Similaridad Genética, Tiempo y Geografía en el Virus del Dengue

Andrea Lizeth Silva Cala

Trabajo de Grado para Optar el título de Bióloga

Director
Daniel Rafael Miranda Esquivel
Dr. Ciencias Naturales

Universidad Industrial de Santander
Facultad de Ciencias
Escuela de Biología
Bucaramanga
2018

Tabla de Contenido

Pág.

Introducción 12

1. Objetivos 15

1.1 Objetivo General 15

1.2 Objetivos Específicos 15

2. Métodos 16

2.1 Construcción de la base de datos 16

2.2 Selección de datos 16

2.3 Cálculo de distancias genéticas sin alineamiento 18

2.4 Distancia Temporal y Geográfica 18

2.5 Relación entre la distancia genética, el tiempo y la geografía 18

2.5.1 Test de Mantel. 19

2.5.2 Análisis de Redundancia Canónica basado en distancias (db-RDA).. 19

2.5.3 Análisis de Procrustes. 20

3. Resultados 20

3.1 Base de datos 20

3.1.1 Serotipos. 21

3.1.2 Gen(es)..... 21

3.1.3 País y Localidad..... 21

3.1.4 Temporalidad. 22

3.2 Datos Seleccionados 27

3.3 Distancias Genéticas sin alineamiento 29

3.4 Relación entre la distancia genética, el tiempo y la geografía..... 29

4. Discusión 33

4.1 Base de datos 33

4.2 Genética y Geografía del virus del Dengue 34

4.3 Estructura Temporal 35

4.4 Gen E Vs Genoma 35

5. Conclusión 36

6. Recomendaciones 36

Referencias Bibliográficas..... 37

Apéndices 46

Lista de Tablas

Tabla 1. Test de Mantel 30

Tabla 2. Análisis de redundancia canónica basado en distancias (db-RDA)..... 31

Tabla 3. Análisis de redundancia canónica basado en distancias (db-RDA) por “términos”.. 32

Tabla 4. Análisis de procrustes. 33

Lista de Figuras

Figura 1. Número de registros presentes en la base de datos por serotipo. 23

Figura 2. Número de registros presentes en la base de datos por países. 24

Figura 3. Distribución global de los serotipos por países. 25

Figura 4. Distribución global de los serotipos por localidades. 26

Figura 5. Número de datos del gen E y del genoma por serotipo. 28

Lista de Apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en la base de datos de la biblioteca UIS)

Apéndice A: Script utilizado para obtener información del virus del Dengue desde el GenBank.

Apéndice B: Script utilizado para depurar la información obtenida del GenBank.

Apéndice C: Script utilizado para obtener las coordenadas geográficas de los países y localidades reportados en la base de datos.

Apéndice D: Script utilizado para descargar las secuencias del genoma y gen E del Dengue.

Apéndice E: Scripts utilizados para identificar clones y recombinantes en secuencias del genoma y gen E usando UCLUST v4.0.

Apéndice F: Script utilizado para calcular las distancias genéticas del genoma y gen E usado el método libre de alineamiento k-mer.

Apéndice G: Script utilizado para calcular el delta absoluto entre los años de reporte del Dengue.

Apéndice H: Scripts utilizados para calcular las distancias geográficas entre países y localidades donde se ha reportado el virus del Dengue.

Apéndice I: Script utilizado para aplicar el test de mantel a las distancias genéticas, geográficas y temporales.

Apéndice J: Script utilizado para el análisis de redundancia canónica basado en distancias (db-RDA).

Apéndice K: Script utilizado para el análisis de Procrustes.

Apéndice L: Información sobre el contenido de cada columna presente en la base de datos.

Apéndice M: Información sobre los datos del gen E y genoma usados en este estudio.

Apéndice N: Estadísticos y gráficas de los datos del gen E y genoma usados en este estudio.

Apéndice O: Script utilizado para calcular la distancia mínima, máxima y media de las matrices de distancia genética.

Apéndice P: Matrices de distancia genética, geográfica y temporal, obtenidas en el estudio.

RESUMEN

TÍTULO: SIMILARIDAD GENÉTICA, TIEMPO Y GEOGRAFÍA EN EL VIRUS DEL DENGUE*.

AUTOR: ANDREA LIZETH SILVA CALA**

PALABRAS CLAVE: DENGUE, K-MER, MÉTODOS LIBRES DE ALINEAMIENTO, DISTANCIA GENÉTICA, DISTANCIA GEOGRÁFICA, DISTANCIA TEMPORAL.

DESCRIPCIÓN:

En los últimos años el número de secuencias virales disponibles ha aumentado, dando paso a nuevos estudios de virus patógenos donde se centran en integrar datos genéticos con datos geográficos y temporales para comprender las dinámicas evolutivas de los virus. En este trabajo evaluamos la relación entre las distancias genéticas del virus del Dengue, calculadas mediante el método libre de alineamiento k-mer, con los patrones geográficos y temporales, a través de tres análisis estadísticos: el test de Mantel, el análisis de redundancia canónica basado en distancias (db-RDA) y el análisis de procrustes. Los tres análisis muestran que la variabilidad genética del genoma del Dengue no es explicada por su geografía ni por su temporalidad, mientras que, los resultados del db-RDA y el test de Mantel muestran correlación significativa ($p = 0.047$; $p = 0.006$ respectivamente) entre el gen de la envoltura (Gen E) y la temporalidad del virus del Dengue, lo que sugiere una estructura temporal para este gen. Los resultados que obtuvimos con el test de Mantel se destacan porque se encuentra correlación entre las distancias genéticas del gen E con las distancias temporales y geográficas, mientras que ninguna correlación es significativa para el genoma, lo que sugiere que el gen E es un buen candidato para profundizar en estudios sobre las dinámicas evolutivas del virus del Dengue.

---

*Trabajo de Grado

**Universidad Industrial de Santander. Facultad de Ciencias. Escuela de Biología. Director: Daniel Miranda. PHD

ABSTRACT

TITLE: GENETIC SIMILARITY, TIME AND GEOGRAPHY IN THE DENGUE VIRUS*.

AUTHOR: ANDREA LIZETH SILVA CALA**.

KEYWORDS: DENGUE, K-MER, FREE ALIGNMENT METHODS, GENETIC DISTANCE, GEOGRAPHICAL DISTANCE, TEMPORAL DISTANCE.

DESCRIPTION:

The number of available viral sequences has been increasing in the last years. This allowed new research focusing on integrated genetic data with geographic and temporal data. In this study, we analyzed the relationship between genetic distances with geographical and temporal patterns of the Dengue virus. For this: 1. We used the k-mer free-alignment method to calculate the genetic dissimilarity between pairs of sequences. 2. We applied three statistical analyzes to test the relationship between the variables: the Mantel test, the Canonical Redundancy Analysis based on distances (db-RDA) and the Analysis of Procrustes. The results indicate that geography and temporality in dengue don't explain the observed genetic variability in the genome. The results of db-RDA, and Mantel show a significant correlation between the gene-E and the temporality of Dengue ($p = 0.047$, $p = 0.006$). This suggests a temporal structure for this gene. The results of the Mantel test show a correlation between the genetic distances of gene-E and the temporal and geographical distances, while none correlation is significant for the total genome. We suggest that the gene-E is a suitable candidate to deepen on studies on the evolutionary dynamics of Dengue virus.

