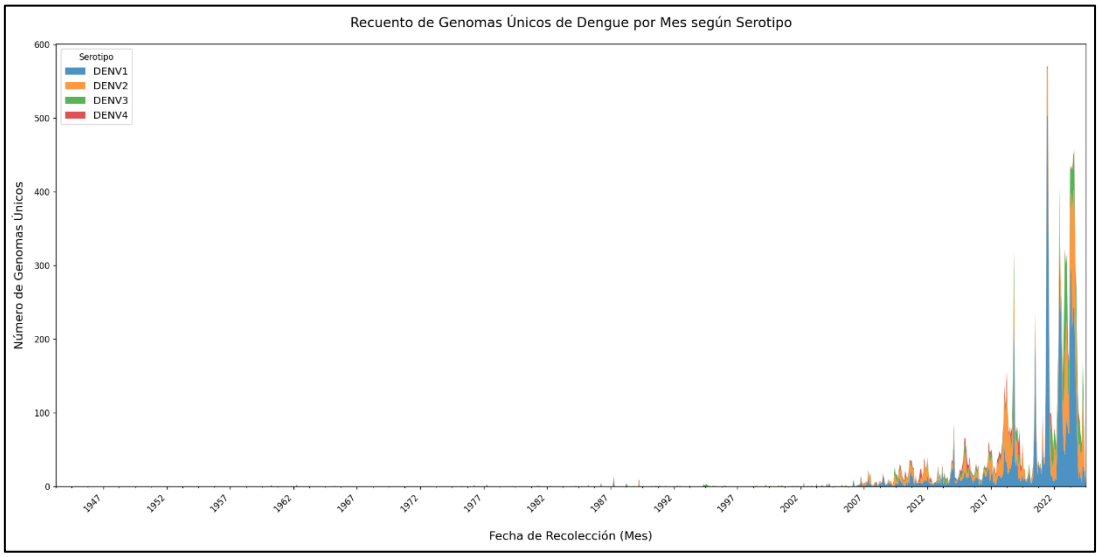


*Figura 14. Distribución temporal de las secuencias completas del genoma del virus del Dengue (DENV) depositadas en GISAID hasta el 30 de abril de 2025.*

La Figura 14 muestra un marcado incremento en la deposición de secuencias genómicas del DENV a partir de la década de 2010, con picos notables en años recientes, reflejando la creciente incidencia global del dengue y la mejora en la vigilancia genómica. A diferencia del patrón más concentrado del ZIKV en torno a su principal epidemia, el DENV presenta una actividad de secuenciación más sostenida y con fluctuaciones en la predominancia de sus cuatro serotipos. Esta distribución temporal y por serotipo es crucial para contextualizar los análisis mutacionales, ya que la diversidad genética observada estará influenciada por la representatividad de cada serotipo y el periodo del que provienen las secuencias, subrayando la naturaleza endémica y epidémica recurrente del dengue.

Una vez comprendida la distribución temporal y por serotipo de las secuencias del virus del Dengue (DENV), el siguiente paso es analizar las consecuencias funcionales de las variaciones genéticas identificadas. De forma similar al análisis realizado para el ZIKV,

la Tabla 5 resume la frecuencia total y la prevalencia mediana de los diferentes tipos de mutaciones genómicas encontradas en los genomas de DENV procesados, clasificadas según su impacto en la poliproteína viral y la co-ocurrencia de SNVs.

*Tabla 5. Frecuencia total y prevalencia media de los tipos de mutaciones genómicas y sus consecuencias funcionales en los genomas analizados de DENV.*

| **Tipo de Mutación** | **Recuento** | **Número medio de genomas (%)** |
|---|---|---|
| Con cambio de sentido | 14633 | 36 (0,14 %) |
| Sin sentido | 326 | 1351 (5,2 %) |
| Sinónimas | 7187 | 1617 (6,2 %) |
| Not alone | 11938 | 279 (1,1 %) |

La Tabla 5 revela que, al igual que en ZIKV, las mutaciones con cambio de sentido en DENV (14633) muestran una prevalencia media extremadamente baja (0.14% vs. 0.18% en ZIKV), sugiriendo una fuerte selección purificadora en ambos virus a pesar del mayor recuento absoluto en DENV, posiblemente debido a más secuencias o mayor diversidad serotípica. A diferencia de ZIKV donde eran mayoritarias, las mutaciones sinónimas en DENV (7187) son el segundo tipo más frecuente por recuento, con una prevalencia media (6.2%) superior a la de ZIKV (0.46%), lo que indica una mayor acumulación, aunque sin alta fijación individual. Las mutaciones sin sentido en DENV (326), aunque menos comunes que en ZIKV (10), presentan una prevalencia media en genomas afectados (5.2%) considerablemente menor que el llamativo 94% de ZIKV. Finalmente, la categoría "not alone" (SNVs co-ocurrentes) es mucho más numerosa en DENV (11938 vs. 1045 en ZIKV) aunque con prevalencia media similarmente baja (1.1% vs. 0.55%), lo que podría reflejar una mayor complejidad en la vinculación de mutaciones en DENV debido a su mayor diversidad global y la cocirculación de múltiples serotipos.

