

Influencia del muestreo de terminales y particiones sobre la reconstrucción filogenética del virus
del dengue usando análisis de inferencia bayesiana

Luis Manuel Calsada Rodríguez

Trabajo de Grado para optar al título de Biólogo

Director

Daniel Rafael Miranda Esquivel

Doctor en Ciencias Naturales

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Biología

Bucaramanga

2019

Dedicatoria

A mi madre, padre y hermanos.

A Manuel Rodríguez, Isabel Cardona, Ana Isbelia Saavedra y María de Jesus Rodríguez.

A mis otros hermanos, Felipe y Fabián Quintero.

A la memoria de Jaime Rodríguez, quién despertó mi interés por la ciencia y definió el rumbo de mi vida.

Agradecimientos

A mi madre por toda su paciencia y sus consejos, por ser el faro que guía mi camino. A mi padre, por su paciencia, por sus abrazos y por siempre estar para mí incluso en la distancia. A mis hermanos por su compañía y por todo lo que me han enseñado a lo largo de mi vida. A toda mi familia por creer en mí. A Xiomara Figueroa, Nancy Osorio y Eusebio Castro por permitirme entrar en sus casas, sentarme en su mesa y sobre todo por darme el mejor regalo de mi vida, mis hermanos.

Al profesor Daniel Miranda por su invaluable aporte a mi formación como biólogo y a mis compañeros de laboratorio por el café, las risas y por ser una parte importante de mi aprendizaje.

A todas las personas que en el transcurso de estos 25 años han ayudado a forjar mi forma de pensar y sentir. A mi tierra adoptiva Aguachica, Cesar y sus maravillosos habitantes, por acogermme y regalarme una identidad, especialmente a la señora Liliana cuyas charlas y consejos llevaré siempre conmigo.

A mis amigos Angélica Rodríguez Rey, Elver José Herrán Salas, Juan José Fernández Castro, Jeferson Aceros Calderón y Withman Jinete Polo, por permitirme hacer parte de sus vidas, por enseñarme que en las diferencias se encuentran esas largas y maravillosas conversaciones que le dan sentido a la vida.

Tabla de Contenido

Introducción	12
1. Objetivos	15
2. Materiales y métodos	16
2.1. Muestreo, particiones y alineamiento	16
2.2. Selección de modelos de sustitución nucleotídica	18
2.3. Selección de prior de reloj y de árbol	18
2.3.1. Test de reloj estricto	18
2.3.2. Ajuste de modelo de generación árbol y reloj relajado	18
2.4. Reconstrucción filogenética	19
2.5. Comparación de filogenias	20
2.5.1. Recuperación de serotipos y genotipos	20
2.5.2. Distancia de las topologías respecto a evidencia total	20
3. Resultados	21
3.1. Muestreo de secuencias	21
3.2. Selección de modelos de sustitución nucleotídica	22
3.3. Selección de prior de reloj y de árbol	23

3.3.1. Test de reloj estricto	23
3.3.2. Modelo de reloj	24
3.3.3. Modelo de generación de árbol	24
3.4. Comparación de topologías	25
3.4.1. Recuperación genotipos y serotipos	25
3.4.2. Nodos comunes respecto a evidencia total	27
4. Discusión	30
4.1. Modelos de sustitución nucleotídica	30
4.2. Reloj y dinámica poblacional en el virus del Dengue	31
4.3. Recuperación de nodos	32
4.3.1. Recuperación de serotipos y genotipos	32
4.3.2. Similitud topológica	33
5. Conclusiones	33
6. Recomendaciones	34
Referencias Bibliográficas	34
Apéndice A	47

Lista de Figuras

Figura 1.	Recuperación de genotipos y serotipos en los dos tamaños de muestreo	25
Figura 2.	Recuperación de genotipos por cada serotipo	26
Figura 3.	Eficiencia de las particiones en la recuperación de los genotipos en los tamaños de muestreo	27
Figura 4.	Porcentaje de nodos comunes respecto a la filogenia de evidencia total según la longitud de la secuencia utilizada	28
Figura 5.	Porcentaje de nodos en común de las topologías obtenidas de los tipos de genes	29

Lista de Tablas

Tabla 1.	Regiones del genoma utilizadas para la reconstrucción de las filogenias	17
Tabla 2.	Tabla de número de secuencias por serotipos y sus respectivos genotipos	21
Tabla 3.	Valores del alfa de las distribuciones Gamma de cada modelo	23
Tabla 4.	Priors seleccionados por cada partición	24
Tabla 5.	Modelos con mayor peso AIC por réplica para cada partición	47
Tabla 6.	Matriz de transición y frecuencias de bases en cada partición	48
Tabla 7.	Frecuencias de bases y número de tipos de cambio de los modelos por partición	49
Tabla 8.	Test de ajuste a reloj estricto por gen para las tres réplicas	50

Resumen

Título: Influencia del muestreo de terminales y particiones sobre la reconstrucción filogenética del virus del dengue usando análisis de inferencia bayesiana *

Autor: Luis Manuel Calsada Rodríguez **

Palabras Clave: Virus del Dengue, Análisis bayesiano, Muestreo, Terminales, Genoma.

Descripción: El estudio de la filogenia del virus del Dengue (DENV) ha permitido comprender su historia evolutiva, su dinámica poblacional, además de ser útil en la clasificación taxonómica del mismo. En la estimación de estas filogenias se han utilizado distintas particiones y un número variable de terminales, sin embargo no es claro el criterio de selección de los datos o el efecto que podrían tener estas variables al realizar la reconstrucción filogenética. Por tanto, el propósito de este trabajo fue evaluar el efecto que tienen el tamaño del muestreo taxonómico y la partición utilizada en la estimación filogenética de DENV, para esto utilizamos una base de datos de 328 secuencias de genoma completo la cual submuestreamos en tamaños del 20 % y el 70 % de los terminales, a partir de estos muestreos construimos filogenias con diferentes regiones del genoma (10 genes, 2 combinaciones y la región de la Proteína 2K) que comparamos con la Evidencia Total (ET). Encontramos que la recuperación de genotipos es directamente proporcional al tamaño de la partición, aunque esta recuperación también se ve afectada por el tamaño de muestreo ya que las filogenias obtenidas a partir de un menor número de terminales presentan porcentajes más altos de recuperación. La similitud topológica respecto a ET fue más alta al usarse un menor número de terminales y una secuencia de mayor longitud. En conclusión encontramos que las particiones de mayor longitud son más eficientes en la identificación taxonómica de DENV; sin embargo, éstas se ven afectadas por el número de terminales usados en la estimación de la filogenia.

* Trabajo de grado

** Facultad de Ciencias. Escuela de Biología. Director: Daniel Rafael Miranda Esquivel, Doctor en Ciencias Naturales.

Abstract

Title: Effect of partition and taxon sampling on dengue virus phylogeny: an analysis using Bayesian inference *

Author: Luis Manuel Calsada Rodríguez **

Keywords: Dengue Virus, Bayesian inference, Sampling, Taxon, Genome.

Description: The study of the phylogeny of dengue virus (DENV) has allowed us its taxonomic classification, and the understanding of its evolutionary and dynamic history. In the phylogenetic estimates, different genomic regions and a variable number of taxa have been included, however, the criteria for selecting the data or the effect of these variables on phylogenetic reconstruction is not clear. Therefore, the purpose of this work was to evaluate the effect of taxonomic sampling size and genomic region on the phylogenetic estimation of DENV. We use a database of 328 complete genome sequences that we sample in sets of 20 and 70 %, and the phylogeny was reconstructed using different genomic partitions (10 genes, 2 combinations of genes and the 2K protein region) that we compared with a phylogeny obtained from a Total Evidence analysis (ET). The recovery of the genotype in the partitions is directly proportional to the size of the genomic region, although this is also affected by the number of taxa since the phylogeny obtained from a smaller number has higher recovery rates. The topological similarity with the total evidence was greater when a smaller number of taxa and a longer sequence size were used. In conclusion, longer genomic regions are more efficient in the taxonomic identification of DENV; however, this is affected by the number of taxa included in the phylogeny estimate.

* Bachelor Thesis

** Facultad de Ciencias. Escuela de Biología. Director: Daniel Rafael Miranda Esquivel, Doctor en Ciencias Naturales.

Introducción

El virus del Dengue (DENV) es un arbovirus de la familia *Flaviviridae* perteneciente al género *Flavivirus* (Knipe y Howley, 2007; Mukhopadhyay et al., 2005), es uno de los agentes infecciosos con mayor reemergencia a nivel mundial y se estima que causa 50 millones de infecciones al año (WHO, 2000), por lo cual se ha convertido en uno de los principales intereses de la medicina actual (Gubler, 2002). Su genoma es de tipo ARN monocatenario positivo cuya región codificante es de aproximadamente 11kb (Mukhopadhyay et al., 2005; Chambers et al., 1990), esta región comprende 10 genes de los cuales 3 codifican proteínas estructurales: proteína de la capsida, la envoltura (Gen E) y de la membrana (PrM); los 7 genes restantes codifican proteínas no estructurales (NS1, NS2A, NS2B, NS3, NS4A, NS4B y NS5) (Guzman et al., 2016; Perera y Kuhn, 2008).

DENV está subdividido en 4 grandes grupos antigénicamente diferenciados llamados serotipos (DENV1-4) que han sido considerados como grupos monofiléticos (Holmes y Twiddy, 2003; Vasilakis y Weaver, 2008) y a su vez estos serotipos están subdivididos en grupos estrechamente relacionados llamados genotipos (Weaver y Vasilakis, 2009; Holmes y Twiddy, 2003; Rico-Hesse, 1990). Las relaciones entre y dentro de estos grupos han sido ampliamente estudiadas debido a su importancia médica (Mackenzie et al., 2004; Leng et al., 2009; Barban et al., 2012), su dinámica poblacional y evolutiva (Zanotto et al., 1996; Wittke et al., 2002; Dos Santos Carmo et al., 2017), e incluso por su distribución geográfica (Kyle y Harris, 2008; Cherian y Mourya, 2016).