---

*Bachelor Thesis

** Universidad Industrial de Santander. Facultad de Ciencias. Escuela de Biología. Director: Daniel Miranda. PHD

Introducción

El Dengue es un virus patógeno de la familia flaviviridae, su genoma tiene alrededor de 11000 pares de base (pb) y presenta 10 genes: 3 estructurales, los cuales codifican la cápside (C), la membrana (M) y la envoltura (E) y 7 no estructurales, responsables de la replicación (Guzmán et al., 2016). Presenta cuatro serotipos (DENV1, DENV2, DENV3, DENV4) diferenciados antigénicamente que comparten un mismo ancestro y alrededor del 65% de su información genética (Guzmán et al., 2010). Actualmente los 4 serotipos cocirculan en Asia, África y las Américas, siendo endémicos en más de 100 países (Hasan, et al., 2016).

El virus del Dengue presenta altas tasas de mutación (~7,6 x 10^{-4} sustituciones/sitio/año, Costa et al., 2012), lo que conlleva a que su genoma acumule diferencias genéticas en un corto periodo de tiempo (Pybus y Rambaut, 2013).

Actualmente existen dos formas de cuantificar la similaridad entre los genomas virales, así como identificar regiones conservadas, inserciones y deleciones: (1) los métodos basados en alineamiento y (2) los métodos libres de alineamiento (Chauve et al., 2013). Los métodos basados en alineamiento presentan varias desventajas como, asumir colinealidad y no tener en cuenta eventos de recombinación o transferencia horizontal de genes, así como la disminución en la precisión del alineamiento en secuencias con baja identidad, y los altos requerimientos de cómputo que limitan el análisis de secuencias a escala genómica (Zielezinski et al., 2017).

Los métodos libres de alineamiento surgen como alternativa frente a los métodos tradicionales, ya que se pueden aplicar en secuencias con baja identidad, donde el alineamiento no es confiable, y a secuencias donde ocurren eventos de recombinación (Cong et al., 2016; Vinga, 2014). Además, son computacionalmente menos costosos, tanto en tiempo como en cómputo, adecuados para análisis de secuencias genómicas (Zielezinski et al., 2017).

Uno de los métodos libres de alineamiento más común para calcular la similaridad/disimilaridad entre secuencias, es el método de k-mer o k-word (Blaisdell, 1986; Lu et al., 2008; Vinga y Almeida, 2003; Zielezinski et al., 2017). Donde un k-mer es una subsecuencia de longitud k y la frecuencia de un k-mer es igual al número de veces que este aparece a lo largo de una secuencia (Song et al., 2013; Lu et al., 2008). De este modo, dos secuencias con alta identidad tendrán frecuencias de k-mers similares (Reyes-Prieto et al., 2011).

El cálculo de las distancias de k-mer no se basa en modelos evolutivos, y por esta razón no son consideradas distancias genéticas ni distancias evolutivas, así mismo, se considera inapropiado derivar afirmaciones evolutivas a partir de estas distancias (Fan, Ives, Surget-groba, & Cannon, 2015); sin embargo, se ha encontrado correlación entre las distancias genéticas y las distancias de k-mer (Sun et al., 2009). Para efectos de discusión en este trabajo, aunque nuestro enfoque no es evolutivo, nos referiremos a la distancia de k-mer como distancia genética en concordancia con lo propuesto por Murray et al. (2016).

El aumento de la disponibilidad de secuencias del genoma viral en los últimos años, así como la mejora en la calidad de los datos geográficos y temporales, han permitido generar nuevas

hipótesis sobre los orígenes, la dispersión y las dinámicas evolutivas del virus del Dengue (Costa et al., 2012; Pybus, Tatem, y Lemey, 2015; Ratanawong et al., 2016). Sin embargo, pese a los esfuerzos de los métodos filogeográficos para describir y entender las dinámicas temporales y geográficas, así como las dinámicas de dispersión del Dengue, han generado poco conocimiento sobre la estructura espacio-temporal del virus (Salje et al., 2017).

Las bases de datos proporcionan información sobre los reportes del virus del Dengue, y permiten acceder a datos genéticos, geográficas y temporales, así como a información clínica sobre la fuente de aislado, la edad y el género del paciente (Sharma et al., 2015). Esta información es usada para desarrollar diversas investigaciones, sin embargo, las bases de datos disponibles suelen presentar estos datos de manera desorganizada y sin depurar (Benson et al., 2005; Messina et al., 2014; Pickett et al., 2012). Por esto, se hace necesario recopilar estos datos y organizarlos de manera que sean accesibles y de utilidad para futuras investigaciones.

1. Objetivos

1.1 Objetivo General

Estimar la relación entre la distancia genética, la distancia temporal y geográfica para el virus del Dengue.

1.2 Objetivos Específicos

Construir una base de datos del virus del Dengue a partir de la metadata disponible en el GenBank.

Evaluar la relación de la distancia genética entre el genoma del virus del Dengue con la distancia geográfica y la distancia temporal.

Evaluar la relación de la distancia genética entre el gen de la envoltura (Gen E) del virus del Dengue con la distancia geográfica y la distancia temporal.

2. Métodos

2.1 Construcción de la base de datos

Para la construcción de la base de datos del virus del Dengue, accedimos a la información registrada en el GenBank (Benson et al., 2005), bajo el criterio de búsqueda "Dengue Virus Organism" usando el paquete Rentrez en R (Winter, 2016) (Apéndice A). Descargamos información sobre el número de acceso, serotipo, genotipo, gen(es), longitud de la secuencia y lugar del reporte (país – localidad). Organizamos la información en columnas usando funciones base de R v 3.3.3 (R Core Team, 2017) (Apéndice B) y terminamos su depuración en OpenRefine (Ham, 2013). Por último, realizamos la búsqueda de las coordenadas geográficas (Longitud – Latitud) de cada país y localidad a través de la API de Google Maps v 3.30 (Google Inc., 2017), y las añadimos a la base de datos (Apéndice C).

2.2 Selección de datos

De la base de datos construida, seleccionamos las secuencias correspondientes al gen de la envoltura (Gen E), y al marco abierto de lectura (Genoma) teniendo en cuenta las siguientes reglas de decisión:

- No estar designados por el GenBank como:
    - Quimeras
    - Clones

    - No Verificados
- Reportar el serotipo, año de registro y país

Genoma:

- Secuencias de longitud mayor a 10100 pb. Teniendo en cuenta que la longitud del genoma de Dengue reportada en el GenBank por la secuencia de referencia "NC_001474.2" es de 10723 pb.
- Secuencias que reporten el inicio y final del CDS (Secuencia de ADN codificante) en el GenBank.

Gen E:

- Secuencias de longitud mayor a 1300 pb, y menor a 1600 pb. Teniendo en cuenta que la longitud del gen E reportada en el GenBank por la secuencia de referencia "NC_001474.2" es de 1484 pb.

Descargamos las secuencias que cumplieron con estas reglas a partir de su número de acceso. Para las secuencias del Gen E usamos el paquete ape en R (Paradis, Claude, & Strimmer, 2004), y para las secuencias del genoma usamos la función downloadCDSgb (Romero-Alarcon, 2015) en R (Apéndice D). Una vez descargadas, identificamos clones y recombinantes con 97% de similaridad usando el algoritmo UCLUST v4.0.38 (Edgar, 2010), y las eliminamos de los datos (Apéndice E).

2.3 Cálculo de distancias genéticas sin alineamiento

Para el cálculo de las distancias genéticas entre las secuencias del gen E y entre las secuencias del genoma, usamos el método libre de alineamiento k-mer (distancias de k-mer) (Blaisdell, 1986), tomando a k con longitud igual a 3. Primero, calculamos la frecuencia de todas las tripletas posibles a lo largo de cada secuencia, y luego, calculamos las distancias de k-mer tomando las frecuencias como entrada para la ecuación "Fractional common k-mer count" (Edgar, 2004b), que calcula la disimilaridad entre pares de secuencias (Apéndice F).