Tras examinar la naturaleza y frecuencia de los tipos de mutaciones funcionales en el conjunto de los serotipos del DENV (Tabla 5), es importante investigar la variabilidad en la carga mutacional total por genoma entre los diferentes serotipos. Esta perspectiva nos permite entender si ciertos serotipos tienden a acumular más o menos variantes nucleotídicas en general, especialmente considerando que todas las secuencias han sido comparadas contra una única secuencia de referencia del serotipo DENV2.

La Figura 15 presenta diagramas de caja que ilustran la distribución del número total de SNVs detectadas por genoma para cada uno de los cuatro serotipos de Dengue.

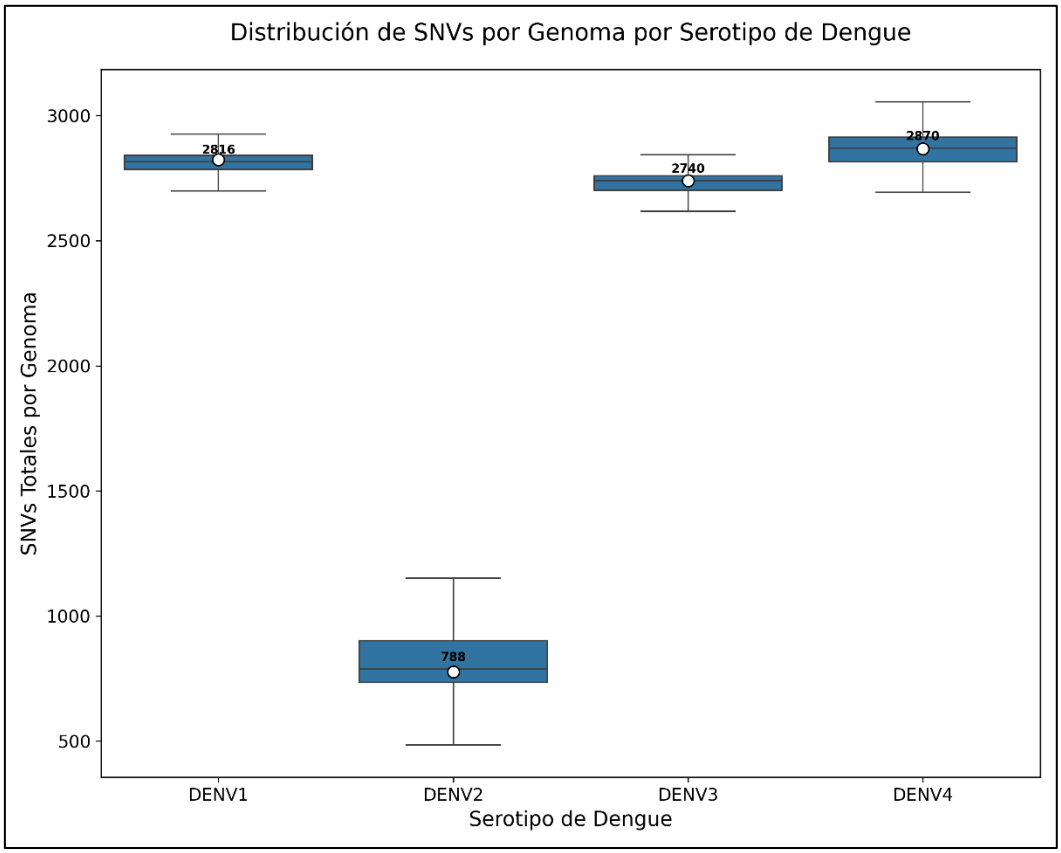


*Figura 15. Distribución del número total de SNVs por genoma para cada serotipo de DENV.*

La Figura 15 evidencia la divergencia genética entre los serotipos del virus del Dengue al compararlos con una referencia DENV2. Como es de esperar, DENV2 muestra la menor cantidad de SNVs (mediana de 788), reflejando su variabilidad intra-serotipo respecto a la cepa de referencia específica. En marcado contraste, DENV1 (mediana 2816 SNVs), DENV3 (mediana 2740 SNVs) y DENV4 (mediana 2870 SNVs) presentan cargas mutacionales significativamente más altas, lo que cuantifica su distancia evolutiva acumulada frente a DENV2. La dispersión relativamente compacta de SNVs en DENV1 y DENV3 (IQR de 57 para ambos) sugiere una distancia genética más homogénea de estos serotipos hacia DENV2, mientras que DENV4 (IQR 97.25) muestra una variabilidad interna ligeramente mayor en este aspecto. Estos valores no indican necesariamente tasas de mutación intrínsecas diferentes, sino la diversidad acumulada

inter-serotípica, con medianas y medias muy cercanas dentro de cada grupo, sugiriendo distribuciones bastante simétricas de SNVs por genoma al usar DENV2 como referencia.