En el estudio de las relaciones evolutivas dentro de DENV se han utilizado diferentes enfoques de reconstrucción ya sea parsimonia (Leitmeyer et al., 1999), máxima verosimilitud (Holmes et al., 1999), distancia (Araújo et al., 2009) e inferencia bayesiana (Jiménez-Silva et al., 2018). El uso de análisis bayesiano para la inferencia filogenética se ha incrementado en los últimos años debido al desarrollo e implementación de sofisticados análisis evolutivos, además del desarrollo de plataformas que facilitan la interacción del usuario con los programas (Nascimento et al., 2017). La posibilidad de estimar distintos parámetros simultáneamente con la filogenia es una de las ventajas del análisis bayesiano sobre los demás métodos (Holder y Lewis, 2003), estos parámetros pueden ser estimaciones relacionadas a reloj molecular (Drummond et al., 2006), tamaño poblacional (Drummond et al., 2005) y el modelo evolutivo (Pagel y Meade, 2006).

En DENV el análisis de inferencia bayesiana ha sido usado para estimar patrones filogeográficos (Nunes et al., 2012), reconstrucción de la historia evolutiva (Villabona-Arenas y de Andrade Zanotto, 2011), así como cálculos de tiempo de divergencia y tasas evolutivas (Kotaki et al., 2016). En estas estimaciones se han utilizado diferentes regiones del genoma como: genoma completo (Hapuarachchi et al., 2016; Nunes et al., 2012), marco abierto de lectura (Schreiber et al., 2009), gen E (De Souza et al., 2011; Jiménez-Silva et al., 2018; Dash et al., 2011), fragmento de gen E (Vicente et al., 2018), unión gen E/NS1 (Franco et al., 2015), adicionalmente estos estudios también presentan variación en el número de terminales incluidos que van desde unos pocos (8) (Tolou et al., 2001) hasta más de mil (Ernst et al., 2015). Dada esta variación en las metodologías utilizadas, no queda claro cuál es el criterio de selección y es esperable que existan un efecto en

la estimación de las relaciones. Sin embargo, aún no se han realizado estudios donde se evidencie cómo la matriz de datos (longitud de secuencia y número de terminales) puede afectar la resolución de la filogenia de DENV, por lo tanto el objetivo del presente trabajo fue evaluar el efecto que tiene el número de terminales y la longitud de la partición sobre la reconstrucción filogenética de DENV utilizando análisis bayesiano.

1. Objetivos

Objetivo general

Cuantificar el efecto del muestreo de terminales y uso de particiones en la reconstrucción filogenética por inferencia bayesiana del virus del dengue (DENV).

Objetivos específicos

Seleccionar los priores adecuados para cada partición.

Generar filogenias de DENV a partir de cada partición en los tamaños de muestreo.

Calcular porcentajes de similitud entre las filogenias generadas tomando en cuenta sus nodos en común.

Competencias

Selección de los priores adecuados para realizar el análisis de inferencia bayesiana.

Reconstrucción de filogenias usando análisis de inferencia bayesiana en el programa BEAST.

Análisis de los resultados obtenidos a partir de las comparaciones de las filogenias.

Redacción de un informe científico derivado del trabajo realizado.

2. Materiales y métodos

2.1. Muestreo, particiones y alineamiento

La base de datos de referencia fue generada previamente en un proyecto de grado (Silva Cala y Miranda Esquivel, 2018) y contiene información relevante como: el número de acceso en GenBank (Benson et al., 2012), la región del genoma, el genotipo, el serotipo y la longitud de la secuencia. Para este trabajo tomamos de la base de datos las secuencias de genoma completo cuyo serotipo y genotipo estuvieran determinados.

El muestreo para la selección de modelos de sustitución nucleotídica y el test de ajuste a reloj estricto lo hicimos tomando el 50 % de los terminales disponibles en tres réplicas con una semilla fija para cada una (10, 20, 30). En la exploración de priores solo utilizamos la primera de las réplicas debido al esfuerzo computacional requerido por este método. Las filogenias para las comparaciones fueron generadas a partir de muestreos del 20 % y 70 % de los terminales en tres réplicas con una semilla diferente en cada una (10, 20, 30). Todos los muestreos los hicimos de tipo estratificado sobre los números de acceso de las secuencias teniendo en cuenta las proporciones iniciales de los genotipos. Adicionalmente, si el genotipo contaba con una o dos secuencias siempre lo incluimos dentro del muestreo.

Para el alineamiento descargamos las secuencias del GenBank, alineamos los cuatro serotipos de manera independiente, particionamos los diez genes y la región P2K teniendo en cuenta

las secuencias de referencia de cada serotipo disponibles en la plataforma RefSeq (Briser et al., 2014). Posteriormente alineamos las particiones de nuevo, esta vez con los serotipos en conjunto. Hicimos todos los alineamientos usando el programa MUSCLE v3.8.3 (Edgar, 2004) con los parámetros por defecto y los visualizamos en el programa AliView v1.25 (Larsson, 2014).

Las particiones usadas para la estimación de las filogenias corresponden a los genes estructurales (gen de la Envoltura, Capside y Premembrana), genes no estructurales (NS1, NS2A, NS2B, NS3, NS4A, NS4B y NS5), las combinaciones de E+NS1 y genes estructurales en conjunto, adicionalmente tuvimos en cuenta la región de la proteína 2K ubicada entre los genes NS4A y NS4B (ver Tabla 1).

Tabla 1

Regiones del genoma utilizadas para la reconstrucción de las filogenias

Región del genoma	Nomenclatura
Gen de la Capside	Capside
Gen Pre-membrana	PrM
Gen de la Envoltura	GenE
Gen no estructural 1	NS1
Gen no estructural 2A	NS2A
Gen no estructural 2B	NS2B
Gen no estructural 3	NS3
Gen no estructural 4A	NS4A
Región proteína 2K	P2K
Gen no estructural 4B	NS4B
Gen no estructural 5	NS5
Combinación genes estructurales	EST
Combinación gen de la Envoltura y el gen no estructural 1	E-NS1

2.2. Selección de modelos de sustitución nucleotídica

Evalúamos los modelos de sustitución nucleotídica para cada uno de los genes individualmente y la región P2K en las tres réplicas del muestreo usando el programa JModelTest v2.1 (Darriba et al., 2012), como criterio de selección del modelo utilizamos el contenido de información de Akaike (Akaike, 1998; Posada y Buckley, 2004).

2.3. Selección de prior de reloj y de árbol

2.3.1. Test de reloj estricto. En la primera parte de la exploración de reloj hicimos una prueba de ajuste de likelihood (Neyman y Pearson, 1933) al modelo de reloj estricto (Zuckerkandl y Pauling, 1965) para cada uno de los genes y la región P2K individualmente en cada una de las réplicas. En esta prueba empleamos los árboles obtenidos de cada uno de los genes y la región P2K a partir de una estimación de máxima verosimilitud usando el programa RAxML-NG v0.9.0 (Kozlov et al., 2019) y calculamos el ajuste de likelihood del modelo en el programa PAUP v4.0a165 (Swofford, 2003).

2.3.2. Ajuste de modelo de generación árbol y reloj relajado. Si el gen no ajustaba al modelo de reloj estricto, evaluamos los modelos de reloj relajado no correlacionado log-normal y no correlacionado exponencial (Drummond et al., 2006). En la exploración del modelo de generación de árbol tuvimos en cuenta tres: coalescente de población constante (Kingman, 1982), coalescente de crecimiento exponencial (Drummond et al., 2002) y el modelo coalescente no para-

métrico skyline (Drummond et al., 2005). Para todos los modelos anteriores cuantificamos el ajuste mediante la estimación del likelihood marginal bajo un muestreo de stepping-stone en el programa BEAST v2.5.2 (Bouckaert et al., 2019) y calculamos factores bayesianos para seleccionar el modelo de mejor ajuste (Kass y Raftery, 1995).

2.4. Reconstrucción filogenética

A las matrices de caracteres generadas de los muestreos agregamos como grupos externos terminales del Virus Usutu (NC006551), Encefalitis Japonesa (NC001437) y el Virus del Nilo Occidental (NC009942). Las tasas de sustitución, las frecuencias de bases y el valor Alfa de la distribución Gamma los tomamos del resultado de la selección de modelos, dejando como valor de inicio el promedio entre las tres réplicas; los modelos de árbol y de reloj utilizados fueron los de mejor ajuste en la exploración de priores. Como árbol de inicio utilizamos la topología obtenida de un análisis de máxima verosimilitud. Adicionalmente, generamos una topología de evidencia total en el cual incluimos todos los genes con sus respectivos modelos y todas las secuencias disponibles (328), además de los outgroups anteriormente mencionados.

Realizamos la estimación filogenética utilizando el software BEAST v2.5.2, para esto creamos los archivos ejecutables usando la herramienta Bayesian Evolutionary Analysis Utility (BEAUti) que viene incluida en BEAST. Comprobamos la convergencia del análisis a una distribución estacionaria revisando los valores de tamaño efectivo de muestreo (Nascimento et al., 2017) y la gráfica de probabilidades marginales utilizando el programa Tracer v1.5 (Rambaut et al., 2018).

Generamos filogenias para las tres réplicas de los dos tamaños de muestreo de terminales (20% y 70%) a partir de los 10 genes individualmente, la región P2K, la combinación de genes estructurales y la combinación del gen de la envoltura y el gen NS1. En total obtuvimos 79 filogenias incluyendo evidencia total.

2.5. Comparación de filogenias

Todas la comparaciones entre las topologías fueron hechas a partir de los árboles de máxima credibilidad de las estimaciones, estos fueron calculados usando la herramienta TreeAnnotator v2.6.0 incluida en BEAST.

