# Análisis composicionales de genomas virales

**Lic. Diego Simón**

Maestría en Bioinformática - PEDECIBA

Laboratorio de Genómica Evolutiva
Facultad de Ciencias
Universidad de la República, Uruguay

Montevideo
2021

Orientador: Dr. Héctor Musto
Co-orientadora: Dra. María Inés Fariello


# Análisis composicionales de genomas virales

**Lic. Diego Simón**

Tesis presentada con el objetivo de obtener el título de Magíster en Bioinformática en el marco del Programa de Desarrollo de las Ciencias Básicas (PEDECIBA)

Orientador: Dr. Héctor Mario Musto Mancebo, Profesor Titular
Co-orientadora: Dra. María Inés Fariello Rico, Asistente

Montevideo
2021

A mi Nona

**Agradecimientos**

Montevideo, noviembre de 2021

Estimable lector:

Espero que disfrute la lectura. Le aseguro que parece más largo de lo que es.

Antes de comenzar quiero agradecer a algunos colectivos que me acompañaron durante estos años de Maestría en Bioinformática.

Gracias familia. Siempre he sentido muy cercano el apoyo incondicional.
Especialmente a Carla, por todo el amor y por mucho mate salvador.
Carmen, Raúl y Elena: los quiero mucho.

Héctor: por tu confianza y ayudarme a crecer en la academia y como persona.
Maine: por acompañarme en este proceso y por tu empuje. Gracias a AVIVEN.

Al Piso 4. Por enseñarme más sobre la vida que por sobre todo lo demás.
A Matías Rodríguez por su solidaridad. Al Pelo por tanto. Al Yuyo por otro tanto.
A María Noel, Lucía, Luisa, Pantera, Andrés y Eugenio. Disculpen algún olvido...
Al Laboratorio de Neurociencias, con mención a Tony, Daniel, Leo, Felipe y Fran.
A Ana, Betta y demás circadianos por sumarme. Adriana Migliaro por el apoyo.

A Mora, Pili y todo el LVM. Aún queda mucho por delante y lo mejor está por venir.

A la Facultad de Ciencias. Infinitas gracias a compañeros, docentes y funcionarios.
Al Instituto Pasteur y su gente.

A PEDECIBA Bioinformática por la oportunidad de aspirar a este título de Magíster.
Muchas gracias también por la financiación brindada en varias oportunidades.

A la ANII, gracias por el apoyo otorgado para acompañar el desarrollo de mi tesis.

*«Yo creo que en el "mundo orgánico" cada caso complejo se basa en cosas más simples, y éstas a su vez sobre cosas aún más simples».*

*«Cada caso debe ser reducido a sus términos más simples».*

ø

Friedrich Miescher (1844-1895)

*Fragmentos de una de sus cartas a Rudolf Böhm (1871)*

---

ø © Biblioteca en la Universidad de Basilea (fotografía reproducida en Dahm, 2010).

## Resumen

La aplicación de técnicas cromatográficas derivó en el descubrimiento de que tanto las secuencias de ADN como de ARN muestran composiciones genómicas altamente variables. Los primeros estudios composicionales preceden al modelo de doble hélice de Watson y Crick. Más aún, la dilucidación de la estructura tridimensional del ADN tuvo como insumo fundamental algunos resultados de estudios composicionales. Décadas más tarde, con los primeros genomas disponibles fue posible estudiar en mayor detalle su heterogeneidad, tanto intra-genómica como entre genomas. La genómica composicional, ahora más computacional que molecular, fue una de las ramas que más aprovecharon esta disponibilidad de secuencias creciente.

El objetivo general de esta tesis fue realizar un análisis exhaustivo de la composición nucleotídica y de uso de codones de toda la diversidad viral conocida, con genoma completo secuenciado. Se presentan los principales aspectos detrás del código desarrollado durante la tesis, así como el software utilizado, los datos generados y diferentes bases de datos accedidas. Posteriormente, se describen los análisis composicionales llevados a cabo, seguido de los análisis estadísticos. Por último, se presentan y discuten resultados no publicados, así como las principales contribuciones de los dos artículos publicados en revistas arbitradas hasta el momento de la presentación de esta tesis. Los análisis composicionales, al margen de su contribución histórica, siguen siendo pertinentes. Más aún en sistemas hospedero-virus, donde queda de manifiesto el impacto de la composición nucleotídica y el sistema inmune del hospedero en el genoma de sus virus.



**Palabras clave:** composición genómica, sesgos composicionales, dinucleótidos, uso de codones, frecuencia de aminoácidos, interacción hospedero-virus, clasificación de Baltimore

## Abstract

Chromatographic methods led to the discovery that both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are compositionally heterogeneous. Early compositional studies predate Watson and Crick's double helix model. Moreover, the elucidation of the three-dimensional structure of DNA involved compositional studies as a crucial insight. Decades later, with the first genomes available it was possible to study in greater detail their heterogeneity, both within and between genomes. Compositional genomics, now more computational than molecular, was one of the fields that took most advantage of this growing availability of sequences.

The overall aim of this thesis was to perform a comprehensive analysis of the nucleotide composition and codon usage of all known viral diversity. The main aspects behind the code developed are presented, as well as the data generated. Subsequently, the compositional analyses carried out are described, followed by the statistical analyses. Finally, unpublished results are presented and discussed, as well as the main contributions of the two articles published in peer-reviewed journals up to the time of the presentation of this thesis. Apart from their historical contribution, compositional analyses remain relevant. Even more so in host-virus systems, given that viral genomes are exposed to the nucleotide composition and the immune system of the hosts that infect.



**Keywords:** genomic composition, compositional biases, dinucleotides, codon usage, amino acid frequencies, host-virus interaction, Baltimore classification

**Tabla de contenidos**

# Introducción

## Un poco de historia

El conocimiento de los ácidos nucleicos se inicia con Friedrich Miescher en 1868 (Dahm, 2010). Estos inicios fueron en la Universidad de Tubinga, en el laboratorio de Felix Hoppe-Seyler, uno de los más grandes y mejor equipados de la época. Hoppe-Seyler estudiaba la composición química de las células, en una época en que el concepto de célula aún se debatía (Chan y Conova, 2011b). A Miescher se le encargó estudiar los linfocitos. De núcleos de leucocitos logró aislar una sustancia de propiedades diferentes a las proteínas, a la que llamó «nucleína» (Dahm, 2005).

Los trabajos de Miescher bajo el mando de Hoppe-Seyler fueron interrumpidos abruptamente por la guerra franco-prusiana (1870-1871). Miescher, de regreso a su ciudad natal, ocupó un cargo en la Universidad de Basilea a partir de 1872, continuando allí con sus aislamientos de núcleos y sus derivados (Greenstein, 1943). El término «ácido nucleico» fue creado en 1889 por Richard Altmann, un antiguo estudiante de Miescher (Altmann, 1889).

Hoppe-Seyler se mudó en 1872 a la Universidad de Estrasburgo.[1] En 1877, Albrecht Kossel se convierte en asistente de Hoppe-Seyler y «hereda» parte de los trabajos de Miescher. Kossel describió que la nucleína estaba formada por un componente proteico y por un grupo prostético, de naturaleza ácida. Entre 1885 y 1901, su laboratorio consiguió aislar por hidrólisis a las cinco bases nitrogenadas que típicamente conforman a los ácidos nucleicos: adenina (A), guanina (G), citosina (C), timina (T) y uracilo (U) (Nobel Prize, 1910). Kossel, en solitario, recibió el Premio Nobel en Fisiología o Medicina (1910) «en reconocimiento a las contribuciones al conocimiento de la química celular a través de sus trabajos sobre las proteínas, incluidas las sustancias nucleicas» (Nobel Prize, 1910).

---

[1] La Universidad de Estrasburgo, como consecuencia de la derrota francesa en la guerra franco prusiana, volvía a adoptar el alemán, idioma que había mantenido desde su fundación hasta la Revolución francesa. La universidad recupera definitivamente el idioma francés luego de la Primera Guerra Mundial.

## Agentes filtrables

La virología comienza posiblemente con Adolf Mayer a finales de 1870 en la Universidad de Wageningen, Países Bajos. Mayer describió la enfermedad del mosaico del tabaco y demostró que esta podía transmitirse a plantas sanas simplemente inoculando la savia de plantas enfermas (Mayer, 1886).

En Rusia, Dmitri Ivanovsky realizó entre 1887 y 1892 varias contribuciones al estudio de las enfermedades infecciosas de la planta del tabaco (Ivanovsky y Polovtsev, 1890; Ivanovsky, 1892). Determinó que el agente infeccioso descrito por Mayer era capaz de pasar por un filtro de Chamberland, diseñado para retener bacterias (Ivanovsky, 1892).

Hacia finales del siglo XIX, Martinus Beijerinck reprodujo independientemente los experimentos de filtrado de Ivanovsky. Beijerinck evidenció que el agente causal de dicha enfermedad se replicaba en las plantas vivas, y que este agente era de naturaleza no bacteriana: un virus (del latín *virus*, «veneno»; Beijerinck, 1898).

En su definición actual, un virus es un conjunto de moléculas de ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN), encapsuladas por una cubierta proteica. Son agentes infecciosos que solamente pueden reproducirse dentro de una célula hospedera. La partícula infectiva o virión debe ingresar a una célula viva y utilizar la maquinaria de síntesis de dichas células para replicarse. Infectan todas las formas de vida conocidas, e incluso existen aquellos que parasitan otros virus.

## ADN y ARN

En 1905, Phoebus Levene ingresa al Instituto Rockefeller en Nueva York, hoy Universidad Rockefeller, donde se abocará el resto de su vida académica a identificar los componentes de los ácidos nucleicos. Cabe destacar que Levene, previo a recibir su puesto en Rockefeller, tuvo un breve pasaje por el laboratorio de Kossel, la autoridad indiscutida en ácidos nucleicos en esa época (Chan y Conova, 2011c). Levene se convertiría en las décadas siguientes en la nueva autoridad. Términos como nucleósido, nucleótido o polinucleótido son algunos de sus aportes.

Levene y su estudiante Walter Jacobs reconocieron a la D-ribosa como un componente esencial de los ácidos nucleicos (Levene y Jacobs, 1909). Posteriormente, se determinó que la D-ribosa era el azúcar constituyente del ARN. En aquel entonces no se tenía conocimiento de las diferencias químicas y biológicas entre los ácidos nucleicos. Simplemente se los distinguía por los materiales de los que se aislaba. De esta forma, el ARN era el «ácido nucleico de la levadura», mientras que el ADN era el «ácido nucleico del timo» (obtenido de terneros; es decir, mollejas). Levene junto a Efim London identificaron a la desoxirribosa (Levene y London, 1929), el azúcar constituyente del ADN.

Levene propuso, erróneamente, que el ADN estaba formado por cantidades iguales de A, G, C y T. Su hipótesis de los «tetranucleótidos» sostenía que el ADN era una repetición monótona de los cuatro nucleótidos en sucesión; esta homogeneidad le había llevado a concluir que el ADN no podía almacenar la información genética.

Alrededor de 1940, Alexander Todd comenzó a investigar los nucleótidos y los nucleósidos, obteniendo por sus descubrimientos el Premio Nobel en Química (Nobel Prize, 1957). Hacia 1950, había descubierto y sintetizado las coenzimas trifosfato de adenosina (ATP, de *adenosine triphosphate*) y el dinucleótido flavina-adenina (FAD, de *flavin-adenine dinucleotide*).

## El material genético

Con cada descubrimiento se conocía más sobre la química de los ácidos nucleicos, pero no estaba muy clara su implicancia biológica. Un descubrimiento fundamental fue el «principio transformador» ("*transforming principle*") por Frederick Griffith (1928). Griffith concluyó que algún componente aislado de bacterias virulentas muertas por calor «transformaba» cepas bacterianas no virulentas en virulentas, pero no logró identificar la naturaleza química de este «principio» (más allá del hecho de que era capaz de sobrevivir al tratamiento térmico).

Hubo que esperar hasta 1944 cuando Oswald Avery, Colin MacLeod y Maclyn McCarty determinaron que el ADN de bacterias muertas podía «transformar» bacterias vivas (Avery et al., 1944). Este experimento situó al ADN como el portador

de la información genética, en lugar de las proteínas como se creía hasta ese entonces.

En 1920, Hans Winkler había creado el término «genoma», definido como todo el material genético de un individuo o de una especie (Winkler, 1920). En los organismos de vida celular consiste en una o varias moléculas de ADN. También en muchos virus es el ADN el ácido nucleico que compone su genoma, pero existen otros cuyo genoma está compuesto de ARN. Dicho genoma, además, puede ser de cadena simple o doble, lineal o circular, no segmentado o segmentado, etc.

En los virus, la información genética suele estar codificada de forma muy compacta. Los denominados «virus gigantes», en cambio, exhiben genomas tan grandes y complejos como los procariotas o algunos eucariotas parásitos (Philippe et al., 2013).