2.4 Distancia Temporal y Geográfica

Obtuvimos las distancias temporales calculando el delta absoluto entre los años de reporte de las secuencias (Apéndice G), y calculamos las distancias geográficas euclidianas y geodésicas, a partir de coordenadas (longitud - Latitud), entre los países y entre las localidades reportadas usando la función pointDistance del paquete Raster en R (van Etten, 2012) (Apéndice H).

2.5 Relación entre la distancia genética, el tiempo y la geografía

Para evaluar la relación entre la distancia genética, el tiempo y la geografía en el virus del Dengue, realizamos tres análisis estadísticos: El test de Mantel (Mantel, 1967), El análisis de redundancia canónica basado en distancias (db-RDA) (Legendre & Anderson, 1999), y el

análisis de procrustes (Peres-Neto & Jackson, 2001). Para los tres análisis, tomamos un valor p de 0.05 para los test de significancia.

2.5.1 Test de Mantel. Calculamos la correlación entre las distancias genéticas y las distancias geográficas euclidianas, y entre las distancias genéticas y las distancias temporales del gen E y el genoma del Dengue. Para esto, organizamos los valores de las distancias en matrices de disimilaridad, y las tomamos como datos de entrada para el test de Mantel. Calculamos la significancia de cada correlación con un test de permutación sobre la matriz genética con 999 réplicas. Para realizar estas correlaciones usamos la función mantel del paquete vegan en R (Oksanen et al., 2017) (Apéndice I).

2.5.2 Análisis de Redundancia Canónica basado en distancias (db-RDA). Exploramos la relación entre la distancia genética con la geografía y el tiempo usando db-RDA. Para esto, tomamos la matriz de distancia genética como variable de respuesta, y los datos de las coordenadas geográficas (Longitud y Latitud) y los años de reporte como variables explicativas. Para este análisis usamos la función capscale del paquete vegan en R (Oksanen et al., 2017). Calculamos la significancia del análisis con el test de permutación por “términos” con 999 réplicas para obtener el efecto de cada una de las variables explicativas sobre la variable de respuesta. Para realizar este test usamos la función anova del paquete vegan en R (Oksanen et al., 2017). Realizamos este análisis tanto para los datos del gen E como para los de genoma (Apéndice J).

2.5.3 Análisis de Procrustes. Evaluamos la relación entre la distancia genética y geográfica a través del análisis de procrustes. Tomamos la matriz de distancia genética y la matriz de distancia geográfica geodésica como datos de entrada para el análisis de ordenación: Escalamiento multidimensional no métrico (NMDS) (Kruskal, 1964). Las dimensiones genéticas resultantes fueron rotadas, escaladas y transpuestas sobre las dimensiones geográficas resultantes, a través del análisis de procrustes. Calculamos la significancia de cada correlación con un test de permutación con 10000 réplicas usando la función protest del paquete vegan en R (Oksanen et al., 2017) (Apéndice K).

3. Resultados

3.1 Base de datos

La base de datos resultante contiene 18389 registros del virus del Dengue reportados en el GenBank hasta septiembre del 2016. Estos registros se describen en 20 columnas en las cuales se encuentra información sobre el Número de acceso, Serotipo, Genotipo, Gen, Longitud de la secuencia, País, Localidad, coordenadas geográficas (Longitud - Latitud), mes y año (Apéndice L). La base de datos se encuentra disponible en línea en la página web: https://goo.gl/WmFWq2.

3.1.1 Serotipos. En la base de datos se registran los cuatro serotipos del Dengue. El serotipo 1 presenta el mayor número de registros: 6784, seguido del serotipo 2 con 5374, el serotipo 3 con 3471 y el serotipo 4 con 1623 registros (Figura 1). Así mismo hay 428 registros con serotipo no identificado.

3.1.2 Gen(es). Los registros presentes en la base de datos contienen secuencias nucleotídicas de los 10 genes y los extremos UTR, que presenta el genoma de Dengue. Las secuencias del gen E y el genoma son las más reportadas. 8191 registros reportan la secuencia del Gen E y 4020 reportan el Genoma. También se reportan secuencias que contienen de uno a tres genes tanto estructurales como no estructurales. Las secuencias que contienen cuatro o más genes son llamados “Gen de poliproteína”. Así mismo, hay 546 registros con secuencias no identificadas.

3.1.3 País y Localidad. Del total de registros presentes en la base de datos, 16330 reportan el país. Éstos, corresponden a 109 países diferentes, siendo Vietnam el país con más reportes con 2190, seguido de Brasil con 1707, Tailandia con 1532 e India con 1525 (Figura 2). Los cuatro serotipos están presentes en países de Sudamérica, África y Asia, y se encuentran distribuidos en las regiones tropicales y subtropicales del mundo (figura 3). Encontramos que en la base de datos sólo 6084 registros reportan la localidad. En total, están reportadas 428 localidades diferentes, las que presentan mayor número de registros son: Sur de VietNam con 1455, Bangkok - Tailandia con 594, Managua - Brasil con 340, Aragua - Brasil con 220, São Paulo - Brasil con 157, Província de Kamphaeng Phet - Tailandia con 195, y Guangzhou - China con 139 (Figura 4).

La distribución de los serotipos por localidades nos muestra que los reportes de Dengue en países africanos generalmente se registran sin localidad (Figura 4). No todos los casos presentes en la base de datos reportan el lugar, más de 2000 registros no presentan el país ni localidad.

3.1.4 Temporalidad. La base de datos presenta 14104 registros con año reportado. El registro más antiguo es del año 1944, y el más reciente es del año 2016. El año 2013 presenta mayor número de reportes con 1556, seguido del año 2007 con 1417 y el año 2010 con 1309. Desde la década de los 80 y 90, el número de reportes del virus del Dengue por año ha aumentado. Sin embargo, en la base de datos hay 4285 registros que no reportan el año, y sólo 2922 registros reportan el mes. Los registros con serotipo desconocido no pertenecen a un año específico y más de la mitad no reporta el año.