Tras evaluar la carga mutacional entre serotipos, es crucial investigar los tipos específicos de sustituciones nucleotídicas para entender los mecanismos mutagénicos subyacentes en el DENV. Al igual que en el ZIKV, se espera que ciertos patrones de sustitución puedan revelar la influencia de factores como las enzimas de edición del ARN del hospedador. La Figura 16 ilustra la frecuencia porcentual global de los doce tipos posibles de sustituciones nucleotídicas (considerandos pares complementarios) identificadas en el conjunto de genomas de DENV analizados.

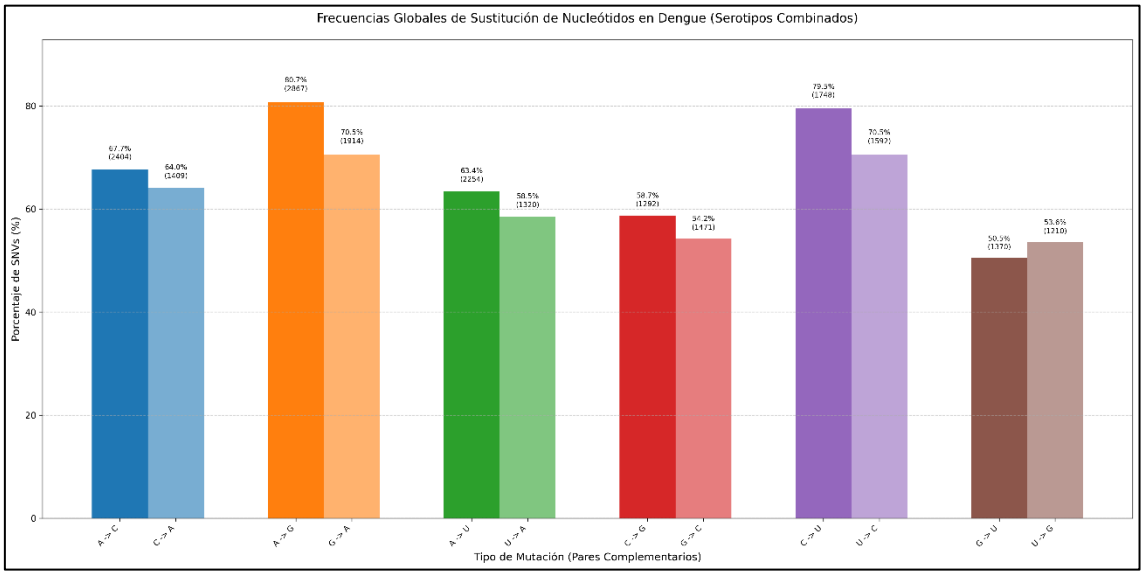


*Figura 16. Frecuencias de mutación de nucleótidos, respecto al número total de veces que aparece cada nucleótido inicial en el genoma de referencia, en el virus del Zika (serotipos combinados).*

La Figura 16 revela que en DENV, al igual que en ZIKV, las transiciones son mucho más frecuentes que las transversiones, destacando los picos A->G (80.7%), U->C (70.5%), C->U (79.5%) y G->A (70.5%). Estas mismas cuatro transiciones dominaron en ZIKV, aunque en DENV las frecuencias de A->G y U->C parecen ser relativamente mayores. Las demás sustituciones muestran frecuencias considerablemente menores en ambos virus.

Esta predominancia de A->G/U->C y C->U/G->A en DENV refuerza la hipótesis, ya discutida para ZIKV, de la acción de enzimas de edición del ARN del hospedador: ADARs (responsables de A->G y U->C) y APOBECs (responsables de C->U y G->A).

La magnitud de estos picos en DENV sugiere un impacto significativo de estos mecanismos en su evolución, similar a ZIKV, y la menor frecuencia del resto de sustituciones apunta a la tasa de error basal de la polimerasa y mutaciones aleatorias.

Tras analizar los patrones generales de sustitución nucleotídica, es fundamental desglosar cómo se distribuyen estas mutaciones y sus consecuencias funcionales a lo largo de los diferentes genes del DENV. Este análisis permite identificar regiones genómicas con mayor o menor variabilidad y comparar estos perfiles con los observados en ZIKV. La Figura 17 presenta las tasas de mutación normalizadas por cada 100 nucleótidos para cada gen y región UTR del DENV, desglosadas por tipo de mutación.