Bacteriófagos modelos

Frederick Twort realizó el primer descubrimiento de virus que atacan a colonias bacterianas en el *Brown Institution* en Londres (Twort, 1915). Independientemente, Félix d'Hérelle el Instituto Pasteur en París también descubrió los virus de bacterias (d'Herelle, 1917). Fue d'Hérelle quien los llamó bacteriófagos (del griego, «devoradores de bacterias»). Además, d'Hérelle fue pionero en la «fagoterapia» y en proponer el uso de «cócteles» ("*phage cocktails*") para sobreponerse a las bacterias que se vuelven resistentes contra un fago aplicado en solitario.

Es indudable el papel protagónico de los virus en la historia de la biología molecular.[2] La biología molecular provoca una verdadera revolución en la manera de aproximarse al estudio de la vida. Viejas preguntas de la biología comienzan a ser abordadas desde otras ciencias como la física y la química. La simplicidad de los sistemas bacteria-fago los posiciona como modelos experimentales estupendos.

---

[2] Término atribuible a Warren Weaver en 1938 (Weaver, 1970).

El máximo exponente en cuanto al uso de virus como modelo fue el *Phage group* («Grupo de los fagos»; a partir de aquí será el «Grupo»). Una breve reseña al Grupo es útil para reflejar lo que se estaba generando entre las décadas de 1930 y 1940.

Max Delbrück, formado en física, se interesa en entender el mecanismo por el cual la infección de una sola partícula generaba algunos cientos una media hora más tarde. Delbrück publica en 1939 junto con Emory Ellis «El crecimiento de bacteriófagos» (Ellis y Delbrück, 1939).

En 1940 Delbrück conoce a Salvador Luria y comienzan una firme colaboración. Junto a Alfred Hershey, forman el Grupo en 1943 con el propósito de comprender el mecanismo de replicación de los fagos. Delbrück, como líder del Grupo, acordó que la investigación debía concentrarse en un conjunto de siete fagos (T1-T7), todos los cuales infectan al mismo hospedero (*Escherichia coli* B), y utilizando condiciones experimentales estandarizadas ("*phage treaty*"; Comfort y Goldstein, 1995).

En 1952, Hershey y Martha Chase realizaron un experimento que afianzó el estatus del ADN como el material genético, apoyando los resultados del experimento de Avery, MacLeod y McCarty (1944). Utilizando al bacteriófago T2 lograron confirmar que es el ADN el portador de la información genética (Hershey y Chase, 1952).

Delbrück, Hershey y Luria compartieron el Premio Nobel en Fisiología o Medicina (1969) «por sus descubrimientos sobre el mecanismo de replicación y la estructura genética de los virus» (Nobel Prize, 1969).

Además de los nombrados, entre los numerosos miembros del Grupo de los fagos, destacaron los Premios Nobel James Watson, Renato Dulbecco y Sydney Brenner, el dúo Matthew Meselson y Franklin Stahl (1958) y también Seymour Benzer[3].

---

[3] Recomiendo sumergirse en los primeros capítulos de «Tiempo, amor, memoria» (Weiner, 1999). Los inicios de Benzer vinculado al *Phage group* permiten un corto viaje por aquellos años (con una breve visita a Thomas Morgan, Alfred Sturtevant y demás miembros del *Drosophila group*).

Cromatografía en papel

Los avances en la química de los ácidos nucleicos, principalmente mediante la aplicación de técnicas cromatográficas, llevaron al conocimiento de que tanto las secuencias de ADN como de ARN muestran una gama de composiciones características de su origen (i.e., de donde fueron aislados).

Erwin Chargaff, doctorado en química en la Universidad Técnica de Viena, da inicio en 1935 a cuatro décadas de trabajo en la Universidad de Columbia, en Nueva York. El experimento de Avery, MacLeod y McCarty (1944) fue de gran influencia para Chargaff, a tal punto que reorganizó su laboratorio para testar su nueva hipótesis: las diferencias genéticas deben reflejarse en diferencias químicas entre ADNs (según testimonio de Seymour Cohen, estudiante de Chargaff en Columbia; Cohen, 2004). Chargaff fue pionero en el estudio de la composición del ADN por técnicas cromatográficas (Chargaff et al., 1949).

Roy Markham y John Smith, en la Universidad de Cambridge, fueron también fundacionales en los estudios cromatográficos en ácidos nucleicos, tanto de ADN (1949a) como de ARN (1949b). También en 1949, Markham y Smith analizaron cinco cepas del virus del mosaico del tabaco, posiblemente el primer estudio composicional en virus. Describen que la composición nucleotídica de estas cepas es muy similar, si bien detectaron diferencias significativas entre tres de las cepas (Markham y Smith, 1950).

En 1950, los análisis realizados por Chargaff y sus colaboradores demostraron que la composición del ADN, definida en sus trabajos seminales como «las proporciones molares de las bases de purina y pirimidina», es característica de la especie; además, es constante para los diferentes tejidos de una especie (Chargaff et al., 1950).

Llegar a obtener estos resultados permitió, a su vez, importantes avances en técnicas cromatográficas y en métodos de hidrólisis. Gerard Wyatt, estudiante de Smith, logró identificar otras moléculas relevantes, que también son constituyentes naturales del ADN, como la 5-metilcitosina (Wyatt, 1950) y la 5-hidroximetilcitosina (Wyatt y Cohen, 1953).

La dóble hélice

Cabe destacar que los primeros estudios composicionales preceden al modelo de doble hélice del ADN de Watson y Crick (1953a). La publicación de la doble hélice fue realizada desde los Laboratorios Cavendish en la Universidad de Cambridge (www.phy.cam.ac.uk), uno de los departamentos más renombrados en cuanto a descubrimientos científicos se refiere (e.g., el electrón, el neutrón, los isótopos, la fisión nuclear artificial, las primeras desintegraciones nucleares controladas inducidas por partículas aceleradas de alta energía, el desarrollo de la cristalografía de rayos X). No ha de sorprender que por Cavendish pasaran 30 Premios Nobel en su casi siglo y medio de historia.

La dilucidación de la estructura del ADN tuvo dos insumos fundamentales, uno proveniente de estudios composicionales y otro de los patrones de difracción de rayos X del ADN.

Watson y Crick asistieron a la presentación de Chargaff en la Universidad de Cambridge en 1952 (Betz, 2011). Estos resultados de Chargaff pueden resumirse como:

(**i**) A = T ≠ G = C

Esto dejaba sin fundamento la hipótesis de los «tetranucleótidos» de Levene, aceptada por décadas al no existir evidencia en contra.

Wyatt, por su parte, enseñó sus resultados a Watson en un encuentro que tuvieron en el Instituto Pasteur de París, brindando más evidencias aún a las relaciones estequiométricas descritas por Chargaff. Muchos de sus resultados fueron obtenidos en colaboración con Cohen y Hershey, y utilizando fagos de ADN (e.g., T2, T4, T6) (Watson, 1968).

El otro insumo fue la famosa fotografía 51 tomada en el *King's College* en Londres en mayo de 1952. Rosalind Franklin capturó por difracción de rayos X las fibras de ADN cristalizadas por Raymond Gosling (Franklin y Gosling, 1953). Esta fotografía, mostrada a Watson por Maurice Wilkins, sin consentimiento de sus autores (Betz, 2011), evidenció la estructura helicoidal del ADN.

La publicación de la doble hélice, que en realidad consistió de dos trabajos en serie, separados por pocos meses, fue un hito para la biología (Watson y Crick, 1953a; 1953b). Su impacto trascendió incluso a la disciplina y permeó en la cultura y la sociedad.

Los resultados de Chargaff y de Wyatt, pese a ser citados como corresponde en ambas publicaciones, fueron discutidos allí solo como evidencia experimental al modelo de la doble hélice y no como insumos necesarios para conceptualizarlo. Lo antedicho, sin ánimo de menoscabar el mérito de Watson y Crick, pretende balancear cuánto pudo haber de ingenio e intuición parte de estos talentosos científicos, y cuánto de inspiración en resultados parciales de otros colegas.

Crick, Watson y Wilkins obtuvieron el Premio Nobel en Fisiología o Medicina 1962 «por sus descubrimientos sobre la estructura molecular de los ácidos nucleicos y su relevancia para la trasmisión de información en los seres vivos» (Nobel Prize, 1962).

Crick y Watson, junto a otros célebres científicos de la época, formaron parte del *RNA tie club*.[4] Liderados por el físico George Gamow, se dedicaron a otro tema candente de la biología molecular: «resolver el enigma de la estructura del ARN y entender cómo se construyen las proteínas» (Chan y Conova, 2011a).

## Polimerasas

Marianne Grunberg-Manago, realizando un posdoctorado en el laboratorio de Severo Ochoa en la Escuela de Medicina de la Universidad de Nueva York, descubrió en 1954 la enzima polinucleótido fosforilasa (Grunberg-Manago y Ochoa, 1955). Grunberg-Manago se encontraba estudiando el ATP a partir de componentes celulares extraídos de la bacteria *Azotobacter vinelandii*. De manera fortuita, identificó dicha enzima, capaz de sintetizar cadenas de polinucleótidos a partir de ribonucleótidos en ausencia de un molde.

---

[4] Sugiero complementar esta parte de la historia en «El Club de la Corbata ARN», capítulo de Héctor Romero (2010) en el libro «Biología: Unidad en la diversidad».

En el laboratorio de Arthur Kornberg, que años atrás fuera investigador posdoctoral de Ochoa, se descubrió en 1955 que la replicación del ADN requiere de una enzima específica: la ADN polimerasa (Kornberg, 1957).

Ochoa y Kornberg serían galardonados con el Premio Nobel en Fisiología o Medicina (1959) «por su descubrimiento de los mecanismos de la síntesis biológica del ARN y del ADN» (Nobel Prize, 1959).

Lo cierto es que la enzima aislada por Grunberg-Manago no era una ARN polimerasa propiamente dicha. Una polimerasa es una enzima que sintetiza cadenas (o polímeros) de ácidos nucleicos (ADN o ARN) a partir de una secuencia complementaria como molde (Case y Hingorani, 2017).

Los primeros reportes sobre las ARN polimerasas fueron publicados, en un mismo número, por Audrey Stevens (1960) y por Jerard Hurwitz y colaboradores (1961).

## El código genético

Para descifrar el código genético, o cómo se traducía la información genética desde el ADN a las proteínas, fueron fundamentales otros abordajes y diferentes aportes de otros investigadores que no pertenecían al *RNA Tie Club*.

Marshall Nirenberg llevó adelante una serie de experimentos simples y elegantes. Utilizando extractos de *E. coli,* con la polinucleótido fosforilasa como protagonista, el abordaje inicial fue obtener secuencias de ARN sintéticas para determinar la correspondencia entre dichas secuencias y los aminoácidos.

Con Heinrich Matthaei sintetizaron una secuencia de ARN compuesta solo por U. Dicha secuencia poliU fue agregada en veinte tubos conteniendo cada uno un aminoácido diferente; el resultado obtenido fue que solamente ocurrió reacción para fenilalanina (Nirenberg y Matthaei, 1961).

Junto a Philip Leder, incorporan secuencias de ARN en una relación 2:1 entre C y U, obteniendo mayoritariamente los aminoácidos serina, leucina y fenilalanina. Pero no era posible obtener una correspondencia directa entre aminoácidos y nucleótidos. Nirenberg y Leder identifican que en la interacción ARN-ribosomas existen

oligonucleótidos que se corresponden de manera específica para la formación de enlaces peptídicos, con un tamaño mínimo de tres nucleótidos o trinucleótidos (Nirenberg y Leder, 1964).

Un pequeño paréntesis al respecto de los trinucleótidos y el código genético. Considerando que las bases en el ARN (también en el ADN) son cuatro, y que los aminoácidos son veinte, no parecía posible que correspondiera con dos nucleótidos (combinaciones de 4 tomados de a 2 = 16). Tal como fue propuesto originalmente por Gamow, el código se leía de a tripletes, de manera no solapada y sin pausas. Esto fue demostrado experimentalmente en *E. coli*, utilizando fagos T4 (Crick et al., 1961).[5]

Otras correspondencias entre tripletes y aminoácidos se fueron develando en los siguientes años, gracias fundamentalmente a los aportes que Har Khorana hizo al construir diferentes cadenas de ARN y producir polipéptidos (Khorana et al., 1965). Las secuencias de aminoácidos sintetizadas a partir de moldes de ARN sintéticos fueron ayudando a resolver todo el código.

Robert Holley, por su parte, fue el primero en aislar un ARN de transferencia (ARNt), y en determinar su secuencia (Holley et al., 1965a) y su estructura (Holley et al., 1965b).

Nirenberg, Khorana y Holley repartieron el Premio Nobel en Fisiología o Medicina 1968 por sus «interpretaciones del código genético y su función en la síntesis de proteínas» (Nobel Prize, 1968).

## Transcripción inversa

El descubrimiento de la transcriptasa inversa fue un hito de la biología molecular. Fue realizado en 1970 de manera simultánea por David Baltimore (1970) y por Howard Temin (junto a Satoshi Mizutani; Temin y Mizutani, 1970).

---

[5] Como testimonio de lo que se conocía hasta el momento sobre el código genético, está disponible la exposición (Nobel Lecture) ofrecida por Francis Crick al recibir su Premio Nobel (1962).