2.5.1. Recuperación de serotipos y genotipos. Para evaluar la eficacia de las particiones en la recuperación de grupos monofiléticos, cuantificamos el porcentaje promedio de genotipos recuperados por las topologías obtenidas a partir de los distintos genes o combinaciones en las tres réplicas de los dos tamaños de muestreo.

2.5.2. Distancia de las topologías respecto a evidencia total. Cuantificamos la similitud entre las topologías por medio de conteo de nodos comunes entre los árboles de máxima credibilidad de cada uno de los submuestreos (incluidas las tres réplicas de cada uno) y la topología de evidencia total, para dicha comparación eliminamos los terminales de la topología de evidencia total que no se encontraran en la obtenida a partir del submuestreo.

3. Resultados

3.1. Muestreo de secuencias

La matriz de terminales disponibles tuvo un total de 328 números de acceso de los cuales; 104 (31.90 %) correspondieron a DENV1 pertenecientes a 4 genotipos; 129 (39.27 %) a DENV2 correspondientes a 6 genotipos (El genotipo V solo tuvo un representante); DENV3 con 60 (18.40 %) representantes de 3 genotipos; finalmente, DENV4 tuvo un total de 35 (10.43 %) números de acceso correspondientes a 3 genotipos (el genotipo III con solo un representante)(ver Tabla 2).

Tabla 2

Tabla de número de secuencias por serotipos y sus respectivos genotipos

Serotipo	Genotipo	Secuencias disponibles	Total
DENV1	I	5 (4,8 %)	104
	III	62 (59,6 %)	
	IV	8 (7,7 %)	
	V	29 (27,9 %)	
DENV2	IV	4 (3,1 %)	129
	Cosmopolitan	81 (62,8 %)	
	Asian II	10 (7,7 %)	
	American	21 (16,3 %)	
	V	1 (0,8 %)	
	American-Asian	12 (9,3 %)	
DENV3	I	16 (26,7 %)	60
	II	3 (5,0 %)	
	III	41 (68,3 %)	
DENV4	I	2 (5,7 %)	35
	III	1 (2,9 %)	
	II	32 (91,4 %)	
Total secuencias			328

3.2. Selección de modelos de sustitución nucleotídica

Los genes NS5, gen E, NS2A, Pre-Membrana, NS2B y la región P2K ajustaron al mismo modelo en las tres réplicas, mientras que los genes de la Capside, NS3, NS1, NS4B y NS4A repitieron modelo en dos de las tres réplicas. Sin embargo, los valores de las frecuencias de base y de la matriz de transición fueron similares entre las réplicas aun cuando el modelo de mejor ajuste no haya sido el mismo (ver Anexo A); las frecuencias de base en dichos modelos fueron desiguales en todos los genes exceptuando la región P2K, las frecuencias G-C se mantuvieron cercanas a 0.25 mientras que la frecuencia de adenina en todas las particiones fue superior a 0.3. Las tasas de sustitución de la matriz de cambio son diferenciales dependiendo de la partición, aunque en todos los casos se observa que las transiciones tienen una tasa más alta (ver Apéndice A).

La complejidad del modelo no estuvo determinada por la longitud de la secuencia, aunque la partición de menor número de caracteres tiene las frecuencias iguales y solo tres tipos de cambio, en las otras particiones el número de parámetros del modelo es independiente de la longitud (ver Anexo A). Aunque en algunas ocasiones el modelo seleccionado para el gen no fue el mismo entre las replicas se evidencia que los principales modelos de mejor ajuste se mantienen aún cuando las secuencias hayan cambiado debido al muestreo (ver Apéndice A).

Todos los modelos de los genes presentaron heterogeneidad Gamma entre los sitios, el Alfa de esta distribución estuvo entre 0.2 y 0.5, entre estos el mayor valor fue el del gen NS2A con un

promedio de 0.455 mientras que el gen NS4B tuvo el menor Alfa promedio el cual fue 0.2195, con estos valores las formas de las distribuciones son semejantes, por lo tanto la heterogeneidad en los genes es similar y no está determinada por la longitud de estos. Por otra parte, entre las réplicas del mismo gen no hubo diferencias considerables entre los valores obtenidos, siendo la diferencia absoluta mayor 0.066 correspondiente al gen NS2A (ver Tabla 3).

Tabla 3

Valores del alfa de las distribuciones Gamma de cada modelo

Partición	Alfa mínimo	Alfa máximo	Media	Diferencia
P2K	0,371	0,397	0,384	0,026
Capside	0,301	0,336	0,3185	0,035
NS4A	0,325	0,33	0,3275	0,005
NS2B	0,315	0,319	0,317	0,004
PrM	0,298	0,334	0,316	0,036
NS2A	0,422	0,488	0,455	0,066
NS4B	0,211	0,228	0,2195	0,017
NS1	0,254	0,274	0,264	0,02
GenE	0,337	0,356	0,3465	0,019
NS3	0,251	0,256	0,2535	0,005
NS5	0,284	0,293	0,2885	0,009

Nota: La columna Diferencia, representa la diferencia absoluta entre las réplicas.

3.3. Selección de prior de reloj y de árbol

3.3.1. Test de reloj estricto. Ninguno de los genes ajustó al modelo de reloj estricto (ver Apéndice A), el valor de p fue lo suficientemente bajo como para no dejar dudas sobre la incompatibilidad del modelo con los datos (Wasserstein et al., 2016).

3.3.2. Modelo de reloj. En este caso la región P2K, los genes NS2A, NS2B, NS3 y NS4B presentaron un modelo de reloj relajado no correlacionado exponencial, mientras que el gen de la Capside, de la Envoltura, Pre-membrana, NS5, NS4A, y NS1 ajustaron a un reloj log-normal (ver Tabla 4). La complejidad del modelo del prior de reloj relajado no correlacionado no estuvo determinado por la longitud de la secuencia.

Tabla 4

Priors seleccionados por cada partición

Partición	Reloj estricto	Modelo reloj relajado	Modelo árbol
P2K	No	Exponencial	Coalescente exponencial
Capside	No	Log-normal	Coalescente exponencial
NS4A	No	Log-normal	Bayesian Skyline
NS2B	No	Exponencial	Coalescente constante
PrM	No	Log-normal	Bayesian Skyline
NS2A	No	Exponencial	Coalescente exponencial
NS4B	No	Exponencial	Bayesian Skyline
NS1	No	Log-normal	Bayesian Skyline
Gen E	No	Log-normal	Coalescente constante
NS3	No	Exponencial	Bayesian Skyline
NS5	No	Log-normal	Bayesian Skyline

3.3.3. Modelo de generación de árbol. El gen de la envoltura y el NS2B se ajustaron al modelo de población constante, mientras que la mayoría de genes presentaron modelos con población variable a través del tiempo. El gen de la Capside, NS2A y la región P2K ajustaron al modelo coalescente de crecimiento exponencial; NS4A, el gen de la premembrana, NS4B, NS1, NS3 y NS5 ajustaron al modelo skyline (ver Tabla 4).

3.4. Comparación de topologías

3.4.1. Recuperación de genotipos y serotipos. La recuperación de genotipos se incrementó al usar secuencias con una mayor longitud, con el tamaño de submuestreo del 20% encontramos un porcentaje de recuperación de genotipos más alto en comparación con los de mayor número de terminales; las particiones más largas lograron una recuperación promedio de hasta del 70% mientras que en el submuestreo del 70% y la filogenia de evidencia total solo se recuperaron el 50% y el 40% de los genotipos respectivamente. Por otra parte, los serotipos en las filogenias obtenidas del tamaño de muestreo del 20% siempre se recuperaron, mientras que utilizando las dos particiones de menor longitud con un tamaño de muestreo del 70% estos grupos no son monofiléticos en algunas de las réplicas (ver Figura 1).

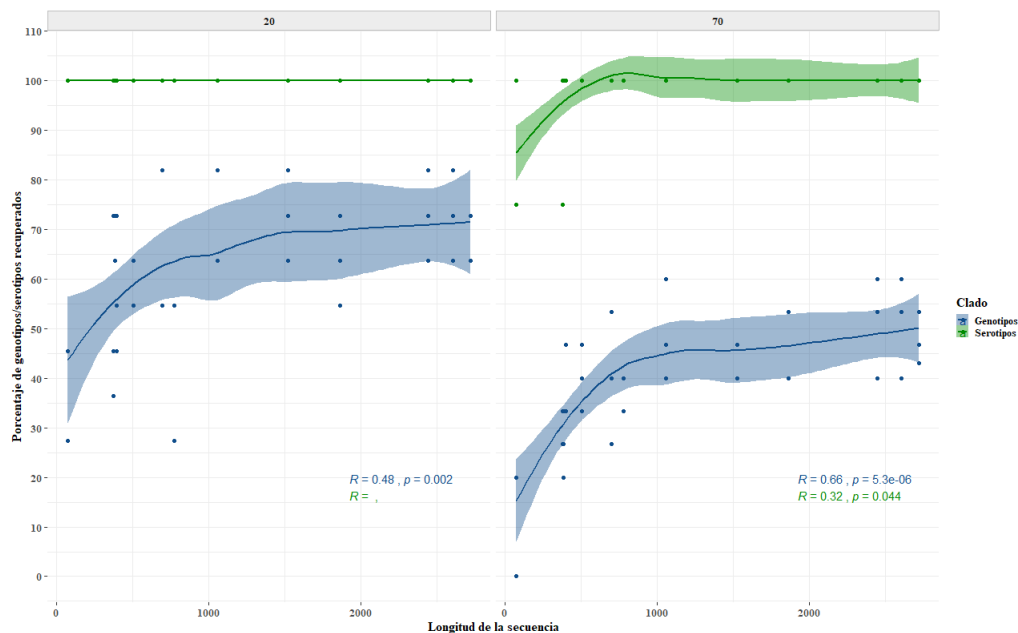


Figura 1. Recuperación de genotipos y serotipos en los dos tamaños de muestreo. Porcentajes de recuperación de genotipos(azul) y serotipos (verde) en los dos tamaños de muestreo utilizados.