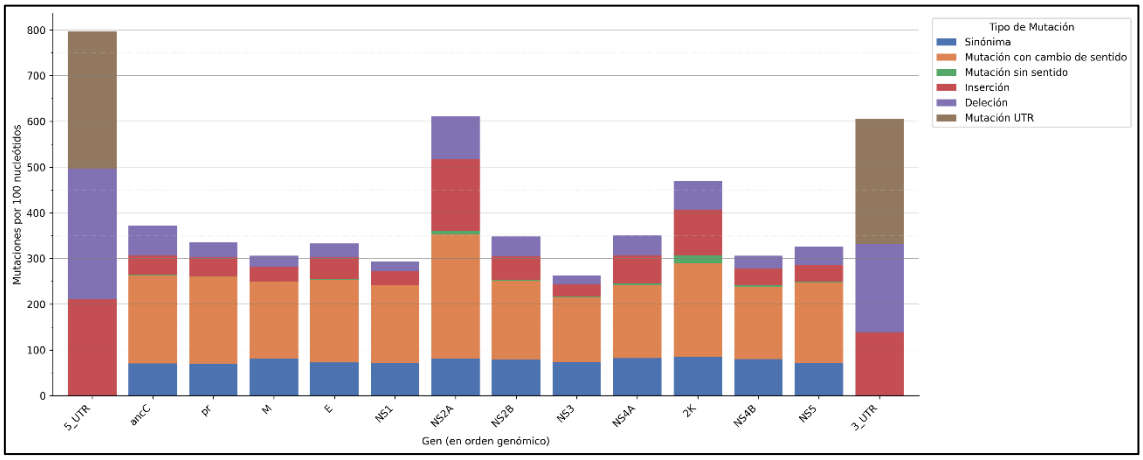


*Figura 17. Tasas de mutación por gen en el genoma del virus del dengue (serotipos combinados).*

El análisis de la distribución de mutaciones por gen en DENV (Figura 17) revela una heterogeneidad significativa en la carga mutacional a lo largo de su genoma. En términos generales, las tasas de eventos de mutación únicos observadas en DENV son considerablemente más altas que las detectadas en ZIKV (comparado con la Figura 13 para Zika). Esta diferencia cuantitativa se correlaciona directamente con el volumen de secuencias analizadas, que es sustancialmente mayor para DENV (25.917 genomas) en comparación con ZIKV (1.093 genomas), lo que naturalmente permite la detección de un espectro más amplio de variantes. Al igual que en ZIKV, las regiones no codificantes 5'UTR y 3'UTR son las que acumulan la mayor densidad de mutaciones (796,88 y 605,54 mutaciones por 100 nucleótidos, respectivamente), subrayando la alta plasticidad de estas regiones reguladoras en ambos flavivirus.

En relación con los genes codificantes de DENV, se observa una variabilidad distintiva. Los genes NS2A (611,77 mutaciones/100 nt) y el pequeño péptido 2K (469,57 mutaciones/100 nt) destacan como los más variables. La elevada tasa en 2K es probablemente influenciada por su corta longitud, donde un número moderado de mutaciones resulta en una tasa normalizada alta. La alta tasa en NS2A podría indicar una menor restricción estructural o una fuerte presión selectiva adaptativa, superando la variabilidad de genes como E o M que eran más prominentes en ZIKV. En contraste, genes que codifican enzimas virales esenciales como NS3 (262,03 mutaciones/100 nt) y NS1 (293,94 mutaciones/100 nt) exhiben las tasas de mutación más bajas entre los genes codificantes de DENV analizados. Esta menor variabilidad relativa es consistente con la necesidad de una mayor conservación funcional, similar a la lógica aplicada para los genes conservados de ZIKV (como ancC, NS3 y NS5), aunque las tasas absolutas en DENV siguen siendo más elevadas, reflejo del ya mencionado mayor tamaño muestral y la diversidad inherente a los cuatro serotipos.

Finalmente, al examinar los tipos de mutación, el gran volumen de datos de DENV destaca el comportamiento de las mutaciones sinónimas (barras azules en la Figura 17). Al no cambiar aminoácidos, estas mutaciones están menos restringidas por la selección y su frecuencia refleja más directamente la tasa de error de la polimerasa y la deriva genética, operando de forma más puramente aleatoria. Esto resulta en una contribución relativamente constante de mutaciones sinónimas a la carga mutacional total en los genes codificantes del DENV. La gran cantidad de secuencias analizadas simplemente hace que este "ruido de fondo" de variabilidad silente se manifieste con notable consistencia, un patrón menos evidente en ZIKV debido a su muestreo más limitado. En contraste con ZIKV, DENV también muestra una mayor proporción de inserciones y deleciones en varios genes, especialmente NS2A, 2K y NS4B, probablemente por una mejor detección de estos eventos más raros gracias al mayor tamaño muestral.

# 7. CONCLUSIONES

Este Trabajo de Fin de Grado se centró en analizar la variabilidad genómica de los virus Zika (ZIKV) y Dengue (DENV) para identificar patrones mutacionales clave útiles en el diseño de estrategias biotecnológicas, partiendo de la hipótesis de que dicho análisis masivo aportaría información valiosa para este fin. Los resultados respaldan esta hipótesis y cumplen los objetivos planteados.

Se diseñó e implementó con éxito un algoritmo computacional paralelo en C++, demostrando mediante pruebas de rendimiento que la combinación de OpenMP, SIMD (Parasail) y una estrategia de análisis segmentado por genes reduce drásticamente los tiempos de ejecución, con *speedups* totales de hasta 69x para ZIKV y entre 43x-65x para DENV, validando su eficiencia para grandes volúmenes de datos.