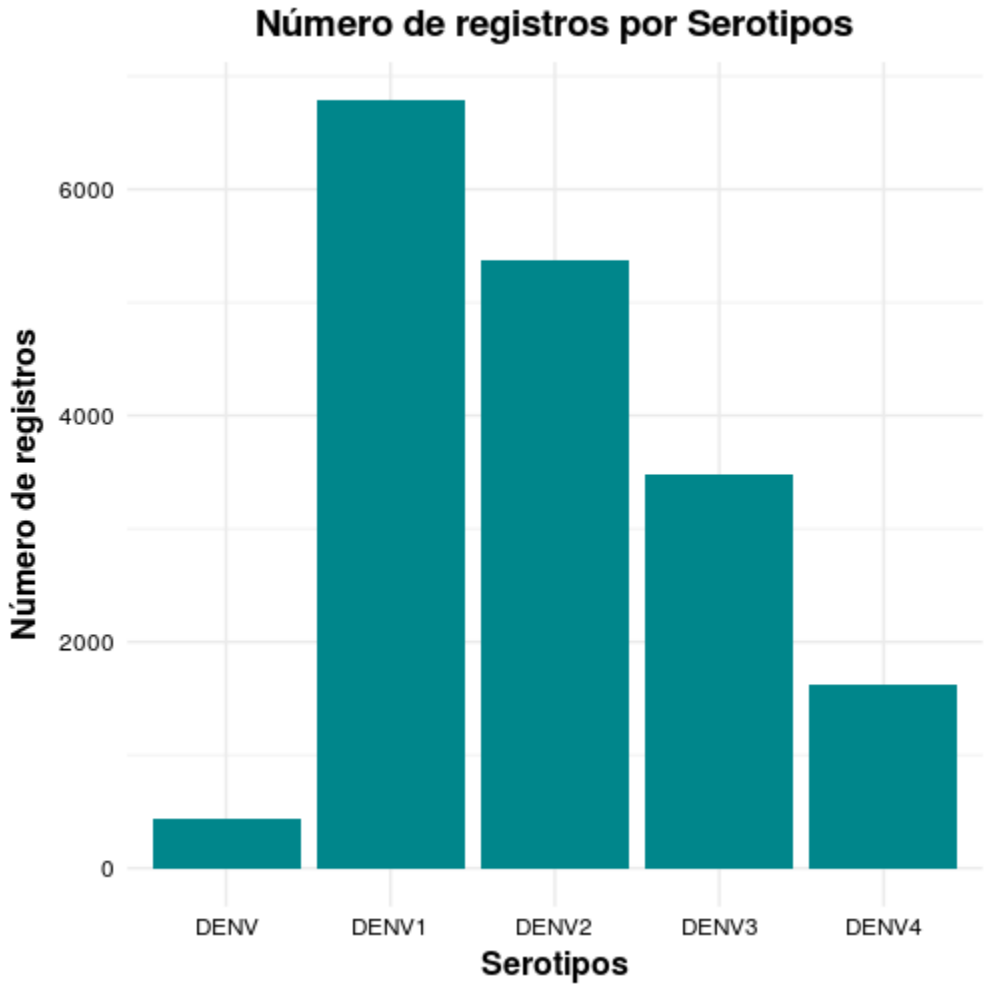

Figura 1. Número de registros presentes en la base de datos por serotipo. La base de datos contiene 18.389 registros de Dengue reportados en el GenBank desde el año 1944 a 2016. La base de datos reporta los cuatro serotipos: serotipo 1 (DENV1) con 6784 registros, el serotipo 2 (DENV2) con 5374 registros, el serotipo 3 (DENV3) con 3471 registros, el serotipo 4 (DENV4) con 1623 registros y 428 registros con serotipo desconocido (DENV).

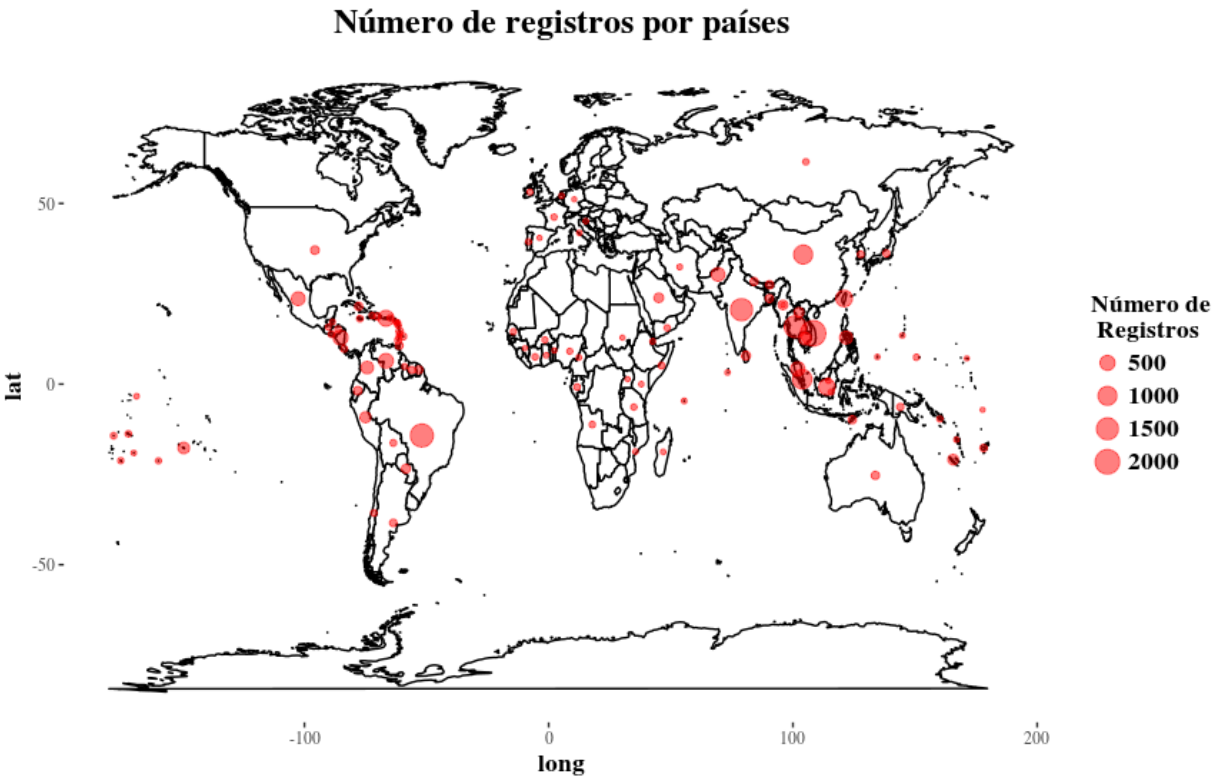

Figura 2. Número de registros presentes en la base de datos por países. El área del círculo es proporcional a la cantidad de registros. Los países con mayor número de reportes son VietNam (2190 registros), Brasil (1707 registros), Tailandia (1532 registros), India (1525 registros), y China (1025 registros).