En los serotipos DENV2 y DENV3 el porcentaje de recuperación de sus genotipos fue superior respecto a los otros, llegándose a recuperar más del 80 % de estos grupos en el muestreo más pequeño y hasta el 70% en el muestreo de mayor tamaño con las particiones más eficientes. Los genotipos dentro de DENV4 se recuperaron menos en relación a los dos anteriores, aunque con un mayor porcentaje respecto a DENV1 que fue el más bajo en el muestreo del 20 % y sus genotipos no se recuperaron en ninguna de las filogenias obtenidas del submuestreo del 70 % (ver Figura 2).

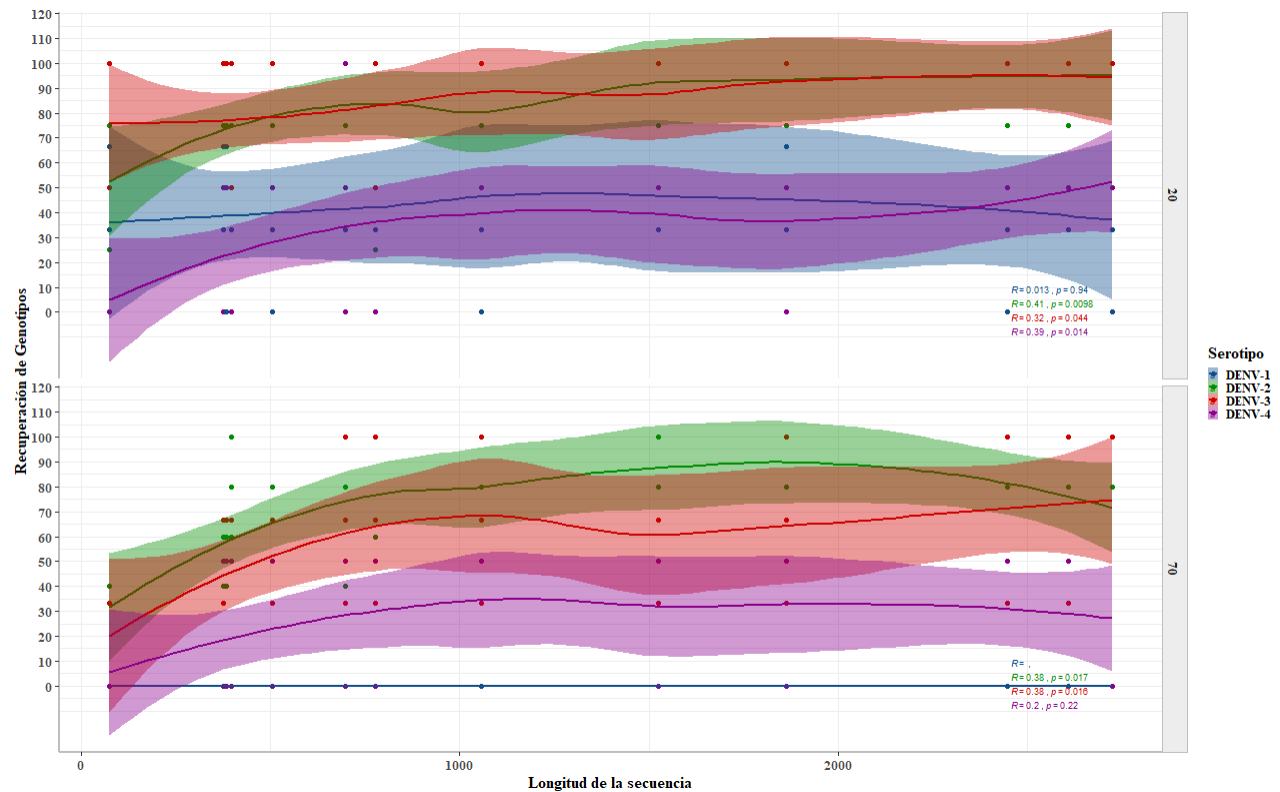


Figura 2. [Recuperación de los genotipos dentro de cada serotipo. Se evidencia un mayor porcentaje de recuperación en los clados de los serotipos 2 y 3]

En cuanto a la región del genoma utilizada, las de mayor tamaño presentaron una media

de recuperación más alta que las de menor longitud, sin embargo el gen NS2A tuvo un porcentaje mayor de recuperación respecto a las demás en el muestreo del 20 % aunque con una desviación mayor. Las combinaciones de genes (estructurales y E-NS1) fueron las más eficientes ya que recuperaron un porcentaje promedio superior al 50 % un porcentaje mayor al de NS5 que posee mayor longitud (ver Figura 3).

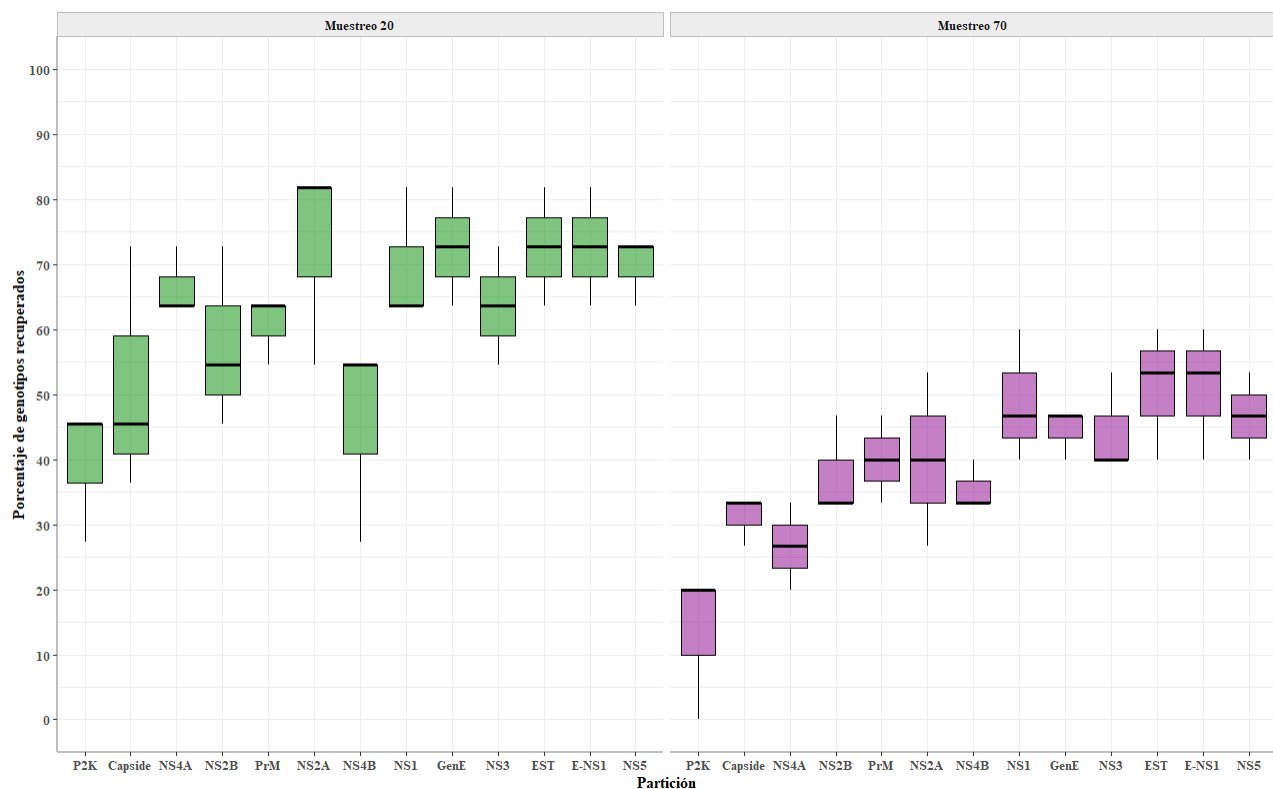


Figura 3. Eficiencia de las particiones en la recuperación de los genotipos en los tamaños de muestreo.

3.4.2. Nodos comunes respecto a evidencia total. La similaridad entre las topologías de los submuestreos y evidencia total estuvo determinada en este caso por la longitud; encontramos

que las topologías generadas con regiones del genoma de mayor longitud reportaban más similitud con la de evidencia total. En cuanto al número de terminales, las estimaciones a partir del muestreo del 20 % tienen más nodos en común con la topología de evidencia total en comparación con las del muestreo de mayor número de terminales (ver Figura 4).