La comparación de estos perfiles reveló que ambos virus, a pesar de diferencias en las tasas generales de mutación, comparten una fuerte selección purificadora sobre mutaciones con cambio de sentido y patrones de sustitución nucleotídica (C>U, U>C, A>G, G>A) consistentes con la acción de enzimas de edición del ARN del hospedador (APOBEC/ADAR). Se identificó una variabilidad heterogénea por gen, con UTRs y proteínas de superficie (E, M en ZIKV; NS2A, 2K en DENV) mostrando mayor plasticidad, lo que posiciona a estas proteínas como puntos de interés para el desarrollo de vacunas. En contraste con esta plasticidad, se observó que genes estructurales internos y enzimáticos cruciales en ambos virus exhibieron una notable conservación funcional. Específicamente, en el virus del Zika, genes como ancC, NS2B, NS3, NS4 y NS5 mostraron tasas de mutación considerablemente menores, indicativo de una fuerte presión selectiva para mantener su función. De manera similar, en el virus del Dengue, los genes que codifican enzimas virales esenciales como NS3 y NS1 destacaron por presentar las tasas de mutación más bajas entre sus regiones codificantes. Esta marcada conservación en genes clave de ambos patógenos los confirma como objetivos especialmente prometedores para el diseño de fármacos antivirales de amplio espectro, ya que las mutaciones en estos sitios serían probablemente deletéreas para el virus.

En conjunto, la herramienta desarrollada y los análisis realizados han proporcionado información valiosa sobre la dinámica evolutiva de ZIKV y DENV, cumpliendo el objetivo de identificar patrones relevantes para futuras intervenciones biotecnológicas.

## 8. BIBLIOGRAFÍA

Antonios, F., & Daelemans Jury members Piet Maes Paul Proost Kevin Arien Jan Munch, D. (2021). *NEW THERAPEUTIC APPROACHES AGAINST EMERGING FLAVIVIRUSES*.

Bhandari, V., Taksande, A. B., Sapkale, B., Bhandari, V., Taksande, A. B., & Sapkale, B. (2023). Disease Transmission and Diagnosis of Zika Virus. *Cureus*, *15*(11). https://doi.org/10.7759/CUREUS.49263

Bhutkar, M., Singh, V., Dhaka, P., & Tomar, S. (2022). Virus-host protein-protein interactions as molecular drug targets for arboviral infections. *Frontiers in Virology*, *2*, 959586. https://doi.org/10.3389/FVIRO.2022.959586/BIBTEX

BOULDJEDJE, S., LEBCIR, R., & AMOURA, H. (2019). *In silico characterization of Zika virus envelope protein structure (2019)*. Centre Universitaire Abdelhafid BOUSSOUF de Mila.

Brady, O., Lim, A., Shearer, F., Sewalk, K., Pigott, D., Clarke, J., Ghouse, A., Judge, C., Kang, H., Messina, J., Kraemer, M., Gaythorpe, K., de Souza, W., Nsoesie, E., Celone, M., Faria, N., Ryan, S., Rabe, I., Rojas, D., … Golding, N. (2024). *The overlapping global distribution of dengue, chikungunya, Zika and yellow fever*. https://doi.org/10.21203/RS.3.RS-4686814/V1

Contreras García, H. A. (2020). *Diseño y optimización de estrategias de detección y tipificación de aislados de virus de Dengue y Zika* [Tesis de licenciatura]. Universidad Nacional Autónoma de México.

Contreras, M., Vásquez Guillén, A., Rincón, M. A., Moreira, R., & Callejas, D. (2021a). Aspectos genéticos del virus del dengue. *QhaliKay. Revista de Ciencias de La Salud ISSN: 2588-0608*, *5*(2), 79. https://doi.org/10.33936/QKRCS.V5I2.3496

Contreras, M., Vásquez Guillén, A., Rincón, M. A., Moreira, R., & Callejas, D. (2021b). Aspectos genéticos del virus del dengue. *QhaliKay. Revista de Ciencias de La Salud ISSN: 2588-0608*, *5*(2), 79. https://doi.org/10.33936/QKRCS.V5I2.3496

Cordo, S. M. (2020). *RNA virus, emergencia y coronavirus*.