Baltimore y Temin reciben, junto a Dulbecco,[6] el Premio Nobel en Fisiología o Medicina (1975) «por sus descubrimientos sobre la interacción entre los virus tumorales y el material genético de la célula» (Nobel Prize, 1975).

[6] Pese a no haber estado implicado en el descubrimiento de la transcriptasa inversa, había sido director de tesis de Temin (1960) y supervisor de Baltimore (1965-1968). Dulbecco estuvo vinculado al *Phage group* como investigador postdoctoral tanto de Delbruch como de Luria (Academic Tree, 2005).

## Clasificación de los virus

Si bien cada vez se describen más virus, conocemos solamente una proporción minúscula de la diversidad viral existente en la naturaleza. Para intentar clasificar esta diversidad extraordinaria es que necesitamos sistemas de clasificación viral.

Un intento pionero, aunque arcaico, fue la clasificación de Francis Holmes (1948). Holmes asignó a los diferentes virus conocidos hasta el momento según qué grupos de hospederos infectaban: «Phaginae» (bacterias), «Phytophaginae» (plantas) y «Zoophaginae» (animales). Este sistema no se popularizó dado que pasaba por alto grandes similitudes morfológicas entre miembros de esos grupos (Kuhn, 2021).

Otro sistema, más aceptado en su momento, fue el denominado LHT (sigla formada por las iniciales de sus creadores: Lwoff, Horne y Tournier, 1962). El sistema LHT dividía a los virus en dos grupos según el material genético presente en el virión: «Deoxyvira» (ADN) y «Ribovira» (ARN); dentro de cada grupo se subdividían según la simetría con la que se disponen las subunidades de la cápside de los viriones (Kuhn, 2021).

En la actualidad suelen utilizarse dos sistemas de clasificación no excluyentes. Ambos sistemas de clasificación serán descritos a continuación.

### Taxonomía viral

En 1966 se fundó el Comité Internacional de Nomenclatura de Virus (ICNV, del inglés *International Committee on Nomenclature of Viruses*) con la misión de «desarrollar una taxonomía de virus mundialmente respetada y aplicable a todos los tipos de virus de todas las formas de vida» (Kuhn, 2021). En 1974 el ICNV se convirtió en el actual Comité Internacional de Taxonomía de Virus (ICTV, del inglés *International Committee on Taxonomy of Viruses*). ICTV organiza la taxonomía de virus y de agentes subvirales (como los viroides y los virus satélites). Además, autoriza modificaciones y/o nuevas especies virales y taxones superiores, a partir de propuestas de la comunidad virológica, que son evaluadas por grupos de expertos (disponible desde talk.ictvonline.org/ictv_wikis).

Desde su primer informe (1971), donde ICTV presentó 2 familias, 27 géneros, 10 subgéneros y 18 «otros grupos», se presentan nuevos informes casi año a año. Junto con el número de virus y secuencias conocidas, el esquema de clasificación se ha ampliado considerablemente desde entonces (ICTV Executive Committee, 2020). Actualmente, la taxonomía de ICTV (presentada en octubre de 2020 y ratificada en marzo 2021) se organiza en 6 imperios, 10 reinos, 17 phyla, 39 clases, 59 órdenes, 189 familias, 2224 géneros y 9110 especies virales (disponible desde talk.ictvonline.org/taxonomy).

El *National Center for Biotechnology Information* (NCBI) intenta seguir el ritmo de actualizaciones del ICTV, incorporándolos a su taxonomía (Schoch et al., 2020; disponible desde ncbi.nlm.nih.gov/taxonomy).

Grupos de Baltimore

Baltimore propuso en 1971 un sistema de clasificación de virus («*viral genetic system*»; Baltimore, 1971b), que agrupaba los virus en seis clases (Baltimore, 1971a), numeradas con números romanos.

La clasificación de Baltimore agrupa a los virus según como la información va desde el genoma viral al ARN mensajero (ARNm; ver **Figura 1**); de esta manera, incorpora información sobre:

(**i**) El tipo de material genético (ADN o ARN).

(**ii**) Si son doble o simple hebra (ss y ds representan simple y doble hebra [*strand*], respectivamente; elegimos usar la notación en inglés por no encontrar un consenso claro en español).

(**iii**) Para el caso de los virus de ARN de simple hebra, de acuerdo a su polaridad (+ o -).

(**iv**) Y si son retro-transcriptos (RT) o no.

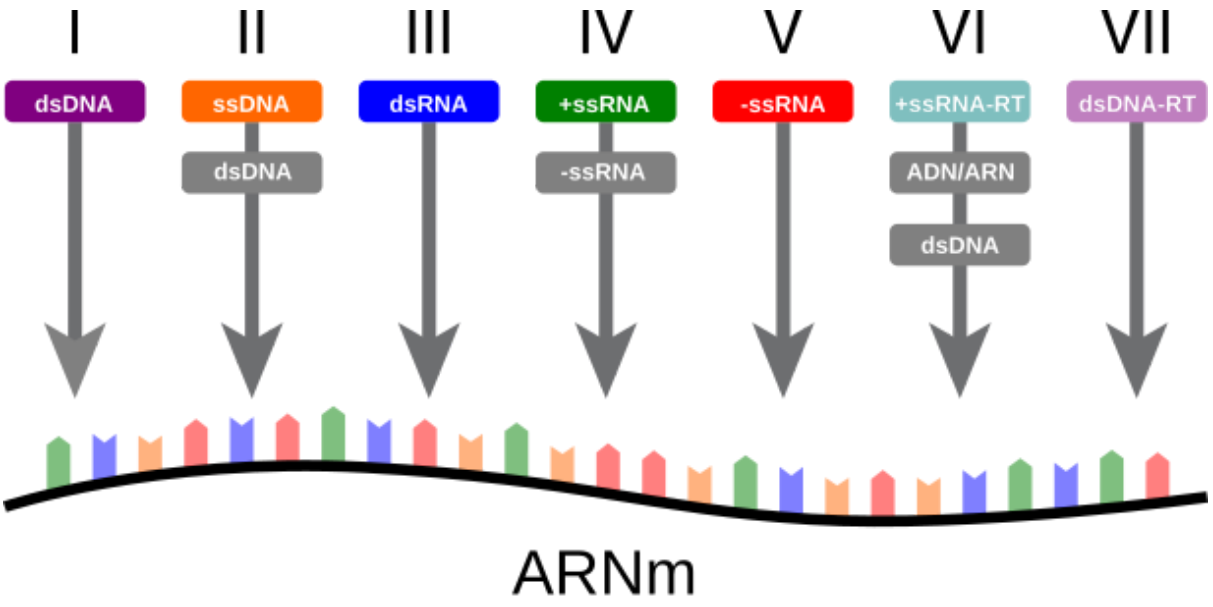



**Figura 1.** Esquema de los caminos que cada grupo de Baltimore recorre para sintetizar su ARNm (modificado de Splettstoesser, 2012).

La clasificación de Baltimore actual se organiza en siete grupos: virus de ADN de doble cadena (dsDNA o grupo I); virus de ADN de cadena única (ssDNA o grupo II); virus de ARN de doble cadena (dsRNA o grupo III); virus de ARN de cadena única positivos (+ssRNA o grupo IV); virus de ARN de cadena única negativos (-ssRNA o grupo V); virus retro-transcritos de ARN de cadena única positivos con intermediarios de ADN (+ssRNA-RT o grupo VI); y los virus retro-transcritos de ADN de doble cadena (dsDNA-RT o grupo VII).

## Bioinformática

El término bioinformática fue compuesto, en la acepción gramatical de la palabra, hace ya medio siglo, siendo definida como el estudio de los «procesos relacionados con la información en los sistemas biológicos» (Hesper y Hogeweg, 1970). Si bien esta definición no refleja el sentido actual, el término bioinformática sirvió a posteriori para reunir diferentes aproximaciones surgidas en las décadas anteriores.

Actualmente, la bioinformática es el desarrollo y la aplicación de técnicas informáticas al estudio de las biomoléculas a escala «ómica» (Gerstein, 1998). Es un área multidisciplinaria donde la biología se nutre de otras ciencias como la computación, la estadística y la matemática.

### La era genómica

En 1976, Walter Fiers y sus colaboradores publicaron el genoma completo del bacteriófago *MS2*. Este virus fue el primer genoma en ser secuenciado (Fiers et al., 1976).

Un año más tarde, Frederick Sanger y sus colaboradores publicaron el genoma completo del bacteriófago *φX174*, el primer genoma de ADN secuenciado (Sanger et al., 1977).

La genómica comparativa emergió como disciplina a mediados de la década de 1980 gracias a la disponibilidad de genomas virales completos. Eckard Wimmer y su grupo compararon genomas de virus ARN de los géneros *Picornavirus* y *Comovirus*, que infectan a animales y plantas, respectivamente; estos géneros comparten una gran similitud de secuencia y, en menor medida, el orden de sus genes (o sintenia), lo que permite agruparlos dentro del orden *Picornavirales* (Argos et al., 1984).

Dumcan McGeoch y Andrew Davison publicaron en 1986 el primer estudio de genómica comparativa en virus de ADN, al comparar los genomas del virus de la varicela-zóster y del virus de Epstein-Barr, ambos pertenecientes a la familia *Herpesviridae*, analizando sus más de 100 genes (McGeoch y Davison, 1986).

Uso de codones

En el código genético estándar hay tres tripletes de parada, restando sesenta y uno para codificar los aminoácidos. Dado que son solamente veinte los aminoácidos canónicos, la mayoría de ellos necesariamente deben estar codificados por más de un codón. Esta redundancia determina que el código genético sea degenerado. Los codones que se traducen al mismo aminoácido son denominados «sinónimos» (Kano-Sueoka y Sueoka, 1969).

Disponer de las primeras secuencias codificantes permitió observar que existe una gran variabilidad entre las distintas especies en el uso de los codones sinónimos (Grantham et al., 1980), lo que se denomina «sesgo en el uso de codones».

Gracias a la mayor disponibilidad de secuencias pudo evidenciarse que los codones sinónimos no son utilizados en un genoma de manera uniforme. Estas diferencias fueron atribuídas en un principio a la selección natural (*translational selection*; Jeffery et al., 1981): los codones «mayores» u «óptimos» se verían favorecidos porque así aumenta la eficiencia y la precisión de la traducción (Ikemura, 1981).

Los genes más expresados tienen una tendencia a usar los codones que coinciden con los ARNt más frecuentes, evitando además el uso de codones «menores». De la misma manera, los ARNt isoaceptores más abundantes son aquellos que reconocen los codones más frecuentes en los genes más expresados (Gouy y Gautier, 1982).

Lo expuesto en las frases anteriores fue estudiado principalmente en bacterias, sobre todo en *E. coli*. Pero también en eucariotas el sesgo en el uso de codones parece ser más extremo en los genes de alta expresión para ajustarse al sesgo de los ARNt (Bulmer, 1991).

En virus, el sesgo en el uso de codones es considerado como uno de los principales determinantes de la eficiencia de la traducción. A su vez, la optimización o des-optimización del uso de codones es una forma común de manipular la capacidad de replicación, la idoneidad y la virulencia de un virus dentro de un hospedero determinado (Burns et al., 2006; Mueller et al., 2006; Plotkin y Kudla, 2010; Jorge et al., 2015).

# Objetivos

## Objetivo general

El objetivo general de esta tesis es realizar un análisis exhaustivo de la composición nucleotídica y de uso de codones de toda la diversidad viral conocida, con genoma completo secuenciado.

## Objetivos específicos

Del objetivo general se desprenden los siguientes objetivos específicos:

**(i)** Integrar diferentes bases de datos biológicas primarias y secundarias.

**(ii)** Desarrollar herramientas bioinformáticas y estadísticas de análisis composicional.

**(iii)** Describir patrones composicionales según la clasificación de Baltimore.

**(iv)** Asociar patrones composicionales a sus hospederos.

## Hipótesis

Si bien esta tesis se presenta como un estudio descriptivo, se sostienen dos hipótesis no excluyentes:

**(i)** Los virus presentan una enorme diversidad evidenciada a muchos niveles. Sus genomas no son la excepción: presentan genomas de ADN o de ARN, que más aún pueden ser de cadena doble o simple, lineales o circulares, etc. Estos factores intrínsecos al genoma viral influyen en su composición nucleotídica.

**(ii)** Los virus dependen de la maquinaria traduccional de la célula que infectan. Sesgos propios de sus hospederos, sobre todo a nivel de su uso de codones, impacta a su vez en el sesgo composicional y el uso de codones de los virus.

## Desarrollo

Luego de la introducción de esta tesis, donde se priorizan aspectos históricos, continúan algunas secciones que requieren ser descritas brevemente. En ellas se incluyen productos de estos años que van desde código desarrollado y/o utilizado, resultados aún no publicados y dos trabajos arbitrados por pares y ya publicados:

> **(i)** *Host influence in the genomic composition of flaviviruses: A multivariate approach* (Simón et al., 2017)
>
> **(ii)** *Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts* (Simón et al., 2021)

Previo a cada uno se presentan los principales aportes y contribuciones. Luego de cada uno se encuentran además las principales conclusiones. El cierre de la tesis presenta además un capítulo con consideraciones finales y algunas perspectivas.