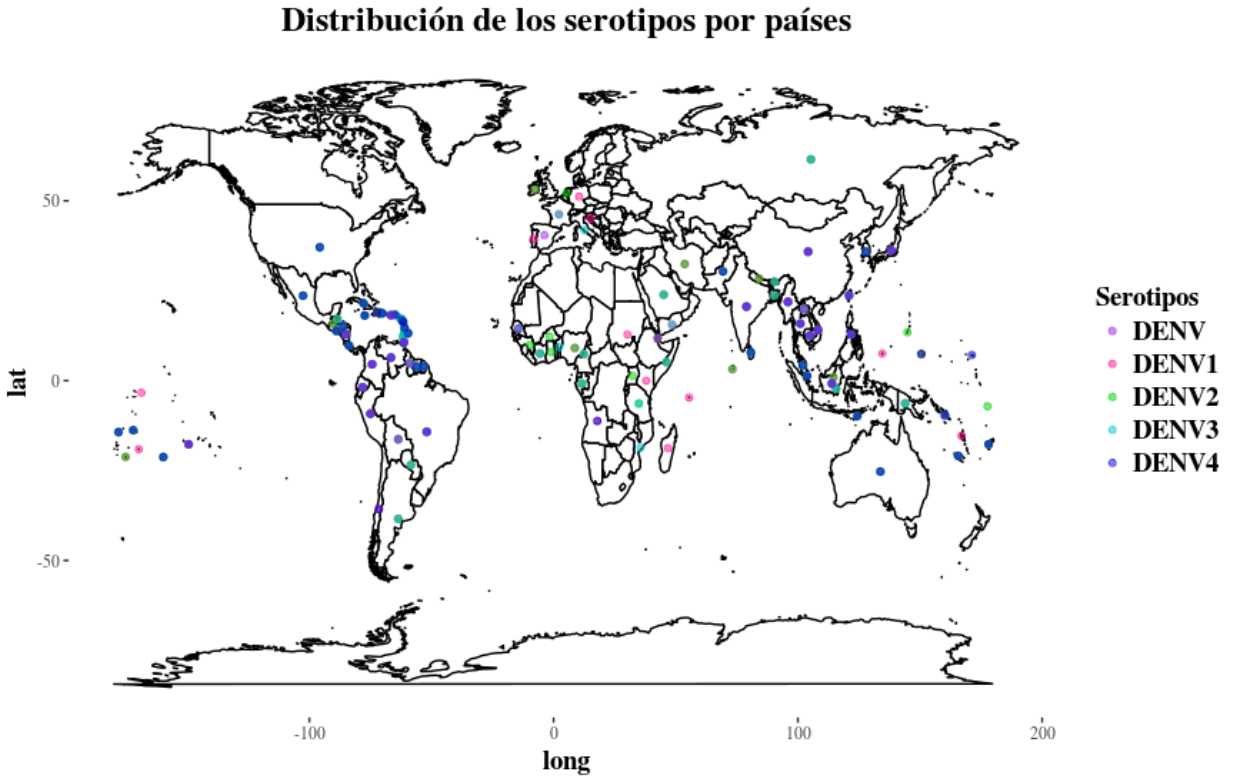

Figura 3. Distribución global de los serotipos por países. En la base de datos se reporta el virus del Dengue en 109 países distribuidos en todos los continentes a excepción de Europa. Los cuatro serotipos (DENV1-4) se encuentran distribuidos por toda la región tropical y subtropical del mundo. Los registros no serotipificados (DENV) no pertenecen a un país en particular.

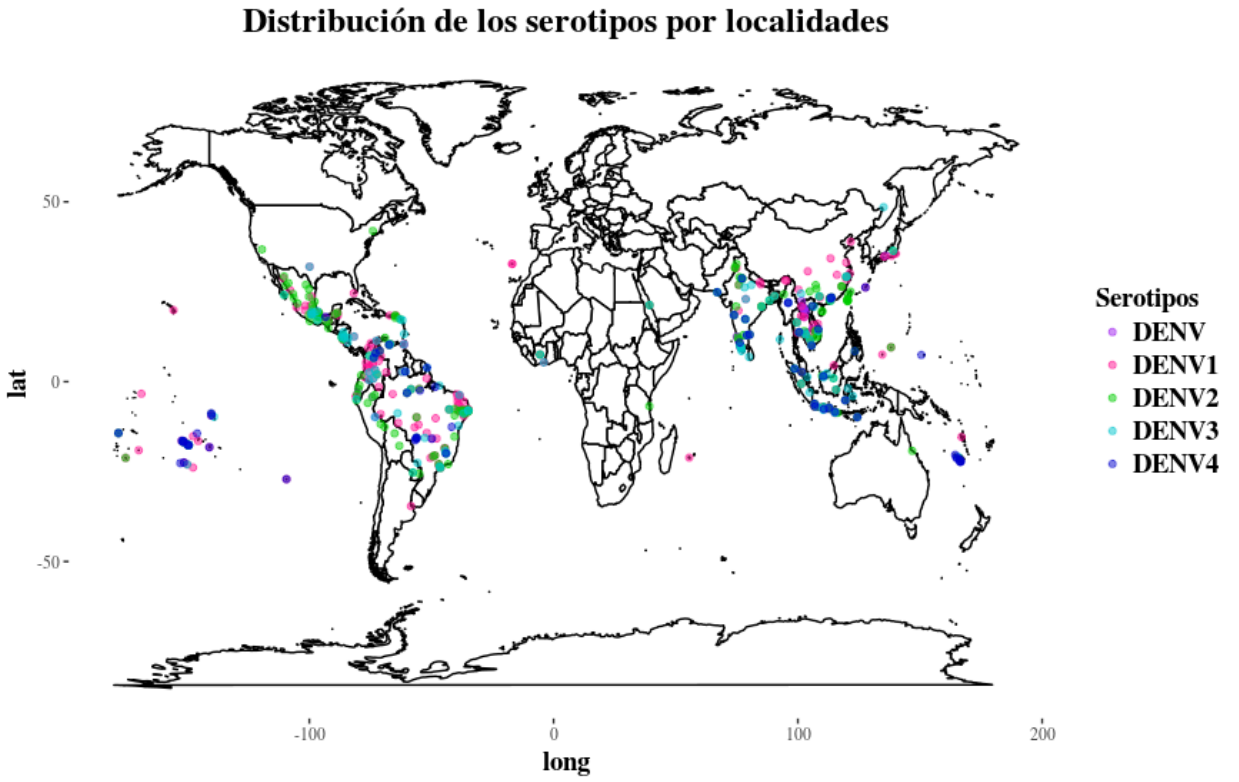

Figura 4. Distribución global de los serotipos por localidades. Los cuatro serotipos (DENV1-4) se reportan en países y localidades de centro y sudamérica así como en Asia y África. Sin embargo, en África muchos de los registros no reportan localidad. Las localidades con mayor número de reportes son Sur de VietNam (1455), Bangkok - Tailandia (594), Managua - Brasil (340), Aragua - Brasil (220), Sao Paulo - Brasil (157), Província de Kamphaeng Phet - Tailandia (195), y Guangzhou - China (139).

3.2 Datos Seleccionados

Obtuvimos 130 secuencias del Gen E y 83 secuencias del Genoma que cumplieron con las reglas anteriormente expuestas y que son 97% disimilares (Apéndice M). Los datos del Gen E contienen secuencias de los cuatro serotipos (Figura 5), reportan 42 países y 19 localidades de Asia, África y sudamérica**,** y presentan una escala temporal desde el año 1945 a 2014. La longitud mínima de las secuencias es 1310 pb, la longitud máxima es 1590 pb y la longitud media es 1486 pb. Los datos del Genoma contienen secuencias de los cuatro serotipos (Figura 5), los cuales están reportados en 28 países y 25 localidades de Asia, África y sudamérica, y así mismo presentan una serie temporal desde el año 1944 a 2016. La longitud mínima de las secuencias es de 10175 pb, la longitud máxima es 10736 pb y la longitud media es 10645 (Apéndice N).