Côrtes, N., Lira, A., Prates-Syed, W., Dinis Silva, J., Vuitika, L., Cabral-Miranda, W., Durães-Carvalho, R., Balan, A., Cabral-Marques, O., & Cabral-Miranda, G. (2023). Integrated control strategies for dengue, Zika, and Chikungunya virus infections. *Frontiers in Immunology*, *14*, 1281667. https://doi.org/10.3389/FIMMU.2023.1281667/XML/NLM

Daily, J. (2016). Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics*, *17*(1), 81. https://doi.org/10.1186/s12859-016-0930-z

Dennehy, J. J. (2017). Evolutionary ecology of virus emergence. *Annals of the New York Academy of Sciences*, *1389*(1), 124–146. https://doi.org/10.1111/NYAS.13304;CSUBTYPE:STRING:SPECIAL;PAGE:STRING:ARTICLE/CHAPTER

Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G., & Conticello, S. G. (2020). Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances*, *6*(25), 5813–5830. https://doi.org/10.1126/SCIADV.ABB5813/SUPPL_FILE/ABB5813_SM.PDF

Diani, E., Lagni, A., Lotti, V., Tonon, E., Cecchetto, R., & Gibellini, D. (2023). Vector-Transmitted Flaviviruses: An Antiviral Molecules Overview. *Microorganisms 2023, Vol. 11, Page 2427*, *11*(10), 2427. https://doi.org/10.3390/MICROORGANISMS11102427

Dutta, S. K., & Langenburg, T. (2023). A Perspective on Current Flavivirus Vaccine Development: A Brief Review. *Viruses*, *15*(4), 860. https://doi.org/10.3390/V15040860/S1

Dwivedi, V. D., Tripathi, I. P., Tripathi, R. C., Bharadwaj, S., & Mishra, S. K. (2017). Genomics, proteomics and evolution of dengue virus. *Briefings in Functional Genomics*, *16*(4), 217–227. https://doi.org/10.1093/bfgp/elw040

Ekins, S., Liebler, J., Neves, B. J., Lewis, W. G., Coffee, M., Bienstock, R., Southan, C., & Andrade, C. H. (2016). Illustrating and homology modeling the proteins of the Zika virus. *F1000Research 2016 5:275*, *5*, 275. https://doi.org/10.12688/f1000research.8213.2

Feng, F., Ma, Y., Qin, P., Zhao, Y., Liu, Z., Wang, W., & Cheng, B. (2024). Temperature-Driven Dengue Transmission in a Changing Climate: Patterns, Trends, and Future Projections. *GeoHealth*, *8*(10), e2024GH001059. https://doi.org/10.1029/2024GH001059

Giné, F. (2020). *Virus Zika: ¿Qué sabemos sobre las causas del síndrome congénito y las complicaciones neurológicas que ocasiona?* [Trabajo de Fin de Máster]. Universidad de La Laguna.

González Almanza, L. K. (2018). *ZIKA Y SU RELACIÓN CON LA MICROCEFALIA* [Monografía de pregrado]. Universidad de los Andes.

Grubaugh, N. D. (2016). *DYNAMICS OF WEST NILE VIRUS EVOLUTION DURING INFECTION OF WILD BIRDS, MOSQUITOES, AND THE HUMAN BRAIN: UNRAVELING THE COMPELEXITIES OF SELECTION, DRIFT, AND FITNESS* [Doctoral dissertation]. Colorado State University.

Guanche Garcell, H., Gutiérrez García, F., Ramirez Nodal, M., Ruiz Lozano, A., Pérez Díaz, C. R., González Valdés, A., & Gonzalez Alvarez, L. (2020). Clinical relevance of Zika

symptoms in the context of a Zika Dengue epidemic. *Journal of Infection and Public Health*, *13*(2), 173–176. https://doi.org/10.1016/J.JIPH.2019.07.006

Gupta, A. K., Kaur, K., Rajput, A., Dhanda, S. K., Sehgal, M., Khan, M. S., Monga, I., Dar, S. A., Singh, S., Nagpal, G., Usmani, S. S., Thakur, A., Kaur, G., Sharma, S., Bhardwaj, A., Qureshi, A., Raghava, G. P. S., & Kumar, M. (2016). ZikaVR: An Integrated Zika Virus Resource for Genomics, Proteomics, Phylogenetic and Therapeutic Analysis. *Scientific Reports 2016 6:1*, *6*(1), 1–16. https://doi.org/10.1038/srep32713

Higuera, A., & Ramírez, J. D. (2019). Molecular epidemiology of dengue, yellow fever, Zika and Chikungunya arboviruses: An update. *Acta Tropica*, *190*, 99–111. https://doi.org/10.1016/J.ACTATROPICA.2018.11.010

Jablunovsky, A. ;, Jose, J., Savic, V., Jablunovsky, A., & Jose, J. (2024). The Dynamic Landscape of Capsid Proteins and Viral RNA Interactions in Flavivirus Genome Packaging and Virus Assembly. *Pathogens 2024, Vol. 13, Page 120*, *13*(2), 120. https://doi.org/10.3390/PATHOGENS13020120

Kinney, R. M., Butrapet, S., Chang, G. J., Tsuchiya, K. R., Roehrig, J. T., Bhamarapravati, N., & Gubler, D. J. (1997). Construction of infectious cDNA clones for dengue 2 virus: strain 16681 and its attenuated vaccine derivative, strain PDK-53. *Virology*, *230*(2), 300–308.