Tal vez la manera más razonable de continuar la lectura es ir a este último artículo. Es el más abarcativo y aquel que mejor se alinea con el objetivo principal de la tesis.

Los resultados no publicados pueden ser vistos como (y no de manera excluyente) análisis complementarios a los publicados, estudios en curso y/o caminos sin salida. Algunos de estos abordajes son a partir de un subset de datos (e.g., algunos virus que utilizan códigos genéticos alternativos, virus de plantas, etc). Esta característica lo asemeja al primer artículo publicado durante la tesis, dado que fue un análisis con un conjunto acotado de virus.

La metodología utilizada en los resultados publicados se encuentra solamente en cada respectivo manuscrito, así como la literatura citada. A continuación se presenta como se realizó el código de análisis composicionales y estadísticos, no pensando en un caso particular, si no en cualquier especie viral a analizar. Empieza con la sección «leéme» (inspirada en lo que puede encontrarse en un archivo README).

## Materiales y métodos

### Léeme

El primer bloque de materiales y métodos describe la motivación y los principales aspectos detrás del código desarrollado durante la tesis, así como el software utilizado, los datos generados y diferentes bases de datos accedidas. Posteriormente, se describirán los análisis composicionales llevados a cabo, seguido de los análisis estadísticos.

Todos los scripts se disponibilizan, alojados en el GitLab de Facultad de Ingeniería (disponible desde gitlab.fing.edu.uy/lompa/squeezeR), junto con las tablas generadas para los genomas virales analizados; `squeezeR` (en inglés, exprimidor) se inspira en este utensilio de cocina (**Figura 2**) para «exprimir» los genomas, aunque en realidad puede ser usado con cualquier secuencia nucleotídica,[7] utilizando R, principalmente, y algo de Python.

 [8]

**Figura 2.** Un exprimidor (*squeezer*) de limas o limones.

---

[7] Si bien se ha desarrollado originalmente para analizar genomas virales, `squeezeR` puede ser utilizado con cualquier secuencia nucleotídica, sin importar su taxonomía.

[8] Foto de Alan Levine (disponible desde flickr.com/photos/cogdog/39225338552).

Descripción del código

Esta tesis ha sido desarrollada por completo en el sistema operativo Ubuntu (Linux). Dado que los *scripts* aquí presentados están escritos en Python (v3) o en R (v4), son independientes del sistema operativo y servirán, en teoría, en otros sistemas.

Tanto Python como R son lenguajes de programación interpretados. Esto refiere a que estos lenguajes son convertidos a un lenguaje que la computadora entiende (interpreta) a medida que es ejecutado (**Figura 3**), a diferencia de otros lenguajes que requieren ser previamente compilados como ocurre con C, C++ o Java.

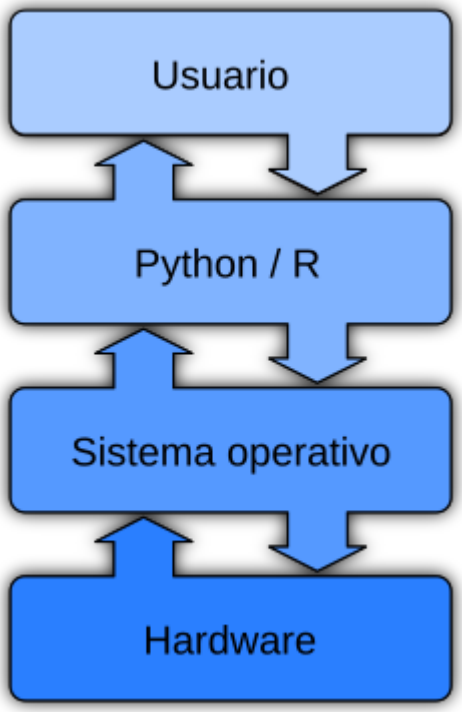



**Figura 3.** Esquema de interacciones entre usuario y los demás componentes informáticos: lenguajes de computación, sistema operativo y hardware (modificado de Golftheman, 2010).

Conda es un sistema de administración y gestión de entornos virtuales (*environments*) para lenguajes de programación (disponible desde docs.conda.io/en/latest/). Es muy popular en la comunidad de usuarios de Python, pero acostumbrarse al gestor de entornos es también muy provechoso para usuarios de R.

Un buen punto de inicio es instalar Anaconda, una distribución de conda lista para usar en diferentes sistemas operativos y con muchos paquetes adicionales (disponible desde anaconda.com). Una vez iniciado un nuevo proyecto, una buena práctica es crear un entorno para dicho proyecto con las versiones de software necesarias; una vez dentro de dicho entorno, toda modificación al código de paquetes de Python o de R no modificarán los paquetes del sistema base (ni de otros entornos). En aquellos casos en que se utilicen funciones extras a los paquetes básicos de Python o R, y se requiera su instalación, se aclarará oportunamente en la documentación.

Python

La principal motivación de usar Python surge más como desafío que por necesidad. En el piso 4 ala norte de la Facultad de Ciencias hay un ambiente muy colaborativo; podemos pasar más tiempo intentando solucionar problemas ajenos, con éxito o no, que problemas propios. En una oportunidad estaba discutiendo algo de mi código (escrito en Bash) con Fernando Álvarez y la conversación derivó a que no estaba respetando las «buenas prácticas» de Bash, un lenguaje principalmente creado para *one-liners*. Este «desafío» hizo que dejara esos largos y feos scripts en Bash y pasara todo ese abordaje a Python. Esos esfuerzos se plasman en el script `gbx.py`. Lo que hace dicho script, y de manera más elegante que lo hecho en Bash, es recorrer los archivos GenBank (gbk o gb), extrayendo regiones genómicas de interés.

Además de, obviamente, Python, requiere Biopython (Cock et al., 2009), que se instala de varias maneras (disponible desde github.com/biopython/biopython).

Si algo no corre como debe, un paso lógico sería actualizar estos paquetes. En muchas situaciones lo lógico es exactamente lo contrario: realizar un *downgrade* (i.e., volver a una versión anterior). Conda permite instalar versiones anteriores. Esto puede hacerse tanto por interfaz gráfica (`navigator`) como desde la terminal.

Luego de crear este entorno con la versión de Python seleccionada, se puede también elegir qué versión instalar de los paquetes. Por defecto, se instalará la

última versión incluida en el repositorio, pero en algunas circunstancias es conveniente, o incluso requerido, instalar versiones anteriores.

R

La gran mayoría de los scripts de R aquí presentados son Rscripts, y se corren desde la terminal de Linux y sin entrar en R. Es decir, ya están automatizados; `squeezeR` requiere tener instalado el paquete `seqinr` (Charif y Lobry, 2007).

Es momento de volver a mencionar los entornos de conda, ya que también pueden utilizarse versiones anteriores de R en caso de que sea necesario.

Secuencias de referencia

RefSeq, acrónimo en inglés de «secuencia de referencia», es una base de datos secundaria alojada en el NCBI (del inglés, *National Center for Biotechnology Information*); pretende ser una sub-base de datos no-redundante y con cierto nivel de curado. Sus registros pueden corresponder a secuencias subidas por cualquier miembro del INSDC (sigla en inglés de *International Nucleotide Sequence Database Collaboration*); esta colaboración internacional es una iniciativa que opera entre los tres principales repositorios de secuencias: DDBJ (Japón), ENA (Europa) y NCBI (Estados Unidos). Se actualiza con una frecuencia bimestral en los meses impares, siendo la versión actual la 209 (1 de noviembre de 2021; disponible desde ncbi.nlm.nih.gov/refseq y/o ftp.ncbi.nlm.nih.gov/refseq).

La base RefSeq, si bien es de las bases con mayor nivel de curado, presenta un alto número de secuencias provisorias que solamente han sido anotadas por algoritmos sin supervisión de un usuario. Los registros RefSeq pueden asignarse a tres categorías excluyentes, ordenadas de acuerdo a un nivel de curado creciente: *provisional* («provisorio»), *validated* («validado») o *reviewed* («revisado»).

En general, para genomas corresponde con la secuencia más larga disponible para un organismo o virus determinado. Al menos esto es así cuando son ingresadas de

manera automática (Pruitt et al., 2020), pero existen excepciones a esta generalización.

Un identificador taxonómico (TaxID, del inglés *Taxonomy ID*) es un número que, como su nombre lo indica, identifica a cada taxón (en un nivel o rango taxonómico). Estos se asignan secuencialmente a medida que se añaden a la base de datos *Taxonomy* del NCBI (disponible desde ncbi.nlm.nih.gov/taxonomy).

En la **Figura 4** se muestra el crecimiento de TaxIDs con genomas totalmente secuenciados con cada nuevo lanzamiento (*release*); 4195 genomas virales conformaron el set de datos de mi tesina de grado (Simón, 2015).

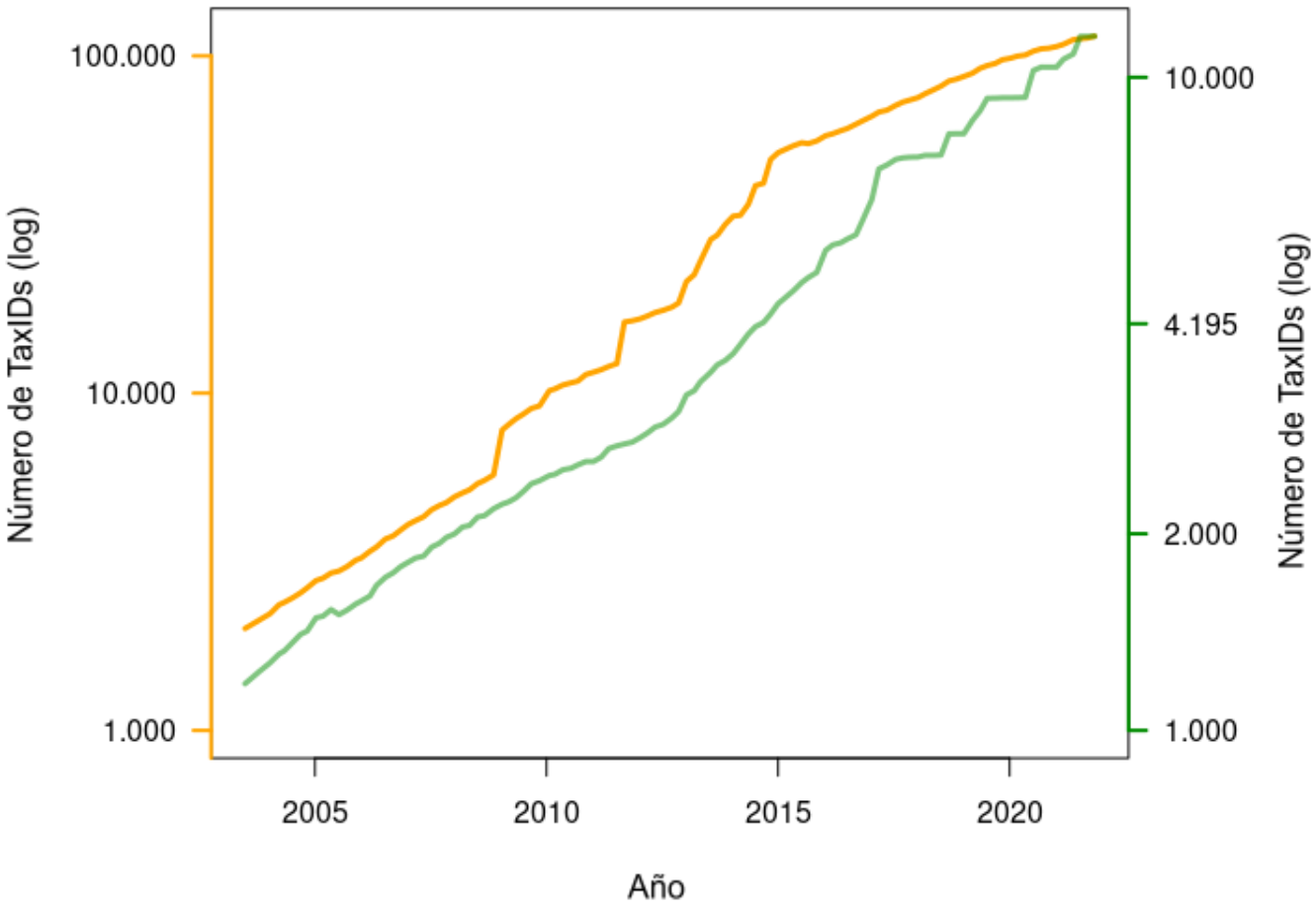



**Figura 4.** Crecimiento en el tiempo del número de TaxIDs (escala log) en total (***complete***, eje izquierdo) y de virus (***viral***, eje derecho), incluidos en cada «lanzamiento» (*release*) de la base de datos RefSeq.