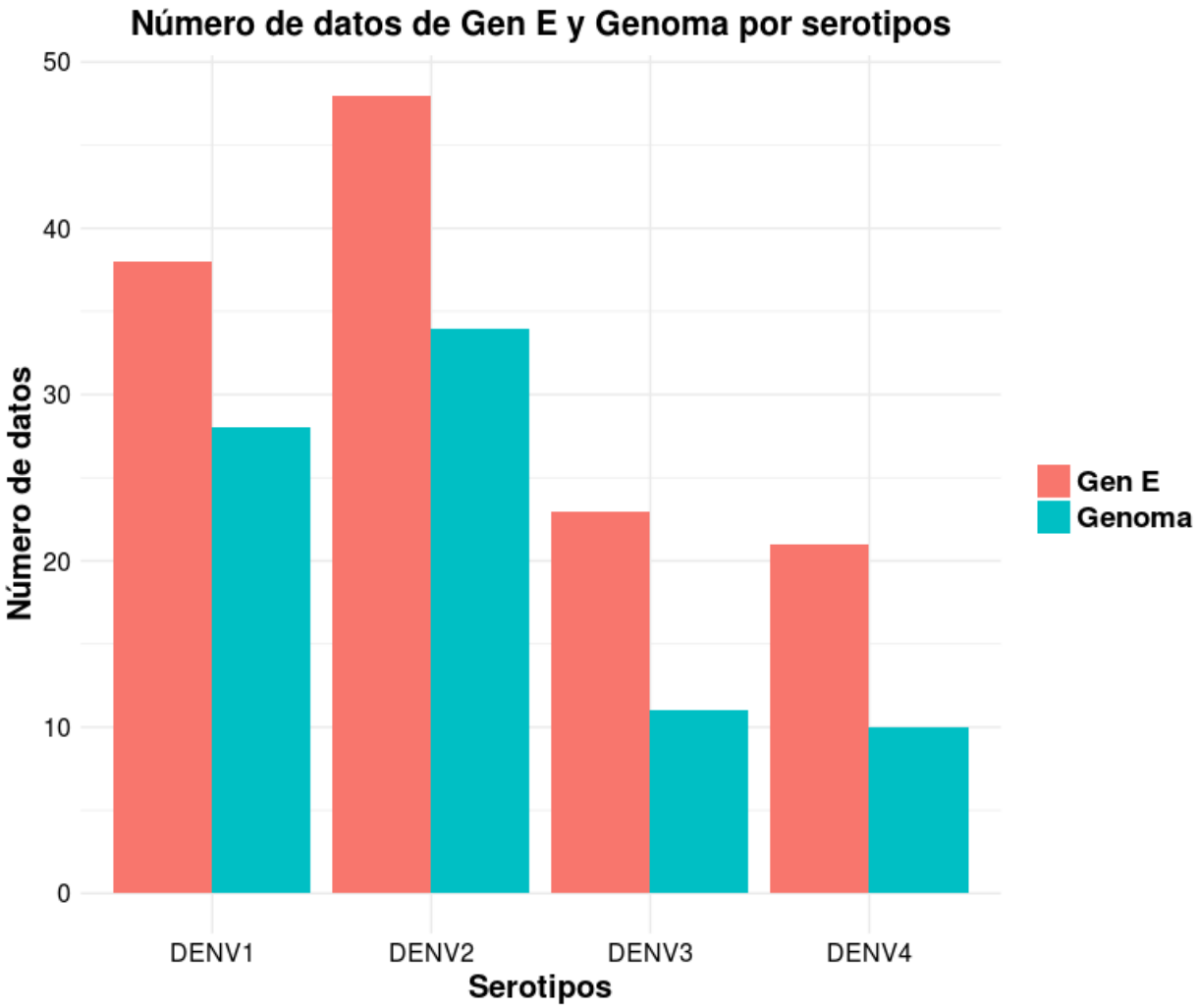

Figura 5. Número de datos del gen E y del genoma por serotipo. Obtuvimos 130 datos del Gen E, reportados en 42 países en una escala temporal de 1945 a 2014; y 83 datos del Genoma, reportados en 28 países en una escala temporal de 1944 a 2016. Los datos del gen E y del genoma presentan los cuatro serotipos: serotipo 1 (DENV1) con 38 y 28 casos, el serotipo 2 (DENV2) con 48 y 34 casos, el serotipo 3 (DENV3) con 23 y 11 casos, el serotipo 4 (DENV4) con 21 y 10 casos respectivamente.

3.3 Distancias Genéticas sin alineamiento

Obtuvimos dos matrices de disimilaridad, una matriz corresponde a las distancias entre pares de secuencias del gen E y la segunda corresponde a las distancias entre pares de secuencias del genoma. De cada una de estas matrices, obtuvimos 4 sub-matrices, resultado de la partición por presencia de Pais y localidad (Apéndice O y P):

1. Matriz del gen E con localidad: La distancia máxima entre las secuencias es 0.1429, la distancia mínima es 0.0227 y la distancia media es 0.0891.
2. Matriz del gen E con país: La distancia máxima entre las secuencias es 0.1469, la distancia mínima es 0 y la distancia media es 0.0861.
3. Matriz del genoma con localidad: La distancia máxima entre las secuencias es 0.0528, la distancia mínima es 0.007 y la distancia media es 0.0329.
4. Matriz del genoma con localidad: La distancia máxima entre las secuencias es 0.0552, la distancia mínima es 0.006 y la distancia media es 0.0345.

3.4 Relación entre la distancia genética, el tiempo y la geografía

El análisis con el test de Mantel nos muestra que las distancias genéticas entre el gen E están correlacionadas de manera positiva con la distancia geográfica tanto a nivel de localidad como a nivel de país, con valor p de 0.015 y 0.011 respectivamente (Tabla 1), y también, nos muestra que están correlacionadas de manera positiva con la distancia temporal con valor p de 0.006

(Tabla 1). Por el contrario, para las distancias genéticas entre el genoma de Dengue obtuvimos que no están correlacionadas ni con la distancia geográfica ni temporal (Tabla 1).

Tabla 1.

Test de Mantel.

Variables	Correlaciones	Estadístico de Mantel r	Significancia
Gen E / Geografía / Tiempo	Dist. Gen E Vs Dist. Geográfica de Localidades	0.1659	0.015 *
	Dist. Gen E Vs Dist. Geográfica de Países	0.07654	0.011 *
	Dist. Gen E Vs Dist. Temporal de Años	0.08715	0.006 *
Genoma / Geografía / Tiempo	Dist. Genoma Vs Dist. Geográfica de Localidades	0.03646	0.275
	Dist. Genoma Vs Dist. Geográfica de Países	0.04584	0.114
	Dist. Genoma Vs Dist. Temporal de Años	0.03644	0.165

Test de Mantel entre las distancias genéticas del Dengue y su geografía y temporalidad. Existe correlación positiva entre las distancias genéticas del gen E con la distancia geográfica y temporal. *Significancia p = 0.005.

Cuando evaluamos la geografía y la temporalidad como variables explicativas en el db-RDA, no encontramos ninguna asociación fuerte entre las distancias genéticas del gen E y entre las distancias genéticas del genoma con la geografía (coordenadas geográficas) y el tiempo (años). La geografía a nivel de localidad y la temporalidad sólo representan el 9% y el 5.7% de la

varianza en la matriz del gen E y del genoma, respectivamente (Tabla 2). La geografía a nivel de país y la temporalidad sólo representan el 2.5% y el 3.6% de la varianza en la matriz del gen E y del genoma, respectivamente (Tabla 2). Sin embargo, cuando evaluamos las variables por separado, encontramos una correlación positiva entre la distancia genética del gen E y los años de reporte con un valor p de 0.047 (Tabla 3).

Tabla 2.

Análisis de redundancia canónica basado en distancias (db-RDA).

Análisis	**Anova RDA (p-Value)**	**Porcentaje de Varianza Explicada por las Variables (%)**
Gen E - Geografía (Localidades) - Tiempo (Años)	0.736	9
Gen E - Geografía (Países) - Tiempo (Años)	0.076	2.5
Genoma - Geografía (Localidades) - Tiempo (Años)	0.819	5.7
Genoma - Geografía (Países) - Tiempo (Años)	0.492	3.6

Análisis de redundancia canónica basado en distancias (db-RDA). db-RDA entre el conjunto de las variables explicativas: coordenadas geográficas y años, y las distancias genéticas del Dengue. No encontramos ninguna correlación positiva, y el porcentaje de la varianza en las distancias genéticas que es explicada por las variables es muy bajo (<10%).

Tabla 3.

Análisis de redundancia canónica basado en distancias (db-RDA) por "términos".

Variable de respuesta	Variable Explicativa	p-Value
Dist. Gen E	Long Localidad	0.36
	Lat Localidad	0.516
	Años	0.047 *****
	Long País	0.139
	Lat País	0.398
Dist. Genoma	Long Localidad	0.61
	Lat Localidad	0.838
	Años	0.446
	Long País	0.178
	Lat País	0.876

Anova sobre cada una de las variables explicativas. Obtuvimos que la variable Años se correlaciona de manera positiva con las distancias del gen E. *****Significancia p = 0.005.