Koile, D. I. (2022). *Plataforma computacional de análisis genómico y diagnóstico de enfermedades raras* [Tesis Doctoral]. Universidad de Buenos Aires.

Kumar, A., Kumar, D., Jose, J., Giri, R., & Mysorekar, I. U. (2022). Drugs to limit Zika virus infection and implication for maternal-fetal health. *Frontiers in Virology*, *2*, 928599. https://doi.org/10.3389/FVIRO.2022.928599/XML/NLM

Kuno, G., & Chang, G. J. J. (2007). Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. *Archives of Virology*, *152*(4), 687–696. https://doi.org/10.1007/s00705-006-0903-z

Lauring, A. S. (2020). Within-Host Viral Diversity: A Window into Viral Evolution. *Annual Review of Virology*, *7*(Volume 7, 2020), 63–81. https://doi.org/10.1146/ANNUREV-VIROLOGY-010320-061642/CITE/REFWORKS

Liang, Y., & Dai, X. (2024). The global incidence and trends of three common flavivirus infections (Dengue, yellow fever, and Zika) from 2011 to 2021. *Frontiers in Microbiology*, *15*, 1458166. https://doi.org/10.3389/FMICB.2024.1458166/BIBTEX

Lorenza Trabalzini, P., & Pini Filippo Dragoni SUPERVISOR Maurizio Zazzi, A. (2019). *Antiviral drug development for treatment of acute and chronic viral infections COORDINATOR OF THE DOCTORAL SCHOOL*.

Martin Reyes-Baque, J. I., Apolo-Pincay, A. I., & Josefina Valero-Cedeño, N. I. (2020). *Factores ambientales y climáticos de la provincia de Manabí y su asociación a la presencia de las Ciencias de la salud Artículo de investigación*. *46*, 453–488. https://doi.org/10.23857/pc.v5i6.1507

Mittal, S., Federman, H. G., Sievert, D., & Gleeson, J. G. (2022). The Neurobiology of Modern Viral Scourges: ZIKV and COVID-19. *Neuroscientist*, *28*(5), 438–452. https://doi.org/10.1177/10738584211009149/ASSET/ED6B3D82-23F9-4D7B-BDCB-A8AE5CC4E402/ASSETS/IMAGES/LARGE/10.1177_10738584211009149-FIG4.JPG

Morgan, J., Strode, C., & Salcedo-Sora, J. E. (2021). Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia. *PLOS Neglected Tropical Diseases*, *15*(3), e0009259. https://doi.org/10.1371/JOURNAL.PNTD.0009259

Naiouf, M. ; D. G. A. ; D. G. L. ; C. F. ; S. V. ; P. A. ; R. E. ; B. M. J. ; S. M. ; C. M. ; G. S. ; F. E. ; G. A. ; C. S. L. (2022). *PARALELIZACION DE ALGORITMOS Y EVALUACION DE RENDIMIENTO EN PLATAFORMAS DE CÓMPUTO DE ALTAS PRESTACIONES*.

Ortega Pérez, C. A., Rivera, N. R., Sandoval López, X., Enrique, C., & Ávila, H. (2022). Origen del perfil de mutaciones presente en las secuencias de SARS-CoV-2 en El Salvador. *Revista Minerva*, *5*(2), 9–22. https://doi.org/10.5377/REVMINERVA.V5I2.15799

Pérez Bernal, M. de J. Eva. (2019). *Aislamiento y tipificación molecular de flavivirus silvestres circulantes en el estado de Yucatán* [Tesis de licenciatura]. Universidad Nacional Autónoma de México.

Ramírez Corona, L. D. (2023). *Comparación In silico de la diversidad y evolución del viroma de los mosquitos vectores Aedes aegypti y Culex pipens* [Tesis de licenciatura]. Universidad Autónoma del Estado de Morelos.

Roy, A., Liu, Q., Yang, Y., Debnath, A. K., & Du, L. (2024). Envelope Protein-Targeting Zika Virus Entry Inhibitors. *International Journal of Molecular Sciences 2024, Vol. 25, Page 9424*, *25*(17), 9424. https://doi.org/10.3390/IJMS25179424

Suárez González, D., & Romero Béjar, J. L. (2024). *TÉCNICAS MULTIVARIANTES PARA EL ANÁLISIS DE DATOS GENÓMICOS*.

Tampere, M. (2021). *ADVANCING ANTIVIRAL STRATEGIES AGAINST EMERGING RNA VIRUSES BY PHENOTYPIC DRUG DISCOVERY* [Doctoral dissertation]. Karolinska Institutet.

Terán Amores, P. I. (2023). *COMPARACIÓN DE HERRAMIENTAS DE IDENTIFICACIÓN DE PLÁSMIDOS* [Tesis de maestría]. Pontificia Universidad Católica del Ecuador.