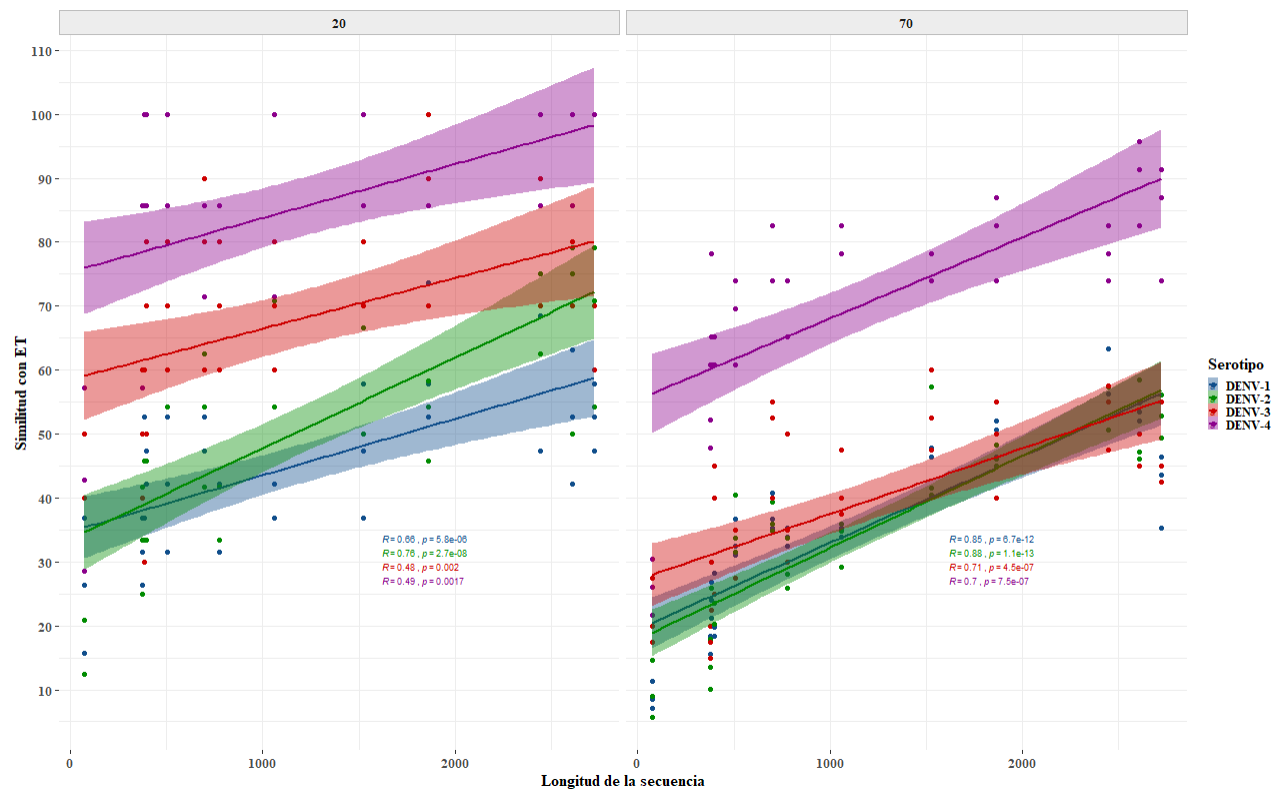


Figura 4. Porcentaje de nodos comunes respecto a la filogenia de evidencia total según la longitud de la secuencia utilizada. Se muestra la similitud entre los árboles de cada uno de los serotipos

El uso de genes estructurales y genes no estructurales en el tamaño del 20 % no afectó la similitud de las topologías del submuestreo respecto a la de evidencia total, esto cambia al aumentar el número de terminales, ya que utilizando el tamaño del 70 % los genes estructurales

presentaron una mayor similitud con evidencia total en comparación con los no estructurales. Por otra parte, el uso de combinaciones de genes tampoco evidenció gran diferencia respecto a estructurales y no estructurales en el tamaño del muestreo del 20%, pero al aumentar el número terminales se obtiene una mayor similitud con evidencia total e incluso con una desviación estándar menor (ver Figura 5).

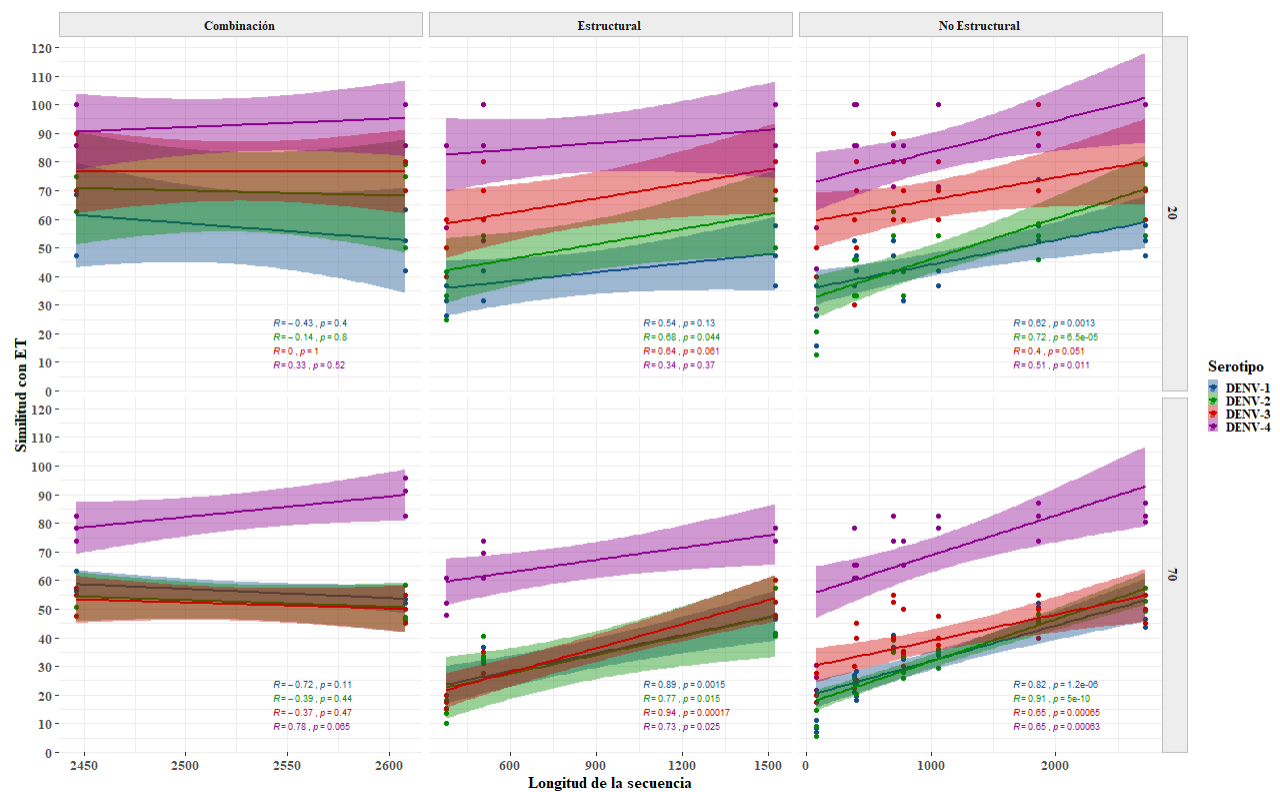


Figura 5. Porcentaje de nodos en común de las topologías obtenidas de genes estructurales, genes no estructurales y combinaciones respecto a evidencia total. Se muestra la similitud entre los árboles de cada uno de los serotipos

En todos los casos la topología de DENV4 tuvo más nodos en común con evidencia total.

La similitud de DENV3 es mayor a DENV1 y DENV2 en el muestreo del 20%, pero al aumentar

el tamaño de muestreo la cantidad de nodos en común es similar en los tres serotipos sin que esto se vea afectado por la longitud ni el tipo de partición que se utilice (ver Figura 5).

4. Discusión

4.1. Modelos de sustitución nucleotídica

Los resultados de la selección de modelos en cuanto a las frecuencias de bases determinaron que son desiguales para todos los genes a excepción de la región P2K, esto concuerda con estudios anteriores donde distintas particiones del genoma de DENV han ajustado a este tipo de modelos; NS5 (Regato et al., 2008; Drumond et al., 2016), Gen E (Afreen et al., 2015; Villabona-Arenas y de Andrade Zanotto, 2013; Twiddy et al., 2002), Genoma completo (Steel et al., 2010) y NS1 (Aryati et al., 2013; Ciccozzi et al., 2014).

Los valores del Alfa de la heterogeneidad Gamma en todas las particiones son inferiores a 1, esto indica una alta heterogeneidad entre las tasas mutacionales de los sitios. Yang (1996) reporta que estos valores corresponden generalmente a genes completos, esto concuerda con nuestros resultados ya que todos los genes usados en este estudio son secuencias completas. Los valores altos en este parámetro también indican una alta tasa mutacional y concuerda con lo propuesto en estudios realizados sobre la alta heterogeneidad en el virus del dengue la cual se debe a los cambios de presiones de selección entre hospederos y a los diferentes frentes evolutivos relacionados con la ubicación geográfica (Zanotto et al., 1996; Lanciotti et al., 1994; Rico-Hesse, 1990).

4.2. Reloj y dinámica poblacional en el virus del Dengue

El ajuste de modelos de reloj relajado no correlacionado era esperable ya que las tasas evolutivas entre los distintos serotipos es diferencial (Costa et al., 2012), adicionalmente las constantes fluctuaciones poblacionales debido al aumento acelerado durante un brote y posterior reducción de la población también alteran las tasas mutacionales además de la diversidad genética (Bennett et al., 2009). Este ajuste de modelos de reloj no correlacionados en DENV ha sido reportado anteriormente (Kumar et al., 2010; Dunham y Holmes, 2007; Carrington et al., 2005). Sin embargo, evaluar el modelo en cada genotipo podría dar una idea más clara sobre el ajuste de reloj estricto a una menor escala, ya que el ajuste de este modelo al interior de los serotipos ha sido reportado anteriormente (Twiddy et al., 2003).

El ajuste de los genes a modelos poblacionales variables a través del tiempo, concuerda con resultados anteriores (Jiménez-Silva et al., 2018), esto puede estar relacionado con la dinámica poblacional producto de la competencia entre los diferentes frentes evolutivos en una misma región, lo cual conlleva a fluctuaciones en la diversidad genética además de cambios en la prevalencia de los serotipos y los genotipos de DENV a través del tiempo (Allicock et al., 2012; Zhang et al., 2005; Bennett et al., 2009).