En el análisis de Procrustes no encontramos ninguna alineación óptima entre las dimensiones resultantes de las distancias genéticas con las dimensiones resultantes de las distancias geográficas a nivel de país ni a nivel de localidad. Esto se demuestra en los resultados que obtuvimos con el test de permutación, donde ninguna comparación es significativa (Tabla 4).

Tabla 4.

Análisis de procrustes.

Variables	Suma de cuadrados (m12) procrustes	Correlación en una rotación procrustes simétrica	Significancia
Gen E Vs Localidades	0.9735	0.1628	0.64314
Gen E Vs Países	0.9983	0.04088	0.90491
Genoma Vs Localidades	0.9943	0.07519	0.77322
Genoma Vs Países	0.9814	0.1365	0.24118

Análisis de procrustes entre las distancias del Dengue y su geografía. No se observa correlación positiva entre las distancias genéticas del gen E y el genoma con su geografía, es decir que no encontramos ninguna alineación óptima entre las dimensiones de los datos genéticos con las dimensiones de los datos geográficos.

4. Discusión

4.1 Base de datos

La base de datos que presentamos aquí es una compilación actualizada y completa sobre los registros del virus del Dengue presentes en el GenBank. Nosotros mostramos una visión general de la información encontrada, centrándonos en los datos temporales y geográficos. Messina et al. (2014) presentan en su estudio una base de datos del virus del Dengue con su información geográfica, sin embargo, los reportes compilados van solo hasta el año 2013, y la base de datos carece de información sobre el serotipo, el genotipo y gen o genes que presenta cada registro.

Sharma et al. (2015) realizaron una revisión sobre las diferentes herramientas computacionales y bases de datos de virus disponibles, donde resaltan la importancia que éstas tienen para el desarrollo de investigaciones en virología.

4.2 Genética y Geografía del virus del Dengue

Los resultados de la relación entre la distancia genética con la geográfica del Dengue difieren según el análisis estadístico. Mientras los resultados del test de Mantel sugieren una estructura geográfica para el gen E a nivel de localidad y a nivel de país, los resultados del db-RDA y el análisis de procrustes sugieren lo contrario. Este resultado no se esperaba, porque se ha evidenciado que el test de Mantel presenta bajo poder estadístico (Guillot y Rousset, 2013), y el análisis de redundancia canónica (RDA) se ha propuesto como alternativa (Legendre y Fortin, 2010). Estudios con otros virus de la familia flaviviridae como el virus de la influenza y el virus de la encefalitis de San Luis (SLEV), han mostrado una correlación entre las distancias genéticas del genoma y las distancias geográficas, sugiriendo una estructura espacial (Carrel et al., 2010; Diaz et al., 2015); sin embargo, estos estudios se han realizado a nivel local. El estudio de Salje et al. (2017) demuestra que la escala espacial influye en la estructura geográfica del virus del Dengue, los casos que se encuentran a menos de 200 metros presentan una estructura dada la geografía, pero esta se pierde cuando los casos se encuentran separados por más de un 1 km de distancia. Esto podría explicar porque nuestros resultados no sugieren una estructura geográfica a nivel mundial.

4.3 Estructura Temporal

Los resultados obtenidos en el db-RDA por "términos" y el test de Mantel, sugieren una estructura temporal para el gen E. Esto nos dice que la variabilidad genética observada en el gen E es explicada a través de la dinámica temporal del virus del Dengue (Cappa et al., 2013). La presencia de estructura temporal es importante, ya que es un control requerido para realizar estimaciones de tasas evolutivas (Holmes, 2016) y para realizar inferencias filogenéticas bajo un modelo de reloj molecular (Murray et al., 2015). Los estudios de Wei y Li (2017), y Tian et al. (2017), apoyan nuestro resultado, ya que encontraron estructura temporal al evaluar las distancias genéticas del gen E con los años de reporte del virus del Dengue, a través del software TempEst (Rambaut et al., 2016). Estos estudios también encuentran una estructura espacial para el gen E usando métodos filogeográficos, sin embargo, los análisis también se llevaron a cabo a escala geográfica local, lo que sugiere, que a diferencia de la estructura geográfica, la estructura temporal del gen E del virus del Dengue se da independiente de la escala geográfica del análisis.

4.4 Gen E Vs Genoma

Cuando evaluamos las distancias genéticas entre el genoma del Dengue, no encontramos ninguna correlación significativa con el tiempo ni la geografía. Sin embargo, el genoma ha sido usado en estudios donde evidencian una estrecha relación entre la genética y la geografía de otros virus de la familia flaviviridae como SLEV y el virus del Nilo occidental (WNV) (Diaz et al., 2015; Di Giallonardo et al., 2016). Los resultados que obtuvimos con el test de Mantel se destacan porque se encuentra correlación entre las distancias genéticas del gen E con las

distancias temporales y geográficas, mientras que ninguna correlación es significativa para el genoma. Esto sugiere que el gen E puede ser un candidato para profundizar en estudios sobre las dinámicas evolutivas del virus del Dengue. El gen E ha sido el gen de referencia para estudios sobre la epidemiología, diversidad y evolución del virus del Dengue (Wei y Li, 2017), así como un blanco en el desarrollo de una vacuna (Liu et al., 2016).

5. Conclusión

Los resultados obtenidos evidencian la presencia de estructura temporal en el gen E. Sin embargo, no esclarecen la relación entre la variabilidad genética y la geografía del Dengue. Nuestro estudio sugiere que el gen E es adecuado para profundizar en estudios sobre la dinámica y estructura espacio-temporal del Dengue.

6. Recomendaciones

- Dado que nuestros resultados muestran estructura temporal para el gen E pero no para el genoma, recomendamos analizar otros genes del virus del Dengue con el fin de evaluar cuáles y en qué medida evidencian estructura geográfica y temporal en el Dengue.

- Nuestro análisis se enfocó en evaluar la estructura geográfica a nivel global, por esto recomendamos explorar la estructura geográfica del Dengue a nivel local y así evaluar si la escala geográfica influye en los resultados.

Referencias Bibliográficas

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, *33*(DATABASE ISS.), 34–38. https://doi.org/10.1093/nar/gki063

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(14), 5155–5159. https://doi.org/10.1073/pnas.83.14.5155

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., & Marcucci Poltri, S. N. (2013). Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: A case study in Eucalyptus globulus. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0081267

Carrel, M. A., Emch, M., Jobe, R. T., Moody, A., & Wan, X. F. (2010). Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam. *PLoS ONE*, *5*(1). https://doi.org/10.1371/journal.pone.0008631

Chauve, C., El-mabrouk, N., & Tannier, E. (2013). *Models and Algorithms for Genome Evolution* (Vol. 19). https://doi.org/10.1007/978-1-4471-5298-9

Cong, Y., Chan, Y. B., & Ragan, M. A. (2016). A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Scientific Reports*, *6*, 1–13. https://doi.org/10.1038/srep30308

Costa, R. L., Voloch, C. M., & Schrago, C. G. (2012). Comparative evolutionary epidemiology of dengue virus serotypes. *Infection, Genetics and Evolution*, *12*(2), 309–314. https://doi.org/10.1016/j.meegid.2011.12.011

Di Giallonardo, F., Geoghegan, J. L., Docherty, D. E., McLean, R. G., Zody, M. C., Qu, J., … Holmes, C. (2016). Fluid Spatial Dynamics of West Nile Virus in the United States : Rapid. *Journal of Virology*, *90*(2), 862–872. https://doi.org/10.1128/JVI.02305-15.Editor