Van Den Elsen, K., Quek, J. P., & Luo, D. (2021). Molecular Insights into the Flavivirus Replication Complex. *Viruses 2021, Vol. 13, Page 956*, *13*(6), 956. https://doi.org/10.3390/V13060956

Varela Tabares, D. (2019). *DISEÑO E IMPLEMENTACIÓN DE UN FLUJO DE TRABAJO BIOINFORMÁTICO EN LA NUBE PARA LA IDENTIFICACIÓN DE VARIANTES ONCOGÉNICAS A PARTIR DE DATOS GENÓMICOS*. Universidad EIA.

Vig, V. J. (2024). Climate Change and the Emergence and Reemergence of Viruses: A Threat in the Future? *Sophia Lucid*, *3*, 162–179.

Villanueva Romero, C. A. (2023). *CORONAVIRUS Y VARIANTES DE SARS-CoV-2 CIRCULANTES EN MÉXICO* [Tesis de licenciatura]. Universidad Nacional Autónoma de México.

Wahaab, A. ;, Mustafa, B. E. ;, Hameed, M. ;, Stevenson, N. J. ;, Anwar, M. N. ;, Liu, K. ;, Wei, J. ;, Qiu, Y. ;, Wahaab, A., Mustafa, B. E., Hameed, M., Stevenson, N. J., Naveed Anwar, M., Liu, K., Wei, J., Qiu, Y., & Ma, Z. (2021). Potential Role of Flavivirus NS2B-NS3 Proteases in Viral Pathogenesis and Anti-flavivirus Drug Discovery Employing Animal Cells and Models: A Review. *Viruses 2022, Vol. 14, Page 44*, *14*(1), 44. https://doi.org/10.3390/V14010044

Wollner, C. J., & Richner, J. M. (2021). mRNA Vaccines against Flaviviruses. *Vaccines 2021, Vol. 9, Page 148*, *9*(2), 148. https://doi.org/10.3390/VACCINES9020148

Ye, Q., Liu, Z. Y., Han, J. F., Jiang, T., Li, X. F., & Qin, C. F. (2016). Genomic characterization and phylogenetic analysis of Zika virus circulating in the Americas. *Infection, Genetics and Evolution*, *43*, 43–49. https://doi.org/10.1016/J.MEEGID.2016.05.004

Zhao, R., Wang, M., Cao, J., Shen, J., Zhou, X., Wang, D., & Cao, J. (2021). Flavivirus: From Structure to Therapeutics Development. *Life 2021, Vol. 11, Page 615*, *11*(7), 615. https://doi.org/10.3390/LIFE11070615

Zhu, Z., Chan, J. F. W., Tee, K. M., Choi, G. K. Y., Lau, S. K. P., Woo, P. C. Y., Tse, H., & Yuen, K. Y. (2016). Comparative genomic analysis of pre-epidemic and epidemic Zika virus

strains for virological factors potentially associated with the rapidly expanding epidemic. *Emerging Microbes and Infections*, *5*(3), 22. https://doi.org/10.1038/EMI.2016.48;

# 9. AUTOEVALUACIÓN

La realización de este Trabajo de Fin de Grado ha representado una oportunidad invaluable para la integración y aplicación práctica de los conocimientos y habilidades adquiridos durante mi formación dual en Biotecnología e Ingeniería Informática. Considero que uno de los mayores logros personales y académicos de este proyecto ha sido precisamente la capacidad de conjugar estos dos campos que me apasionan para abordar un problema científico complejo, como es el análisis mutacional de genomas virales. La aplicación de herramientas y metodologías propias de la ingeniería informática, como el desarrollo de software, la optimización de algoritmos y la gestión de grandes volúmenes de datos, para resolver una cuestión eminentemente biotecnológica, ha reforzado mi convicción sobre la creciente necesidad y el enorme potencial de la bioinformática en la investigación actual.

A lo largo del desarrollo del trabajo, considero que he logrado aplicar de manera efectiva muchos de los conocimientos adquiridos a lo largo del grado en Biotecnología. Aspectos fundamentales como la estructura y función del ADN y ARN, la organización del genoma, los mecanismos de mutación y los principios de la evolución molecular han sido claves para el análisis y la interpretación de los resultados. Asimismo, he podido estructurar el trabajo siguiendo el formato de un artículo científico, cuidando tanto la claridad como el rigor en la exposición de los datos. También, destaco el esfuerzo realizado en la búsqueda, selección y gestión de la bibliografía científica, lo cual me ha permitido contextualizar adecuadamente el problema abordado y fundamentar sólidamente cada una de las secciones del trabajo. Esta experiencia me ha servido no solo para consolidar conocimientos, sino también para desarrollar competencias esenciales en investigación y comunicación científica.

Para concluir, me gustaría resaltar mi satisfacción con el resultado final, ya que una de mis mayores ilusiones al embarcarme en esta doble titulación era precisamente descubrir y explotar las sinergias entre la biotecnología y la ingeniería informática. Este proyecto ha materializado esa aspiración, permitiéndome no solo aplicar los conocimientos de ambas disciplinas, sino también comprender de manera más profunda cómo se pueden complementar para generar soluciones innovadoras y eficientes en el ámbito de la biotecnología.