4.3. Recuperación de nodos

4.3.1. Recuperación de serotipos y genotipos. La recuperación de serotipos y genotipos se vio más afectada por el tamaño del muestreo que por la región del genoma utilizada, trabajos anteriores han demostrado que la adición o extracción de ciertos taxa genera un efecto positivo en la estimación de las relaciones filogenéticas (Hillis, 1998; Zwickl y Hillis, 2002), sin embargo estos estudios fueron hechos con un número bajo de terminales. La baja recuperación de los grupos monofiléticos en el tamaño de muestreo del 70 % y en la topología de evidencia total, puede deberse a que en la estimación de una filogenia usando un número elevado de terminales es más probable que se generen inconsistencias provocando poca exactitud en la topología (Kim, 1996).

Por otra parte, la eficacia de las regiones del genoma de mayor longitud en la recuperación de los serotipos y genotipos se ha comprobado con anterioridad; especialmente los genes NS3, NS5 y gen E han reportado un alto porcentaje en la recuperación de estos grupos (Klungthong et al., 2008). Sin embargo, también se ha evidenciado eficacia diferencial estas regiones en la recuperación de grupos dentro de determinados serotipos, lo cual concuerda con los resultados obtenidos ya que las particiones de longitud intermedia recuperan más genotipos de DENV2 que las de mayor longitud, esto puede ser debido a que ciertas particiones poseen mayor señal filogenética para determinados serotipos (Rosenberg y Kumar, 2001).

4.3.2. Similaridad topológica. Las topologías obtenidas a partir de secuencias de mayor longitud tuvieron un porcentaje más alto de similitud con la topología de evidencia total, esto concuerda con estudios anteriores en donde reportan que la longitud de la secuencia está relacionada positivamente con la similaridad a el árbol de referencia (Kuhner y Felsenstein, 1994). Adicionalmente, el muestreo de menor tamaño presentó mayor similaridad con evidencia total, esto se puede explicar ya que se ha reportado que la adición de taxa a un análisis puede afectar negativamente la estimación (Kim, 1998).

Las topologías obtenidas a partir de combinaciones de genes presentaron un mayor porcentaje de similaridad, este mejor desempeño en la estimación filogenética ha sido reportado anteriormente (Pamilo y Nei, 1988; Corl y Ellegren, 2013). Es importante tener en cuenta esto ya que estas combinaciones tienen menor cantidad de caracteres que el gen NS5, lo cual representa menor tiempo de computo.

5. Conclusiones

Comparamos las filogenias obtenidas a partir de los distintas particiones del genoma y los porcentajes de terminales propuestos, dichas filogenias fueron estimadas en el programa BEAST v2.5.2 utilizando los modelos de sustitución y priores obtenidos durante la exploración. Todos los resultados del trabajo se encuentran en el presente informe. Dado lo anterior, podemos confirmar que todas la competencias fueron adquiridas por el pasante.

La recuperación de serotipos y en mayor medida la de genotipos es sensible a la longitud de la región del genoma utilizada en el análisis y al número de terminales incluidos, esto puede llevar a falsos positivos al tratar de genotipificar una secuencia a través de su filogenia. Por lo tanto, es necesario tener en cuenta estos dos factores al momento de realizar la estimación. Por otra parte, el uso de combinaciones de genes o los genes de mayor longitud y un número de terminales reducido permite un mayor acercamiento al resultado que se tendría usando evidencia total.

6. Recomendaciones

Las particiones de mayor longitud de caracteres especialmente las combinaciones de genes demostraron ser más eficientes en la recuperación de las relaciones filogenéticas, se recomienda su uso para la genotipificación de muestras.

La recuperación de las relaciones al interior de los serotipos es diferencial, se hace necesario la profundización sobre el impacto que genera la partición utilizada en la estimación de estas relaciones.

Referencias Bibliográficas

- Afreen, N., Naqvi, I. H., Broor, S., Ahmed, A., y Parveen, S. (2015). Phylogenetic and molecular clock analysis of dengue serotype 1 and 3 from new delhi, india. *PLoS One*, 10(11):e0141628.
- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.
- Allicock, O. M., Lemey, P., Tatem, A. J., Pybus, O. G., Bennett, S. N., Mueller, B. A., Suchard, M. A., Foster, J. E., Rambaut, A., y Carrington, C. V. (2012). Phylogeography and population dynamics of dengue viruses in the americas. *Molecular Biology and Evolution*, 29(6):1533–1543.
- Araújo, J. M. G. a. d., Bello, G., Schatzmayr, H. G., Santos, F. B. d., y Nogueira, R. M. R. (2009). Dengue virus type 3 in brazil: a phylogenetic perspective. *Memórias do Instituto Oswaldo Cruz*, 104(3):526–529.
- Aryati, A., Trimarsanto, H., Yohan, B., Wardhani, P., Fahri, S., y Sasmono, R. T. (2013). Performance of commercial dengue ns1 elisa and molecular analysis of ns1 gene of dengue viruses obtained during surveillance in indonesia. *BMC Infectious Diseases*, 13(1):611.
- Barban, V., Munoz-Jordan, J. L., Santiago, G. A., Mantel, N., Girerd, Y., Gulia, S., Claude, J.-B., y Lang, J. (2012). Broad neutralization of wild-type dengue virus isolates following immuniza-

- tion in monkeys with a tetravalent dengue vaccine based on chimeric yellow fever 17d/dengue viruses. *Virology*, 429(2):91–98.
- Bennett, S., Drummond, A., Kapan, D., Suchard, M., Munoz-Jordan, J., Pybus, O., Holmes, E., y Gubler, D. (2009). Epidemic dynamics revealed in dengue evolution. *Molecular Biology and Evolution*, 27(4):811–818.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., y Sayers, E. W. (2012). Genbank. *Nucleic Acids Research*, 41(D1):D36–D42.
- Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4):e1006650.
- Brister, J. R., Ako-Adjei, D., Bao, Y., y Blinkova, O. (2014). Ncbi viral genomes resource. *Nucleic Acids Research*, 43(D1):D571–D577.
- Carrington, C. V., Foster, J. E., Pybus, O. G., Bennett, S. N., y Holmes, E. C. (2005). Invasion and maintenance of dengue virus type 2 and type 4 in the americas. *Journal of Virology*, 79(23):14680–14687.
- Chambers, T. J., Hahn, C. S., Galler, R., y Rice, C. M. (1990). Flavivirus genome organization, expression, and replication. *Annual Reviews in Microbiology*, 44(1):649–688.
- Cherian, S. S. y Mourya, D. T. (2016). Phylodynamics of dengue viruses in india vis-à-vis the global scenario. *Dengue*, 39:41.

- Ciccozzi, M., Presti, A. L., Cella, E., Giovanetti, M., Lai, A., El-Sawaf, G., Faggioni, G., Vescio, F., Al Ameri, R., De Santis, R., et al. (2014). Phylogeny of dengue and chikungunya viruses in al hudayda governorate, yemen. *Infection, Genetics and Evolution*, 27:395–401.
- Corl, A. y Ellegren, H. (2013). Sampling strategies for species trees: the effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. *Molecular Phylogenetics and Evolution*, 67(2):358–366.
- Costa, R. L., Voloch, C. M., y Schrago, C. G. (2012). Comparative evolutionary epidemiology of dengue virus serotypes. *Infection, Genetics and Evolution*, 12(2):309–314.
- Darriba, D., Taboada, G. L., Doallo, R., y Posada, D. (2012). jmodeltest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8):772.
- Dash, P., Sharma, S., Srivastava, A., Santhosh, S., Parida, M., Neeraja, M., Subbalaxmi, M., Lakshmi, V., y Rao, P. (2011). Emergence of dengue virus type 4 (genotype i) in india. *Epidemiology & Infection*, 139(6):857–861.
- De Souza, R. P., Rocco, I. M., Maeda, A. Y., Spenassatto, C., Bisordi, I., Suzuki, A., Silveira, V. R., Silva, S. J., Azevedo, R. M., Tolentino, F. M., et al. (2011). Dengue virus type 4 phylogenetics in brazil 2011: looking beyond the veil. *PLoS Neglected Tropical Diseases*, 5(12):e1439.
- Dos Santos Carmo, A. M., Suzuki, R. B., Cabral, A. D., da Costa, R. T., Massari, G. P., Riquena, M. M., Fracasso, H. A. A., Eterovic, A., Marcili, A., y Sperança, M. A. (2017). Co-circulating

serotypes in a dengue fever outbreak: Differential hematological profiles and phylogenetic relationships among viruses. *Journal of Clinical Virology*, 90:7–13.

Drummond, A. J., Ho, S. Y., Phillips, M. J., y Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., y Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Rambaut, A., Shapiro, B., y Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Drumond, B. P., Fagundes, L. G. d. S., Rocha, R. P., Fumagalli, M. J., Araki, C. S., Colombo, T. E., Nogueira, M. L., Castilho, T. E., Silveira, N. J. F. d., Malaquias, L. C. C., et al. (2016). Phylogenetic analysis of dengue virus 1 isolated from south minas gerais, brazil. *Brazilian Journal of Microbiology*, 47(1):251–258.

Dunham, E. J. y Holmes, E. C. (2007). Inferring the timescale of dengue virus evolution under realistic models of dna substitution. *Journal of Molecular Evolution*, 64(6):656–661.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

- Ernst, T., McCarthy, S., Chidlow, G., Luang-Suarkia, D., Holmes, E. C., Smith, D. W., y Imrie, A. (2015). Emergence of a new lineage of dengue virus type 2 identified in travelers entering western australia from indonesia, 2010-2012. *PLoS Neglected Tropical Diseases*, 9(1):e0003442.
- Franco, L., Pagan, I., Del Cor, N. S., Schunk, M., Neumayr, A., Molero, F., Potente, A., Hatz, C., Wilder-Smith, A., Sánchez-Seco, M., et al. (2015). Molecular epidemiology suggests venezuela as the origin of the dengue outbreak in madeira, portugal in 2012–2013. *Clinical Microbiology and Infection*, 21(7):713–e5.
- Gubler, D. J. (2002). Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends in Microbiology*, 10(2):100–103.
- Guzman, M. G., Gubler, D. J., Izquierdo, A., Martinez, E., y Halstead, S. B. (2016). Dengue infection. *Nature Reviews Disease Primers*, 2:16055.
- Hapuarachchi, H. C., Koo, C., Kek, R., Xu, H., Lai, Y. L., Liu, L., Kok, S. Y., Shi, Y., Chuen, R. L. T., Lee, K.-S., et al. (2016). Intra-epidemic evolutionary dynamics of a dengue virus type 1 population reveal mutant spectra that correlate with disease transmission. *Scientific Reports*, 6:22592.
- Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47(1):3–8.
- Holder, M. y Lewis, P. O. (2003). Phylogeny estimation: traditional and bayesian approaches. *Nature Reviews Genetics*, 4(4):275.