Archivos GenBank

Un archivo gbk es un formato de texto plano donde es almacenada la información del genoma; además de la secuencia correspondiente contiene metadatos como información general de la misma (e.g., tipo, tamaño, origen de la muestra, publicaciones, etc.), además de la anotación correspondiente (e.g., genes, regiones codificantes, cambios de marco de lectura, sitios de edición, etc.).

Al ser un formato de texto plano (ver **Anexo A**), este puede ser «parseado». Para esto uno puede desarrollar un script que recorra el archivo. Una alternativa es utilizar funciones ya diseñadas para este fin. Biopython posee la función `parse` implementada en el módulo `SeqIO` que permite esto mismo: recorrer este archivo y extraer información de interés para el usuario.

Cambios en los gbk de algunos genomas, por cómo está codificada la información del ensamblaje, pueden hacer que el uso de este script termine de manera abrupta, sin recorrer el gbk de manera completa; si se presenta este error, es probable que deba actualizarse Biopython.

Un ítem en los archivos gbk que suele presentar errores es la topología del genoma; es decir, si un genoma es circular o lineal. Dado entonces que la topología no es un dato consistente (no siempre está bien anotado), se agruparon en un solo archivo las regiones no codificantes con extensión .non. Es probable que sea sencillo determinar o estimar este dato, pero hacerlo escapaba a los intereses de esta tesis. Por esa razón, se unificó lo no codificante sin importar si es intergénico o flanqueante (también se adicionaron los intrones, en aquellos virus con regiones intrónicas anotadas).

El script gbfx.py (GenBank *feature extractor*), escrito en Python 3.6, toma como *input* un archivo en formato GenBank (extensión *gb* o *gbk*) y extrae secuencias a partir de las anotaciones que se encuentran en los gbk como «atributos» (*features*). Además de presentar la información contenida en las secuencias, aportan información sobre el origen de las secuencias depositadas tanto a nivel biológico y técnico.

A partir de cada virus con genoma completo, se dispone todo el genoma completo en un archivo fasta (con extensión gnm); tener en cuenta que un virus segmentado

tendrá tantos gbk como segmentos. Además del genoma, este script permite extraer distintas porciones genómicas de interés. Siempre que existan, se extraen los 5' y 3' flanqueantes no codificantes (relevante solamente en virus con genoma lineal), regiones codificantes, intrones y regiones intergénicas, siempre que estas regiones existan (**Tabla 1**).

**Tabla 1.** Regiones genómicas extraídas con gbfx.py de los archivos GenBank, extensión asignada a cada uno y observaciones, si corresponden.

| Región genómica | Extensión | Observaciones |
|---|---|---|
| genoma completo | gnm | |
| regiones intergénicas | int | |
| región flanqueante 3’ | fr3 | genomas lineales |
| regiones codificantes | cds | |
| intrones | spl | genes con intrones |
| regiones no codificantes | non | fr5 U int U fr3 U spl |
| cds traducidas in silico | pep | |
| código genético | tbl | si no se explicita, se asigna 1 |

El código genético es universal pero presenta algunas variaciones naturales (Elzanowski y Ostell, 2019). Por defecto, el código genético estándar es el 1, correspondiente a la casi totalidad de los genomas nucleares de eucariotas. El gbk incluye un atributo en caso de utilizar un código genético alternativo (`trans_table=`).

La limitante en este caso es que tan bien anotados estén estos genomas. Es decir, el script depende de aquellas regiones y atributos descritos en el gbk. Si no figura, no existirá para nosotros. Muchos de estos genomas están anotados de manera automática y siempre puede ocurrir que se hayan omitido regiones funcionales. Puede ocurrir también que algunos atributos no existan y sean artefactos. Cuando el volumen de datos es grande, la expectativa es que la señal se sobreponga al ruido.

Con las salidas del archivo gbk, se armaron tres archivos fasta por especie viral:

**(i)** archivo con extensión .gnm (de *genome*) con todo el genoma (concatenando aquellos genomas de virus segmentados).

**(ii)** archivo .cds (de *coding sequences*) concatenando regiones codificantes.

**(iii)** archivo .non (de *non-coding*) concatenando regiones no codificantes.

La base de datos RefSeq suele incluir un único representante por TaxID, aunque hay algunas excepciones. Un identificador único más apropiado en este caso es el código de entrada del ensamblado o accesión (*accession*). Este sí es único y permite además cruzar datos entre GenBank y RefSeq (porque la secuencia numérica de estas entradas coinciden), además de otras bases de datos.

Otras bases de datos

Se podría analizar los datos composicionales junto con información obtenida de los archivos gbk; es decir, utilizando solamente la información presentada en estos (e.g., material genético, taxonomía, origen). Pero en las etapas posteriores de análisis, contar con información de calidad pasa a ser una condición *sine qua non*. Cuanta mayor información biológica se disponga, mayor será el aprovechamiento completo de los datos.

Entonces, otro componente fundamental de esta tesis es la *metadata* asociada a las secuencias analizadas. En paralelo a la obtención de las secuencias, se consultaron otras bases de datos biológicas y/o bibliográficas con la finalidad de complementar la información de las secuencias. Se utilizaron numerosos recursos del NCBI como Virus, Genome, Taxonomy y PubMed. Otras de las bases de datos consultadas fueron ICTV, ViralZone, Virus-Host y CoCoPUTs, con el fin de:

**(i)** Completar y actualizar la taxonomía, para que se presente acorde al nuevo sistema jerárquico de ICTV (ICTV Executive Committee, 2020).

**(ii)** Sobre las familias virales aceptadas (además de hacerlo también para las familias aún no aceptadas y para los géneros aún no asignados en familias).

**(iii)** Se presenta el grupo de Baltimore, su material genético, si son de doble o simple hebra, su polaridad y si son retro-transcriptos o no.

**(iv)** También se incluyen aquellos géneros o familias donde se agrupan entidades subvirales (i.e.; satélites o viroides) y sus virus *helper*.

Ante aquellas situaciones que generaron vacíos en la información (ausente o parcial), o, peor aún, contradictoria, se requirió a publicaciones sobre dichas especies virales (consultando las publicaciones referidas en los gbk y/o recurriendo a PubMed).

En el **Anexo B** se presenta una lista de las bases de datos consultadas y los sitios web donde se accede. Si además son accesibles mediante un sitio FTP (*file transfer protocol*), también se presenta.

**Análisis composicionales**

Al recorrer los inicios del estudio de los ácidos nucleicos y de la biología molecular, le dedicamos algunos párrafos a los primeros estudios composicionales, que utilizaban en aquellos tiempos herramientas cromatográficas (Chargaff et al., 1949; Markham y Smith, 1949a; 1949b).

También vimos que inmediatamente se realizaron los primeros estudios composicionales en virus (Markham y Smith, 1950; Smith y Wyatt, 1951).

Los genomas exhiben una gran diversidad composicional; esta se manifiesta en el uso relativo de sus nucleótidos. Una medida de esta diversidad composicional es el contenido de G y de C (G+C):

$$G+C\% = \frac{G+C}{A+G+T(U)+C} * 100$$

Mucho de lo que sabemos ha sido descrito en procariotas o en eucariotas, existiendo pocos estudios sistemáticos en virus. La mayor parte de los trabajos han sido realizados en bacterias. Fue posible conocer que entre bacterias emparentadas, éstas pueden presentar cambios considerables en su ADN, pero en mucho menor grado en sus ARNs (Belozersky y Spirin, 1958).

Además, Belozersky y Spirin (1958) demostraron que el material genético de las bacterias presenta un fuerte sesgo mutacional y que se observan especies o grupos con gran variación, que va desde valores A+T extremadamente altos a valores de G+C también muy altos. El contenido de G+C varía en los distintos organismos desde un mínimo de 13% hasta un máximo de 75% (Sueoka, 1962; Brocchieri 2014).

En eucariotas, la gran mayoría de los estudios se centran en humanos y/o en vertebrados, principalmente mamíferos y aves, existiendo pocos estudios realizados en eucariotas no-vertebrados.

Los análisis composicionales, por no depender necesariamente de alineamientos de secuencias, brindan la posibilidad de buscar patrones genómicos globales en toda la diversidad viral conocida. Esto es relevante porque en los genomas virales no

existen secuencias conservadas universalmente dado que no comparten un ancestro común, a diferencia de lo que ocurre en el árbol de la vida, donde existen grupos de secuencias que se conservan en todos los genomas celulares, sirviendo de esta manera para estudios taxonómicos y evolutivos.

Comenzar a disponer de más genomas completos permitió realizar búsquedas de patrones composicionales tanto inter- como intra-genomas. De ese modo, se empezó a vislumbrar que si bien en la globalidad de un genoma existe una composición determinada, y que esta era muy similar entre organismos emparentados, dentro de un genoma existe también heterogeneidad.

Lo que se presenta para realizar los análisis composicionales es `squeezeR`, un *Rscript* que se corre desde la línea de comandos, utilizando las funciones `count` y `rho` de `seqinr`. Con estas funciones es posible realizar todos los análisis composicionales pero de manera limitada. Dicho script exporta en formato tabla una multitud de análisis para un archivo fasta dado. Por cómo se diseñó el abordaje de esta base de datos, el script asume, en caso de un archivo multifasta que todas las secuencias corresponden a la misma especie y las agrupará para realizar cálculos composicionales globales. Se comporta de manera diferente según se indique si la secuencia es codificante o no.

Los análisis composicionales se realizaron para el genoma (.gnm), regiones no codificantes (.non) y regiones codificantes (.cds). Debido a que se pretendía obtener un dato global de cada genoma viral, se agruparon aquellas secuencias distribuidas en varios archivos.

Para los virus segmentados, los archivos correspondientes a los segmentos se unificaron en un solo archivo (.gnm), concatenando las secuencias adicionando un nucleótido indefinido (N); dicho detalle es importante a la hora de los análisis composicionales y se aclarará al mencionar los dinucleótidos.

Para genomas segmentados como no segmentados, se hace lo mismo que lo descrito para los archivos gnm pero para cada archivo con secuencias no codificantes (.non). También se adiciona N cada vez que se concatenan secuencias.

Para los archivos con secuencias codificantes (.cds), también se concatenan estos archivos. En esta oportunidad, al concatenar se adiciona NNN (i.e., 3 veces N), para

no alterar el marco de lectura. Esto lo hacemos porque los virus tienen, en promedio, muy pocos genes; a diferencia de lo que ocurre en genomas celulares que cuentan con cientos hasta decenas de miles. Trabajaremos entonces con una secuencia por especie viral.

Cada N adicionado artificialmente se deja registrado en los encabezados de los archivos fasta, para poder luego restarlos al tamaño calculado de los genomas (y también de las regiones codificantes y no codificantes).

El sesgo de dinucleótidos,  es calculado como la frecuencia observada sobre la frecuencia esperada:

$$\rho XY = \frac{fXY}{fX*fY}$$

Este se presenta tanto para la hebra representada en la secuencia (`rho`), como en ambas hebras, agregando el reverso complementario (`kar`). Si se indica un valor de oligonucleótido mayor a 2, se determinan también para k nucleótidos entre 2 y el valor indicado.

Si extendemos el concepto de dinucleótidos a k nucleótidos, también tomados de manera solapada, el número de posibilidades en simple hebra es 4^k. Para k = 3 (trinucleótidos), son 64 posibles (4^3). Además se disponibiliza `revComplement3.r`, que agrupa aquellos tripletes que son opuestos complementarios (e.g., AAA y TTT, AAC y GTT, AAG y CTT, AAT y ATT, ACA y TGT, etc). Para k = 3, los tripletes no redundantes en doble hebra son 32.

En el caso de secuencias codificantes, indicando que así lo son, se extraerán, además, los valores según la posición en el marco de lectura: para bases, 1-2-3, para dinucleótidos 12-23-31. Para k ≥ 6, se extraerá el conteo y la frecuencia de pares de codones.

Correr `squeezeR` desde la línea de comandos independiza al usuario de entrar a R. Genera una nueva subcarpeta donde se irán guardando todas las tablas obtenidas. Además, se brinda un registro de eventos (.log) que permite seguir su funcionamiento.

En cuanto a las secuencias codificantes, estas presentan sesgos en el uso de codones sinónimos. El uso de codones es la frecuencia de cada triplete en el marco de lectura determinado por el codón de inicio. Los virus, como endoparásitos obligados, deben utilizar la maquinaria traduccional de la célula hospedera en la traducción; de esta manera, los virus están expuestos a los sesgos propios del hospedero.

Otra medida es el uso relativo de codones sinónimos (RSCU por sus siglas en inglés, *relative synonymous codon usage*). Este valor se construye como la relación entre la frecuencia observada de los codones y la frecuencia esperada si todos los codones sinónimos se utilizan por igual.

Fuera de R existe CodonW, que permite realizar análisis de uso de codones. Permite correrlo en consola de comandos y ofrece distintas salidas de uso de codones y análisis de correspondencias como análisis multivariados.