Diaz, L. A., Goñi, S. E., Iserte, J. A., Quaglia, A. I., Singh, A., Logue, C. H., … Contigiani, M. S. (2015). Exploring genomic, geographic and virulence interactions among epidemic and non-epidemic St. Louis encephalitis virus (flavivirus) strains. *PLoS ONE*, *10*(8), 1–16. https://doi.org/10.1371/journal.pone.0136316

Ebi, K. L., & Nealon, J. (2016). Dengue in a changing climate. *Environmental Research*, *151*, 115–123. https://doi.org/10.1016/j.envres.2016.07.026

Edgar, R. C. (2004b). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113. https://doi.org/10.1186/1471-2105-5-113

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Fan, H., Ives, A. R., Surget-Groba, Y., & Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, *16:522*. https://doi.org/10.3732/ajb.1100335

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*(4), 336–344. https://doi.org/10.1111/2041-210x.12018

Guzmán, M. G., Gubler, D. J., Izquierdo, A., Martinez, E., & Halstead, S. B. (2016). Dengue infection. *Nature Reviews Disease Primers*, *2*, 1–26. https://doi.org/10.1038/nrdp.2016.55

Guzman, M. G., Halstead, S. B., Artsob, H., Buchy, P., Farrar, J., Gubler, D. J., … Peeling, R. W. (2010). Dengue: a continuing global threat. *Nature Reviews Microbiology*, *8*(12), S7–S16. https://doi.org/10.1038/nrmicro2460

Ham, K. (2013). Electronic resources reviews. *Electronic Resources Reviews*, 233–234.

Hasan, S., Jamdar, S., Alalowi, M., & Al Ageel Al Beaiji, S. (2016). Dengue virus: A global human threat: Review of literature. *Journal of International Society of Preventive and Community Dentistry*, *6*(1), 1. https://doi.org/10.4103/2231-0762.175416

Holmes, E. C. (2016). Complexities of Estimating Evolutionary Rates in Viruses. *Journal of Virology*, *90*(4), 2155–2155. https://doi.org/10.1128/JVI.02570-15

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Legendre, P., & Anderson, M. J. (1999). Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Experiments. *Ecological Monographs*, *69*(1), 1–24. https://doi.org/10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2

Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, *10*(5), 831–844. https://doi.org/10.1111/j.1755-0998.2010.02866.x

Liu, Y., Liu, J., & Cheng, G. (2016). Vaccines and immunization strategies for dengue prevention. *Emerging Microbes & Infections*. Nature Publishing Group. https://doi.org/10.1038/emi.2016.74

Lu, G., Zhang, S., & Fang, X. (2008). An improved string composition method for sequence comparison. *BMC Bioinformatics*, *9 Suppl 6*(Ccv), S15. https://doi.org/10.1186/1471-2105-9-S6-S15

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*(February), 1729–1736.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., … Hay, S. I. (2014). Global spread of dengue virus types: Mapping the 70 year history. *Trends in Microbiology*, *22*(3), 138–146. https://doi.org/10.1016/j.tim.2013.12.011

Murray, G. G. R., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., … Welch, J. J. (2015). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods in Ecology and Evolution*, (2010), 1–10. https://doi.org/10.1111/2041-210X.12466

Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., & Warthmann, N. (2016). kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Computational Biology*, *13*(9), 1–15. https://doi.org/10.1371/journal.pcbi.1005727

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2017). vegan: Community Ecology Package. Retrieved from https://cran.r-project.org/package=vegan

Paradis, E., Claude, J., & Strimmer, K. (2004). A{PE}: analyses of phylogenetics and evolution in {R} language. *Bioinformatics*, *20*, 289–290.

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the Mantel test. *Oecologia*, *129*(2), 169–178. https://doi.org/10.1007/s004420100720

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., … Scheuermann, R. H. (2012). ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, *40*(D1), 593–598. https://doi.org/10.1093/nar/gkr859

Pybus, O. G., Fraser, C., & Rambaut, A. (2013). Evolutionary analysis of the dynamics of viral infectious disease. *Phil. Trans. R. Soc. B.*, *368*(AUgUsT), 20120193. https://doi.org/10.1038/nrg2583

Pybus, O. G., Tatem, A. J., & Lemey, P. (2015). Virus evolution and transmission in an ever more connected world. *Proceedings. Biological Sciences / The Royal Society*, *282*(1821), 20142878. https://doi.org/10.1098/rspb.2014.2878

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007. https://doi.org/10.1093/ve/vew007

Ratanawong, P., Kittayapong, P., Olanratmanee, P., Wilder-Smith, A., Byass, P., Tozan, Y., … Louis, V. R. (2016). Spatial variations in dengue transmission in schools in Thailand. *PLoS ONE*, *11*(9), 1–16. https://doi.org/10.1371/journal.pone.0161895

Reyes-Prieto, F., García-Chéquer, A. J., Jaimes-Díaz, H., Casique-Almazán, J., Espinosa-Lara, J. M., Palma-Orozco, R., … Beattie, K. L. (2011). Lifeprint: A novel k-tuple distance method for construction of phylogenetic trees. *Advances and Applications in Bioinformatics and Chemistry*, *4*(1), 13–27. https://doi.org/10.2147/AABC.S15021

Romero-Alarcon, V. (2015). download.CDS.GenBank. Retrieved from https://github.com/alarconvv/download.CDS.GenBank

Salje, H., Lessler, J., Berry, I. M., Melendrez, M. C., Endy, T., Kalayanarooj, S., … Cummings, D. A. T. (2017). Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science*, *355*(6331), 1302–1306. https://doi.org/10.1126/science.aaj9384

Sharma, D., Priyadarshini, P., & Vrati, S. (2015). Unraveling the Web of Viroinformatics: Computational Tools and Databases in Virus Research. Journal of Virological, 89(3), 1489–1501. https://doi.org/10.1128/JVI.02027-14

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., & Sun, F. (2014). New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, *15*(3), 343–353. https://doi.org/10.1093/bib/bbt067

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., & Farmerie, W. (2009). ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, *37*(March 2017). https://doi.org/10.1093/nar/gkp285

Tian, H., Sun, Z., Faria, N. R., Yang, J., Cazelles, B., Huang, S., … Xu, B. (2017). Increasing airline travel may facilitate co-circulation of multiple dengue virus serotypes in Asia. *PLoS Neglected Tropical Diseases*, *11*(8), 1–15. https://doi.org/10.1371/journal.pntd.0005694

van Etten, R. J. H. & J. (2012). raster: Geographic analysis and modeling with raster data. Retrieved from http://cran.r-project.org/package=raster

Vinga, S. (2014). Editorial: Alignment-free methods in computational biology. *Briefings in Bioinformatics*, *15*(3), 341–342. https://doi.org/10.1093/bib/bbu005

Vinga, S., & Almeida, J. (2003). Alignment-free sequence comparison - A review. *Bioinformatics*, *19*(4), 513–523. https://doi.org/10.1093/bioinformatics/btg005

Wei, K., & Li, Y. (2017). Global evolutionary history and spatio-temporal dynamics of dengue virus type 2. *Scientific Reports*, *7*(February), 1–14. https://doi.org/10.1038/srep45505

Winter, D. J. (2016). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *XX*, 0–2. https://doi.org/10.7287/peerj.preprints.3179v2

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, *18*(1), 1–17. https://doi.org/10.1186/s13059-017-1319-7

Apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en base de datos de la Biblioteca UIS)