- Holmes, E. C. y Twiddy, S. S. (2003). The origin, emergence and evolutionary genetics of dengue virus. *Infection, Genetics and Evolution*, 3(1):19–28.
- Holmes, E. C., Worobey, M., y Rambaut, A. (1999). Phylogenetic evidence for recombination in dengue virus. *Molecular biology and evolution*, 16(3):405–409.
- Jiménez-Silva, C. L., Carreño, M. F., Ortiz-Baez, A. S., Rey, L. A., Villabona-Arenas, C. J., y Ocazonez, R. E. (2018). Evolutionary history and spatio-temporal dynamics of dengue virus serotypes in an endemic region of colombia. *PloS one*, 13(8):e0203090.
- Kass, R. E. y Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- Kim, J. (1996). General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology*, 45(3):363–374.
- Kim, J. (1998). Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology*, 47(1):43–60.
- Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248.
- Klungthong, C., Putnak, R., Mammen, M., Li, T., y Zhang, C. (2008). Molecular genotyping of dengue viruses by phylogenetic analysis of the sequences of individual genes. *Journal of Virological Methods*, 154(1-2):175–181.

- Knipe, D. M. y Howley, P. M. (2007). *Fields Virology*, volume 4, chapter 33, pages 1101–1152. Lippincott-Raven Publishers, Philadelphia, 5 edition.
- Kotaki, T., Yamanaka, A., Mulyatno, K. C., Churrotin, S., Sucipto, T. H., Labiqah, A., Ahwanah, N. L. F., Soegijanto, S., Kameoka, M., y Konishi, E. (2016). Divergence of the dengue virus type 2 cosmopolitan genotype associated with two predominant serotype shifts between 1 and 2 in surabaya, indonesia, 2008–2014. *Infection, Genetics and Evolution*, 37:88–93.
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., y Stamatakis, A. (2019). RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. btz305.
- Kuhner, M. K. y Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.
- Kumar, S. R., Patil, J. A., Cecilia, D., Cherian, S. S., Barde, P. V., Walimbe, A. M., Yadav, P. D., Yergolkar, P. N., Shah, P. S., Padbidri, V. S., et al. (2010). Evolution, dispersal and replacement of american genotype dengue type 2 viruses in india (1956–2005): selection pressure and molecular clock analyses. *Journal of General Virology*, 91(3):707–720.
- Kyle, J. L. y Harris, E. (2008). Global spread and persistence of dengue. *Annual Review of Microbiology*, 62:71–92.
- Lanciotti, R. S., Lewis, J. G., Gubler, D. J., y Trent, D. W. (1994). Molecular evolution and epidemiology of dengue-3 viruses. *Journal of General Virology*, 75(1):65–75.

- Larsson, A. (2014). Aliview: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278.
- Leitmeyer, K. C., Vaughn, D. W., Watts, D. M., Salas, R., Villalobos, I., Ramos, C., Rico-Hesse, R., et al. (1999). Dengue virus structural differences that correlate with pathogenesis. *Journal of virology*, 73(6):4738–4747.
- Leng, C.-H., Liu, S.-J., Tsai, J.-P., Li, Y.-S., Chen, M.-Y., Liu, H.-H., Lien, S.-P., Yueh, A., Hsiao, K.-N., Lai, L.-W., et al. (2009). A novel dengue vaccine candidate that induces cross-neutralizing antibodies and memory immunity. *Microbes and Infection*, 11(2):288–295.
- Mackenzie, J. S., Gubler, D. J., y Petersen, L. R. (2004). Emerging flaviviruses: the spread and resurgence of japanese encephalitis, west nile and dengue viruses. *Nature Medicine*, 10(12s):S98.
- Mukhopadhyay, S., Kuhn, R. J., y Rossmann, M. G. (2005). A structural perspective of the flavivirus life cycle. *Nature Reviews Microbiology*, 3(1):13.
- Nascimento, F. F., dos Reis, M., y Yang, Z. (2017). A biologist's guide to bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1(10):1446.
- Neyman, J. y Pearson, E. S. (1933). IX. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.
- Nunes, M. R. T., Faria, N. R., Vasconcelos, H. B., de Almeida Medeiros, D. B., de Lima, C. P. S., Carvalho, V. L., da Silva, E. V. P., Cardoso, J. F., Sousa Jr, E. C., Nunes, K. N. B., et al. (2012).

- Phylogeography of dengue virus serotype 4, brazil, 2010–2011. *Emerging Infectious Diseases*, 18(11):1858.
- Pagel, M. y Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6):808–825.
- Pamilo, P. y Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583.
- Perera, R. y Kuhn, R. J. (2008). Structural proteomics of dengue virus. *Current Opinion in Microbiology*, 11(4):369–377.
- Posada, D. y Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808.
- Rambaut, A., Drummond, A. J., Xie, D., Baele, G., y Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67(5):901–904.
- Regato, M., Recarey, R., Moratorio, G., de Mora, D., Garcia-Aguirre, L., González, M., Mosquera, C., Alava, A., Fajardo, A., Alvarez, M., et al. (2008). Phylogenetic analysis of the ns5 gene of dengue viruses isolated in ecuador. *Virus Research*, 132(1-2):197–200.
- Rico-Hesse, R. (1990). Molecular evolution and distribution of dengue viruses type 1 and 2 in nature. *Virology*, 174(2):479–493.

- Rosenberg, M. S. y Kumar, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences*, 98(19):10751–10756.
- Schreiber, M. J., Holmes, E. C., Ong, S. H., Soh, H. S., Liu, W., Tanner, L., Aw, P. P., Tan, H. C., Ng, L. C., Leo, Y. S., et al. (2009). Genomic epidemiology of a dengue virus epidemic in urban singapore. *Journal of Virology*, 83(9):4163–4173.
- Silva Cala, A. L. y Miranda Esquivel, D. R. (2018). Similaridad genética, tiempo y geografía en el virus del dengue. Tesis pregrado, Universidad Industrial de Santander. Escuela de Biología.
- Steel, A., Gubler, D. J., y Bennett, S. N. (2010). Natural attenuation of dengue virus type-2 after a series of island outbreaks: a retrospective phylogenetic study of events in the south pacific three decades ago. *Virology*, 405(2):505–512.
- Swofford, D. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0a165*. Sinauer Associates, Sunderland, Massachusetts.
- Tolou, H., Couissinier-Paris, P., Durand, J.-P., Mercier, V., de Pina, J.-J., De Micco, P., Billoir, F., Charrel, R., y De Lamballerie, X. (2001). Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences. *Journal of General Virology*, 82(6):1283–1290.
- Twiddy, S. S., Farrar, J. J., Chau, N. V., Wills, B., Gould, E. A., Gritsun, T., Lloyd, G., y Holmes, E. C. (2002). Phylogenetic relationships and differential selection pressures among genotypes of dengue-2 virus. *Virology*, 298(1):63–72.

- Twiddy, S. S., Holmes, E. C., y Rambaut, A. (2003). Inferring the rate and time-scale of dengue virus evolution. *Molecular Biology and Evolution*, 20(1):122–129.
- Vasilakis, N. y Weaver, S. C. (2008). The history and evolution of human dengue emergence. *Advances in Virus Research*, 72:1–76.
- Vicente, C., Pannuti, C., Urbano, P., Felix, A., Junior, C. C., Herbinger, K.-h., Fröschl, G., y Romano, C. (2018). First phylogenetic analysis of dengue virus serotype 4 circulating in espírito santo state, brazil, in 2013 and 2014. *Epidemiology & Infection*, 146(1):100–106.
- Villabona-Arenas, C. J. y de Andrade Zanotto, P. M. (2011). Evolutionary history of dengue virus type 4: insights into genotype phylodynamics. *Infection, Genetics and Evolution*, 11(5):878–885.
- Villabona-Arenas, C. J. y de Andrade Zanotto, P. M. (2013). Worldwide spread of dengue virus type 1. *PloS One*, 8(5):e62649.
- Wasserstein, R. L., Lazar, N. A., et al. (2016). The asa’s statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.
- Weaver, S. C. y Vasilakis, N. (2009). Molecular evolution of dengue viruses: contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. *Infection, genetics and evolution*, 9(4):523–540.
- WHO (2000). Strengthening implementation of the global strategy for dengue fever/dengue hae-

morrhagic fever prevention and control. report of the informal consultation, 18-20 october 1999, who hq, geneva, switzerland. *Report of the Informal Consultation*.

Wittke, V., Robb, T., Thu, H., Nisalak, A., Nimmannitya, S., Kalayanrooj, S., Vaughn, D., Endy, T., Holmes, E., y Aaskov, J. (2002). Extinction and rapid emergence of strains of dengue 3 virus during an interepidemic period. *Virology*, 301(1):148–156.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372.

Zanotto, P. d., Gould, E. A., Gao, G. F., Harvey, P. H., y Holmes, E. C. (1996). Population dynamics of flaviviruses revealed by molecular phylogenies. *Proceedings of the National Academy of Sciences*, 93(2):548–553.