Dado que el paso previo (el análisis de firmas genómicas, con `squeezeR`) se realizaba en R, se presenta una alternativa para correr utilizando este lenguaje: `ufo.R`(k). Un "*fork*" es en ingeniería de software una derivación independiente a partir de un proyecto preexistente; `ufo.R` consiste en un *shallow fork* de la función `uco` de `seqinr`, que además de brindar el conteo y la frecuencia de codones, permite calcular el RSCU asignando el código genético, fundamental para aquellos genomas que utilizan códigos genéticos alternativos (es decir, que difieran del código estándar o universal). Cabe destacar que la función original de `seqinr`, `uco`, no lo permite.

Una perspectiva es ampliar estos análisis a la secuencia aminoacídica, ya sea con alguna de las funciones de `seqinr` u otros paquetes (e.g., `protr`; disponible desde CRAN.R-project.org/package=protr).

**Análisis estadísticos**

Como medidas de tendencia central se obtienen la media aritmética (M), la media geométrica (GM), la media armónica (HM) y la mediana (Mdn). Estas diferentes medidas de tendencia central pueden ser informativas en el caso de que logren describir mejor la distribución de alguna de las variables. Por ejemplo, la mediana representa el valor de posición central de la variable en un conjunto de datos ordenados. Si el conjunto de datos tiene un número impar de datos, la mediana corresponde al valor elemento que parte la distribución en dos series de igual número de elementos; si por el contrario la serie tiene un número par de datos, la mediana corresponde al valor promedio entre los dos valores centrales.

Al comparar la media y la mediana de una variable que presenta distribución normal, ambos valores serán bastante próximos entre sí (**Figura 5A**). Lo mismo puede ocurrir aún en variables que no sean unimodales (**Figura 5B**).

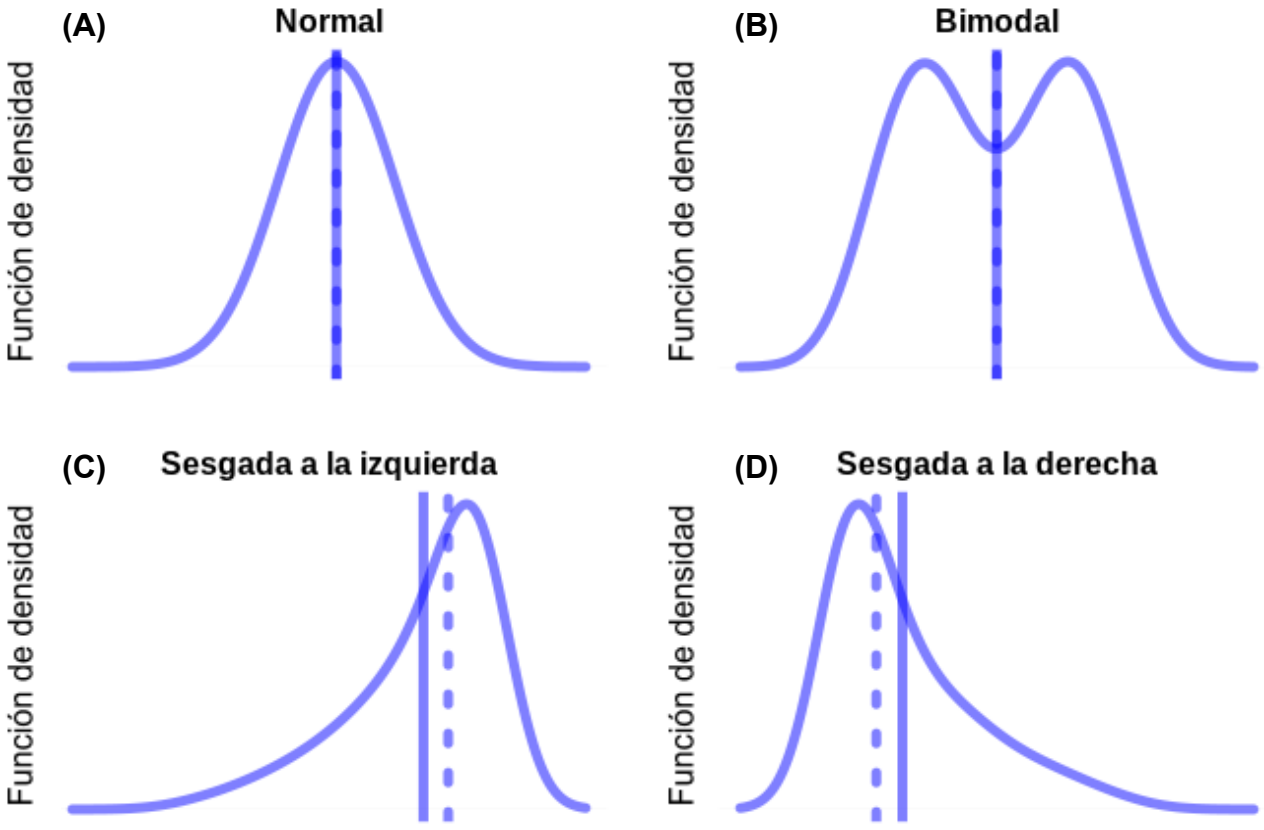



**Figura 5.** Representación de las funciones de densidad de distribuciones normal **(A)**, bimodal **(B)** y sesgadas a la izquierda **(C)** y a la derecha **(D)**; líneas verticales representan la media (**continua**) y la mediana (punteada).

Pero cuando una variable presenta una distribución asimétrica, la media y la mediana tenderán a separarse. A modo de ejemplo, con una distribución que está sesgada a la derecha (**Figura 5D**), la media tendrá un valor más alto que la mediana; una distribución sesgada hacia la derecha tiene un mayor número de valores hacia el extremo inferior y un menor número de valores hacia el extremo superior.

Se calculan, además, diferentes medidas de dispersión y de posición como desviación estándar (SD), desviación media absoluta (MAD), rango (i.e., mínimo [Min] y máximo [Max]), y rango intercuartílico (IQR).

Todos estas medidas se obtienen con R. Se automatizó la obtención de los mismos con la función `univaR`, que debe cargarse desde el script `univa.R`; esta función genera una lista con todos los estadísticos descriptivos mencionados previamente. Además realiza otros test de interés (**Tabla 2**), para los cuales requiere los paquetes `diptest` (Maechler, 2021) y `moments` (Komsta y Novomestky, 2015).

**Tabla 2.** Medidas obtenidas con `univaR` y su notación, por categoría.

| Categoría | Medida | Notación |
|---|---|---|
| Medidas de tendencia central | Media aritmética o promedio | M (*mean*) |
| | Media geométrica | GM (*geometric mean*) |
| | Media armónica | HM (*harmonic mean*) |
| | Mediana | Mdn |
| Dispersión | Rango (mínimo, máximo) | Min, Max |
| | Desviación estandar | SD |
| | Desviación media absoluta | MAD |
| | Rango intercuartílico | IQR |
| Otros | Simetría | |
| | Sesgo | |
| | Normalidad | |
| | Unimodalidad | |

Para determinar valores de coeficiente de correlación entre variables se utiliza la función `cor.test` de R, que permite calcular una correlación con los métodos de Pearson, de Spearman y de Kendall.

El coeficiente de correlación de Pearson, *r*, es el estadístico más utilizado para medir la correlación lineal entre dos variables cuantitativas. Este coeficiente (*r*) puede no ser el adecuado en el caso de variables que posean muchos valores atípicos o que se aparten de la distribución normal. En estos casos es más correcto recurrir a coeficientes de correlación no paramétricos. Tanto el de Spearman como el de Kendall son casos particulares de un coeficiente de correlación más general.

El coeficiente de correlación de Spearman, $\rho$ (rho), es una medida no paramétrica de correlación de rango. Evalúa lo bien que se puede describir la relación entre dos variables mediante una función monótona. La correlación de Spearman entre dos variables es igual a la correlación de Pearson entre los valores de rango de esas dos variables; mientras que la correlación de Pearson evalúa las relaciones lineales, la correlación de Spearman evalúa relaciones monótonas (que pueden ser lineales o no lineales). En el cálculo de Spearman el peso de cada par de rangos es proporcional a la distancia entre dicho par.

El coeficiente de correlación de Kendall, $\tau$ (tau), se diferencia del de Spearman, en que el peso concedido a un par de rangos es pesado por igual (siendo independiente de su distancia). Sustituir los valores de una variable por sus rangos genera que se pierda información de los datos originales, pero estos métodos no paramétricos son apropiados tanto para variables continuas como para variables discretas. La elección de un determinado coeficiente de correlación debe basarse en el tipo de variable que se va a analizar, la distribución de los datos y el tipo de relación prevista. Esta evaluación a priori no siempre es posible.

Como comentario final sobre los coeficientes de correlación y sus significancias: «*Cum hoc ergo procter hoc*» («correlación no implica causalidad», Omar Defeo *dixit*). Estos coeficientes (y sus significancias) se utilizarán de manera cautelosa y con la finalidad de conocer cómo describen las distintas relaciones y comparar sus desempeños.

## Resultados no publicados

### Análisis multivariados

Las propiedades composicionales obtenidas se estudiaron con estadística multivariada, utilizando para este fin la técnica de análisis de componentes principales (PCA, de *principal component analysis*). Un PCA es un procedimiento estadístico para reducir la multidimensionalidad de un conjunto de datos. Estudiamos con este abordaje las propiedades composicionales de genomas virales a nivel de sus frecuencias de dinucleótidos, de codones y de aminoácidos.

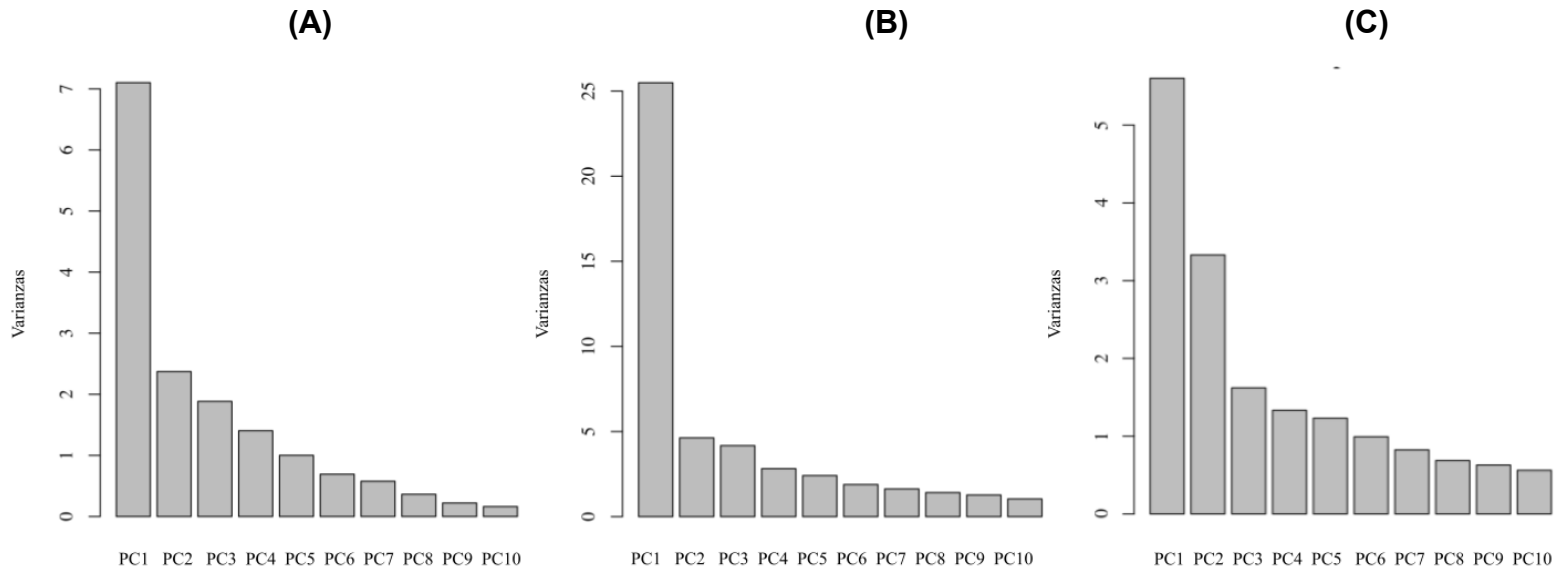



**Figura 6.** Varianzas de los primeros diez componentes principales del PCA para la frecuencia de dinucleótidos **(A)**, codones **(B)** y aminoácidos **(C)**.

Frecuencia de dinucleótidos

Empezando con los dinucleótidos, utilizando solamente regiones codificantes (.cds), en la **Figura 7** se muestra el peso de las variables originales en el plano formado por las dimensiones 1 y 2, PC1 y PC2 respectivamente, que en conjunto explican un 59% de la varianza de los datos originales.

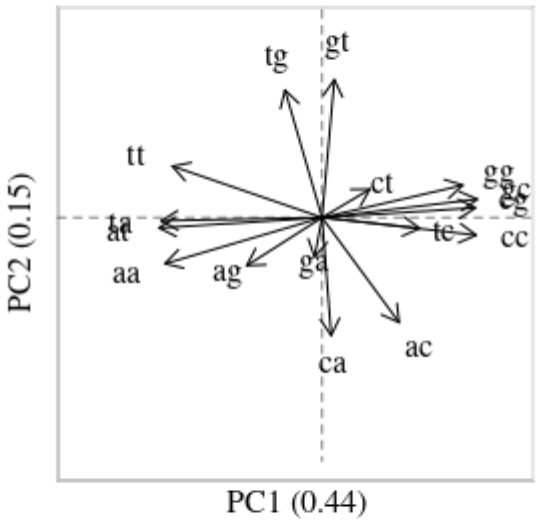



**Figura 7.** Peso de las variables originales sobre las dimensiones 1 y 2, PC1 y PC2 respectivamente, del PCA para la frecuencia de dinucleótidos con todos los genomas analizados (N = 9443).

La variable con más peso en el PC1 es el contenido de G+C (**Figura 7**), evidenciado en la disposición de los dinucleótidos compuestos de A y/o T (U para los virus de ARN) hacia los valores negativos del PC1, y de los dinucleótidos compuestos de G y/o C hacia los valores positivos. PC2 separa AC/CA (-) y GT/TG (+); es interesante notar que estos pares de dinucleótidos son reversos complementarios: AC con GT y CA con TG.

Al observar la distribución de los virus en el plano destaca el patrón presentado por el grupo I (dsDNA; **Figura 8**, panel izquierdo), donde los virus se ubican formando un gradiente con el PC1, que como se dijo anteriormente corresponde al G+C.

Es interesante que el otro grupo de virus con genoma de ADN, el grupo VII (dsDNA-RT; **Figura 8**, panel derecho), «exploran» un sector del plano diferente a los del grupo I. El apartamiento de los virus del grupo VII de la composición «esperada» para los virus dsDNA puede deberse a sesgos propios de la transcriptasa inversa. La transcriptasa inversa del grupo VII de Baltimore presenta homología con las de los virus del grupo VI (Koonin et al., 2020), el otro grupo de virus retro-transcriptos.

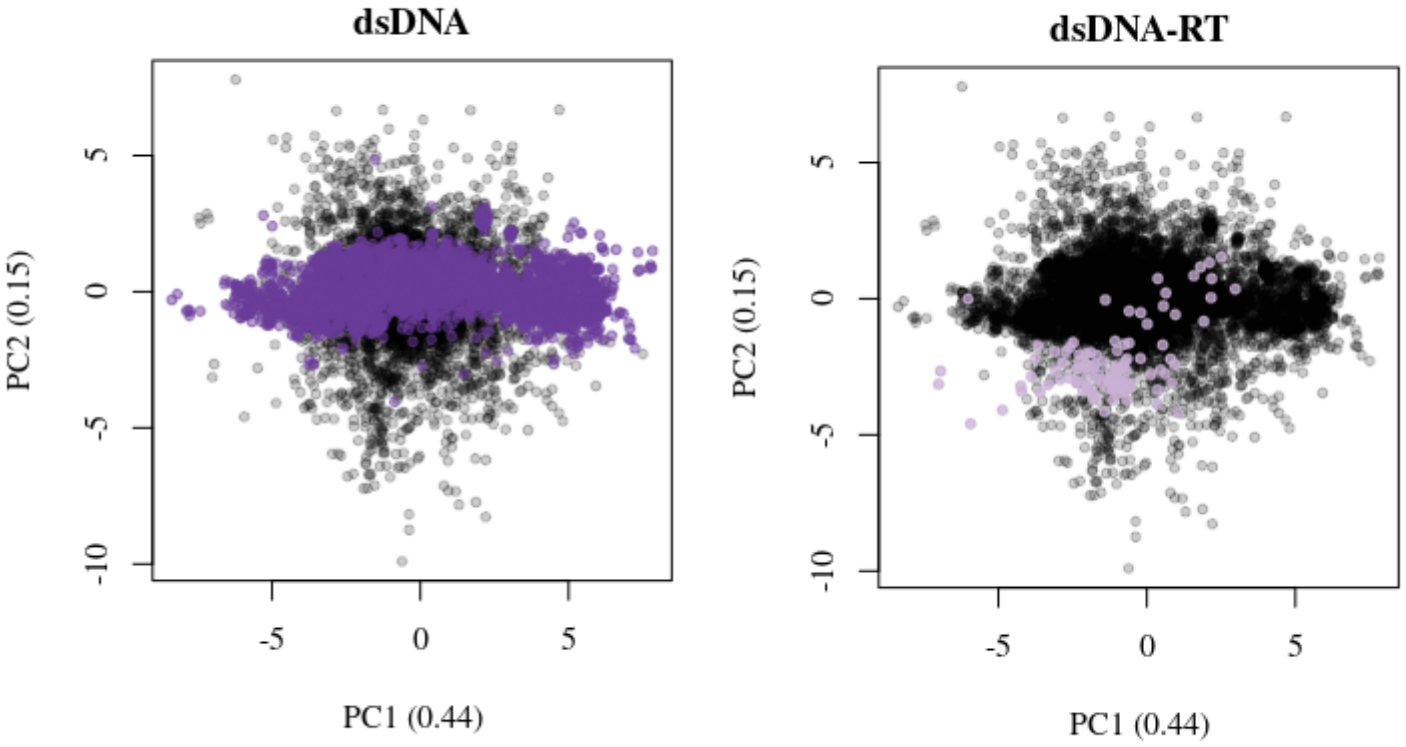



**Figura 8.** Virus del grupo I (**dsDNA**, panel izquierdo) y grupo VII (**dsDNA-RT**, panel derecho) de Baltimore en el plano formado por las dos primeras dimensiones (PC1 y PC2) del PCA para la frecuencia de dinucleótidos (N = 9443).

Frecuencia de codones

Con respecto a los codones, en la **Figura 9** se muestra el peso de variables originales en el plano formado por PC1 y PC2. Estos ejes acumulan un 47% de la varianza de los datos originales.

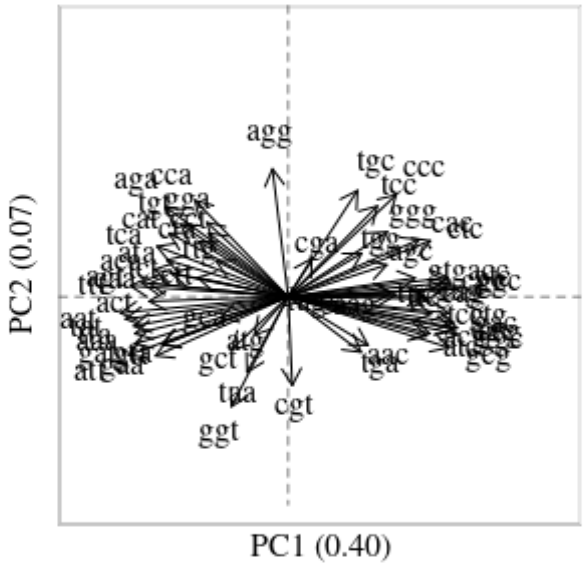



**Figura 9.** Peso de las variables originales sobre PC1 y PC2 del PCA para la frecuencia de codones con todos los genomas analizados (N = 9994).

De forma similar a lo observado para dinucleótidos (**Figura 7**), el PC1 para codones es pautado por el contenido de G+C (**Figura 9**); tripletes ricos en A+T hacia valores negativos y tripletes ricos en G+C hacia los positivos. El PC2 parece relacionarse con el uso de codones del aminoácido arginina; AGG (y también AGA) hacia valores positivos del eje y CGT (o CGU) hacia los negativos; cabe destacar que el porcentaje de varianza explicado por este eje es relativamente bajo (7%).

El uso de codones sinónimos para el aminoácido arginina (es decir, AGR vs CGN) como fuente de variabilidad ha sido descrito en diversos sistemas (por ejemplo: Lynn et al., 2002; Nakamura y Sugiura, 2011; Novoa et al., 2019).

En la **Figura 10** se evidencian dos acúmulos de puntos. Principalmente, corresponden a virus de animales (panel izquierdo) y bacteriófagos (panel derecho), con pocas excepciones en uno y otro caso. Los fagos de árqueas aparecen coincidiendo con una u otra nube (ver **Apéndice B1**), aunque era de esperar un mayor solapamiento con los bacteriófagos (por su calidad de virus de procariotas). Coinciden con los virus animales el resto de los virus de eucariotas (i.e., hongos, plantas y protistas), por lo que esta gran nube corresponde a virus de eucariotas y a algunos virus de árqueas.

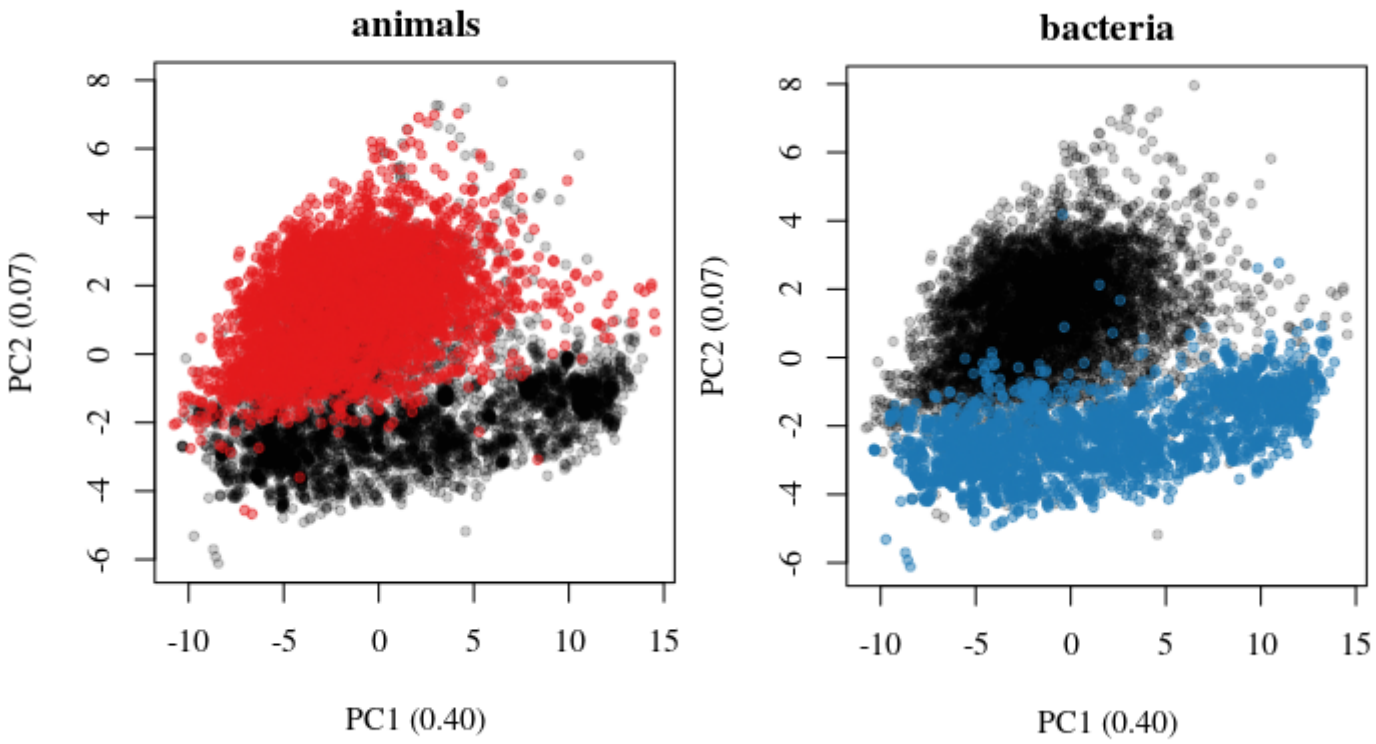



**Figura 10.** Virus de animales (**animals**, panel izquierdo) y de bacterias (**bacteria,** panel derecho) en el plano formado por las dos primeras dimensiones (PC1 y PC2) del PCA para la frecuencia de codones (N = 9443).

Frecuencia de aminoácidos

Al analizar los aminoácidos, PC1 y PC2 acumulan un 45% de la varianza de los datos originales. La **Figura 11** muestra el peso de variables originales en el plano formado por PC1 y PC2.