Zhang, C., Mammen, M. P., Chinnawirotpisan, P., Klungthong, C., Rodpradit, P., Monkongdee, P., Nimmannitya, S., Kalayanaroj, S., y Holmes, E. C. (2005). Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence. *Journal of Virology*, 79(24):15123–15130.

Zuckerkandl, E. y Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier.

Zwickl, D. J. y Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, 51(4):588–598.

Apéndice A

Tabla 5

Modelos con mayor peso AIC por réplica para cada partición

Partición	Número de caracteres	Réplica 1		Réplica 2		Réplica 3	
		Modelo	Peso AIC	Modelo	Peso AIC	Modelo	Peso AIC
P2K	75	TrNef+G	0.2263	TrNef+G	0.2648	TrNef+G	0.2823
		TIM1ef+G	0.0928	TIM1ef+G	0.1371	TIM3ef+G	0.1171
		K80+G	0.0913	TIM2ef+G	0.0984	TIM1ef+G	0.1066
		TIM2ef+G	0.0882	TIM3ef+G	0.0976	TIM2ef+G	0.1039
		TIM3ef+G	0.0846	TrN+G	0.0888	TrN+G	0.0743
Capside	376	TIM2+G	0.7589	TIM2+G	0.5958	TIM1+G	0.3519
		GTR+G	0.1249	GTR+G	0.1876	GTR+G	0.2545
		TrN+G	0.0612	TIM1+G	0.0912	TIM2+G	0.2296
NS4A	383	TIM2+G	0.3959	TrN+G	0.2923	TIM2+G	0.3796
		TrN+G	0.2331	TIM2+G	0.2567	TrN+G	0.2719
		TIM1+G	0.1326	TIM1+G	0.2388	TIM3+G	0.1317
		TIM3+G	0.1200	TIM3+G	0.1080	TIM1+G	0.1263
		GTR+G	0.1182	GTR+G	0.1040	GTR+G	0.0895
NS2B	397	GTR+G	0.8125	GTR+G	0.6389	GTR+G	0.4749
		TIM1+G	0.1602	TIM2+G	0.1545	TIM1+G	0.1965
		TIM2+G	0.0130	TIM1+G	0.0895	TIM3+G	0.1828
PrM	505	TIM2+G	0.5342	TIM2+G	0.7753	TIM2+G	0.6945
		TrN+G	0.1819	GTR+G	0.1172	GTR+G	0.1228
		GTR+G	0.1258	TrN+G	0.0557	TrN+G	0.0967
NS2A	698	TrN+G	0.3740	TrN+G	0.3792	TrN+G	0.4323
		TIM2+G	0.2234	TIM1+G	0.2308	TIM2+G	0.1948
		TIM1+G	0.2045	TIM3+G	0.2047	TIM1+G	0.1733
		TIM3+G	0.1498	TIM2+G	0.1427	TIM3+G	0.1690
NS4B	777	TIM2+G	0.6454	GTR+G	0.5070	TIM2+G	0.3939
		GTR+G	0.2922	TIM2+G	0.4148	GTR+G	0.3447
		TIM3+G	0.0283	TIM3+G	0.0497	TIM3+G	0.1648
		TrN+G	0.0248	TrN+G	0.0177	TrN+G	0.0699
NS1	1058	GTR+G	0.3305	GTR+G	0.7564	TIM2+G	0.3756
		TIM2+G	0.2059	TIM1+G	0.1179	TrN+G	0.2414
		TIM3+G	0.1798	TIM3+G	0.1069	TIM3+G	0.1560
		TIM1+G	0.1744	TIM2+G	0.0121	TIM1+G	0.1183
		TrN+G	0.1094	TrN+G	0.0067	GTR+G	0.1088
GenE	1524	GTR+G	0.9198	GTR+G	0.8639	GTR+G	0.7989
		TIM2+G	0.0802	TIM2+G	0.1361	TIM2+G	0.2010
NS3	1863	GTR+G	0.7323	GTR+G	0.6300	TIM2+G	0.6351
		TIM2+G	0.2668	TIM2+G	0.3587	GTR+G	0.3630
NS5	2723	TIM2+G	0.8049	TIM2+G	0.7838	TIM2+G	0.8414
		GTR+G	0.1950	GTR+G	0.2160	GTR+G	0.1586

Nota: Las denominaciones de los modelos corresponden a los del programa JModelTest v2.1.

Tabla 6

Matriz de transición y frecuencias de bases en cada partición

P2K						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,0000	7,5261:10,8434	1,0000	A	0,2500
C		*	1,0000	16,1641:22,2944	C	0,2500
G			*	1,0000	G	0,2500
T				*	T	0,2500

Capside						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,0000:2,3651	3,7155:9,4700	0,5013:2,3651	A	0,3317:0,3454
C		*	0,5013:1,0000	9,2431:27,3402	C	0,2073:0,2125
G			*	1,0000	G	0,2444:0,2572
T				*	T	0,2036:0,2039

NS4A						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,0000:1,6050	7,2141:10,0724	1,0000:1,6050	A	0,3016:0,3083
C		*	1,0000	22,3772:30,6271	C	0,2458:0,2509
G			*	1,0000	G	0,2265:0,2426
T				*	T	0,2088:0,2104

NS2B						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	3,3575:3,8385	18,8053:23,6601	4,1885:6,4437	A	0,3213:0,3236
C		*	3,9566:4,7678	44,5504:60,9540	C	0,2290:0,2340
G			*	1,0000	G	0,2278:0,2331
T				*	T	0,2124:0,2156

PrM						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,6523:0,9103	9,7300:9,8921	1,6523:1,9103	A	0,3012:0,3072
C		*	1,0000	24,0452:25,2392	C	0,2423:0,2510
G			*	1,0000	G	0,2419:0,2455
T				*	T	0,2045:0,2092

NS2A						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,0000	7,1091:7,5811	1,0000	A	0,3015:0,3046
C		*	1,0000	16,3723:18,0065	C	0,2205:0,2216
G			*	1,0000	G	0,2329:0,2354
T				*	T	0,2410:0,2425

NS4B						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,1990:1,8279	9,7268:11,9672	1,6180:1,9808	A	0,3246:0,3284
C		*	0,7494:1,0000	21,8013:29,6679	C	0,2387:0,2460
G			*	1,0000	G	0,2103:0,2132
T				*	T	0,2159:0,2226

NS1						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	0,9082:1,3594	9,2077:10,2307	1,3594:2,1624	A	0,3475:0,3510
C		*	1,0000:1,1114	24,4753:27,6889	C	0,2298:0,2343
G			*	1,0000	G	0,2215:0,2230
T				*	T	0,1946:0,1995

Gen E						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,7413:1,8202	9,7292:10,9252	2,7718:3,0091	A	0,3400:0,3424
C		*	1,2225:1,3224	34,4559:38,3641	C	0,2190:0,2217
G			*	1,0000	G	0,2429:0,2446
T				*	T	0,1940:0,1954

NS3						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,4404:1,7875	9,0078:12,3770	1,6851:2,5487	A	0,3541:0,3583
C		*	1,0000:1,4291	29,4456:39,6750	C	0,2169:0,2205
G			*	1,0000	G	0,2295:0,2328
T				*	T	0,1917:0,1971

NS5						
Matriz Q					Frecuencias	
	A	C	G	T		
A	*	1,6366:1,7611	8,2050:9,0205	1,6366:1,7611	A	0,3518:0,3532
C		*	1,0000	24,2367:26,9161	C	0,2241:0,2248
G			*	1,0000	G	0,2377:0,2397
T				*	T	0,1824:0,1865

Tabla 7

Frecuencias de bases y número de tipos de cambio de los modelos por partición

Partición	Modelo	Frecuencias de bases	Tipos de cambio	Parámetros libres
P2K	TrNef	Iguales	AC=AT=CG=GT;AG;GT	2
Capside	TIM1uf/TIM2uf	Desiguales	AC=GT;AT=CG;AG;CT	6
			AC=AT;CG=GT;AG;CT	
NS4A	TrN/TIM2uf	Desiguales	AC=AT=CG=GT;AG;GT AC=AT;CG=GT;AG;CT	6
NS2B	GTR	Desiguales	AC;CG;AT;GT;AG;CT	8
PrM	TIM2uf	Desiguales	AC=AT;CG=GT;AG;CT	6
NS2A	TrN	Desiguales	AC=AT=CG=GT;AG;GT	5
NS4B	TIM2uf/GTR	Desiguales	AC=AT;CG=GT;AG;CT AC;CG;AT;GT;AG;CT	6/8
NS1	TIM2uf/GTR	Desiguales	AC=AT;CG=GT;AG;CT AC;CG;AT;GT;AG;CT	6/8
Gen E	GTR	Desiguales	AC;CG;AT;GT;AG;CT	8
NS3	TIM2uf/GTR	Desiguales	AC=AT;CG=GT;AG;CT AC;CG;AT;GT;AG;CT	6/8
NS5	TIM2uf	Desiguales	AC;CG;AT;GT;AG;CT	6

Tabla 8

Test de ajuste a reloj estricto por gen para las tres réplicas

Partición	Réplica 1		Réplica 2		Réplica 3	
	valor p	Ajuste	valor p	Ajuste	valor p	Ajuste
P2K	2,45E-008	No reloj estricto	3,88E-045	No reloj estricto	4,41E-040	No reloj estricto
Capside	1,76E-124	No reloj estricto	2,22E-145	No reloj estricto	8,85E-039	No reloj estricto
NS4A	2,58E-132	No reloj estricto	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto
NS2B	3,95E-159	No reloj estricto	3,72E-265	No reloj estricto	2,37E-165	No reloj estricto
PrM	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	8,45E-101	No reloj estricto
NS2A	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto
NS4B	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto
NS1	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	2,04E-272	No reloj estricto
GenE	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	5,56E-269	No reloj estricto
NS3	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto
NS5	1,01E-073	No reloj estricto	0,00E+000	No reloj estricto	0,00E+000	No reloj estricto