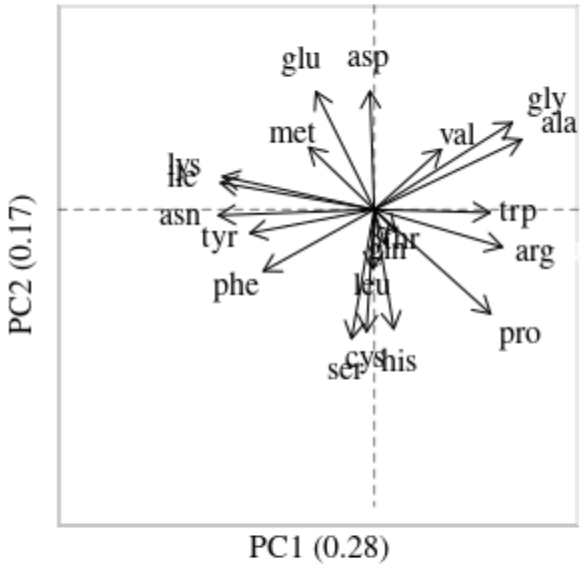
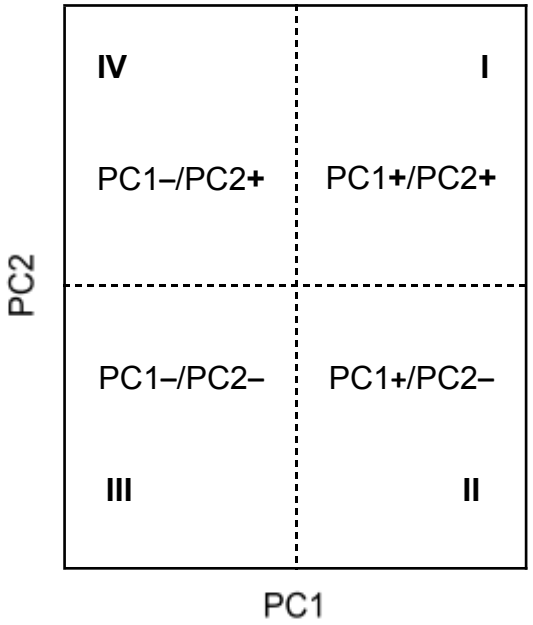



**Figura 11.** Peso de las variables originales sobre PC1 y PC2 del PCA para la frecuencia de aminoácidos con todos los genomas analizados (N = 9994).

La variable con más peso en el PC1 vuelve a ser el contenido de G+C (**Figura 11**), evidenciado en la disposición hacia los valores negativos del PC1 de aminoácidos codificados por codones ricos en A y/o T/U, como asparagina (asn), isoleucina (ile), lisina (lys), tirosina (tyr) y fenilalanina (phe). Hacia los valores positivos del PC1, hállanse aminoácidos ricos en G y/o C como alanina (ala), glicina (gly), arginina (arg), triptófano (trp) y prolina (pro). Respecto al PC2, se destacan hacia los valores positivos los aminoácidos aspártico (asp) y glutámico (glu); ambos aminoácidos están codificados por la cuaterna de tripletes GAN (siendo N cualquier nucleótido). Hacia los valores negativos se disponen serina (ser), cisteína (cys) e histidina (his).

Para describir la disposición de los virus en el plano del PCA para aminoácidos y, tomando como referencia al origen de coordenadas, dividiremos dicho plano en cuadrantes (**Tabla 3**).

**Tabla 3.** Esquema de los cuadrantes en los que se divide el plano PC1-PC2, en sentido horario y comenzando de valores positivos para ambos ejes: «arriba a la derecha» (**I**), «abajo a la derecha» (**II**), «abajo a la izquierda» (**III**) y «arriba a la izquierda» (**IV**).




Los bacteriófagos, al igual que para dinucleótidos y para codones, se extienden a lo largo del PC1, dominando los cuadrantes **I** y **IV**. Los virus de árqueas presentan cierto gradiente en dirección cuadrante **I**-**III** (o **III**-**I**), aunque en general su disposición solapa con el resto de los fagos; marcadamente más que lo que ocurría con los codones (ver **Apéndice C1**).

Los virus de eucariotas predominan en los cuadrantes **II** y **III**. En la **Figura 12** se muestra como los grupos que más área cubren del plano son los virus de animales y de plantas. Estos últimos, sin embargo, se proyectan además hacia el cuadrante IV (**Figura 12**, panel derecho). Virus de hongos y de protistas se suman a la nube de los virus de animales, tal como ocurría para los codones.

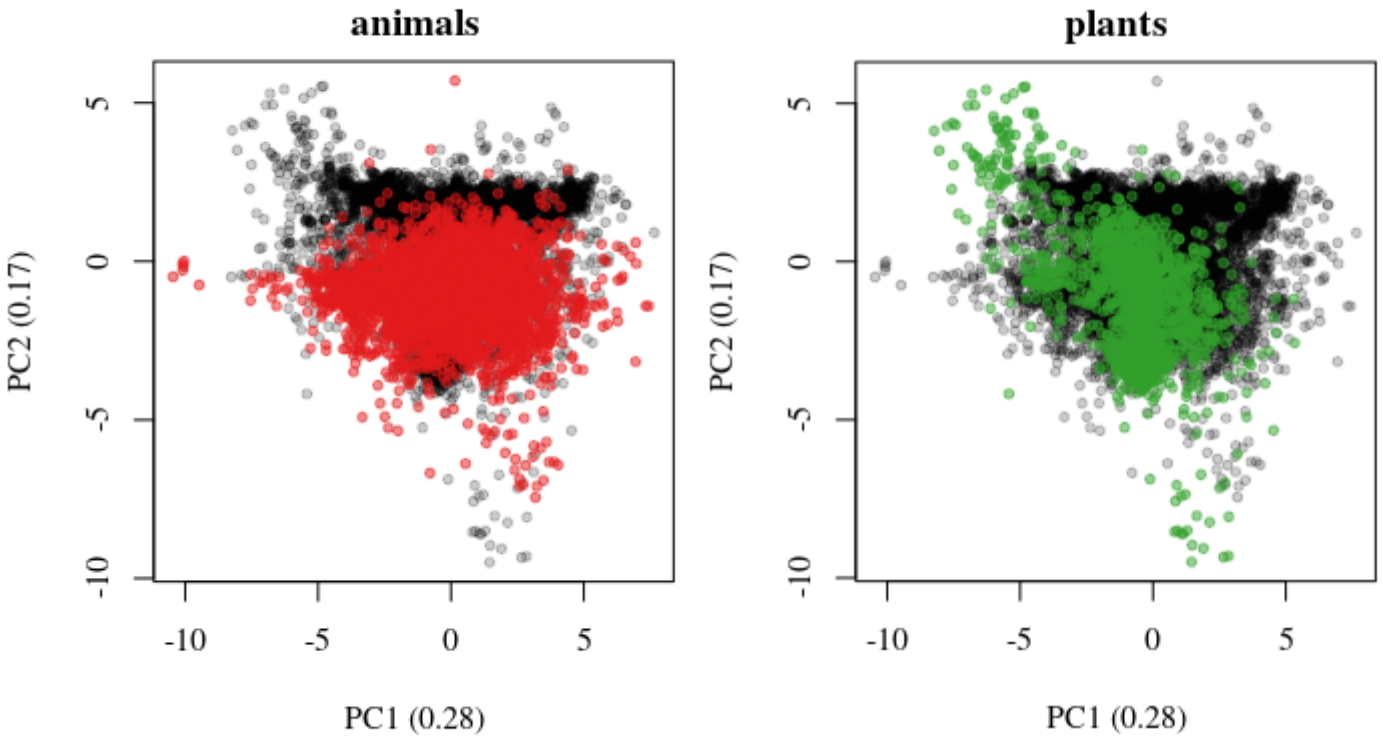



**Figura 12.** Virus de animales (**animals**, panel izquierdo) y de plantas (**plants**, panel derecho) en el plano formado por las dos primeras dimensiones (PC1 y PC2) del PCA para la frecuencia de codones (N = 9443).

En los **Apéndices A-C** se presentan todas las categorías por grupo de hospedero y por grupo de Baltimore.

**Sesgo mutacional**

Los virus comparten una importante característica con los organismos procariotas; al igual que arqueas y bacterias, sus genomas poseen una alta densidad génica, estando sus genes distribuidos de forma compacta. Esto implica que sus genomas son en gran parte codificantes o con funciones regulatorias.

Para evidenciar el efecto del sesgo mutacional, principalmente sobre el contenido de G+C, se estudió cómo varía éste en relación al G+C no codificante. Para evitar artefactos provocados por regiones no codificantes poco representativas, se consideraron solamente para este análisis virus cuyos genomas tuvieran al menos 100 bases no codificantes (fueran estas continuas o no).

La correlación de Pearson entre el G+C no codificante y el G+C genómico es 0,88, con un $R^2$ de 0,77, con una pendiente de regresión lineal entre variables de 0,80. Además, todos los grupos de Baltimore presentan correlaciones positivas (**Tabla 4**). Independientemente de su material genético y de su polaridad, todos los grupos presentan correlación entre el G+C no codificante y el G+C genómico o global. Aunque heterogéneas en cuanto a sus pendientes, todos estas correlaciones son entre moderadas y altas. La relación para todos los virus puede observarse en la **Figura 13**.

**Tabla 4.** Correlaciones de Pearson (**r**), pendiente de la recta de regresión lineal (**m**) y coeficiente de determinación ajustado ($\mathbf{R^2_a}$) para todos los virus analizados y por grupo de Baltimore.

| | r | m | $R^2_a$ |
|---|---|---|---|
| todos | 0,88 | 0,80 | 0,77 |
| dsDNA | 0,97 | 0,96 | 0,94 |
| ssDNA | 0,73 | 0,46 | 0,53 |
| dsRNA | 0,66 | 0,56 | 0,44 |
| +ssRNA | 0,64 | 0,50 | 0,41 |
| -ssRNA | 0,63 | 0,38 | 0,40 |
| +ssRNA-RT | 0,82 | 0,86 | 0,67 |
| dsDNA-RT | 0,73 | 1,06 | 0,51 |

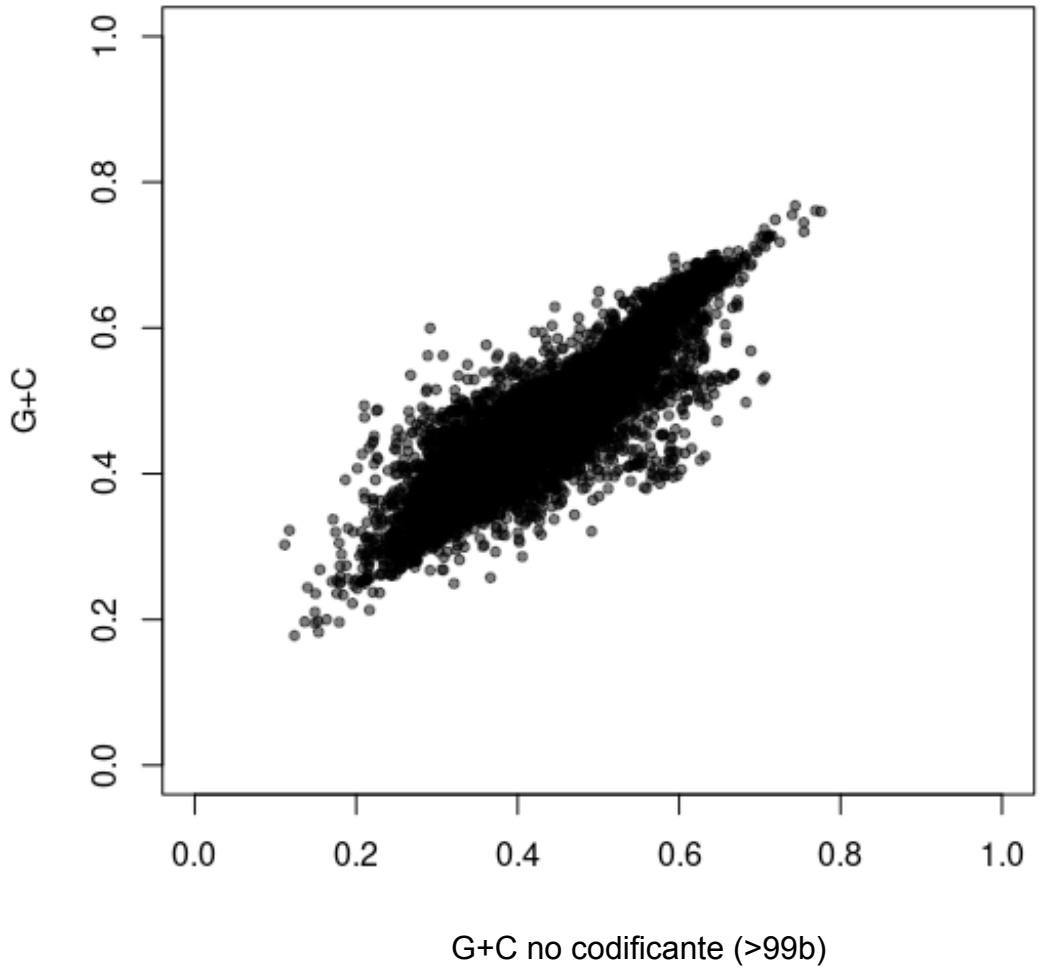



**Figura 13.** Dispersión del G+C no codificante (con largos de 100 o más bases) y el G+C genómico para todos los genomas virales analizados (N = 9443).

El valor más bajo de pendiente de los diferentes grupos de Baltimore es m = 0,38, para los virus del grupo V (-ssRNA; ver **Tabla 4**). De entre los miembros de este grupo (**Figura 14**), podemos observar un grupo de puntos que se apartan de toda la nube. Y es este grupo el responsable de que el valor de la pendiente sea tan bajo, de tal manera que si repetimos el modelo para los -ssRNA sin los *Arenaviridae*, el valor de la pendiente es m = 0.80.

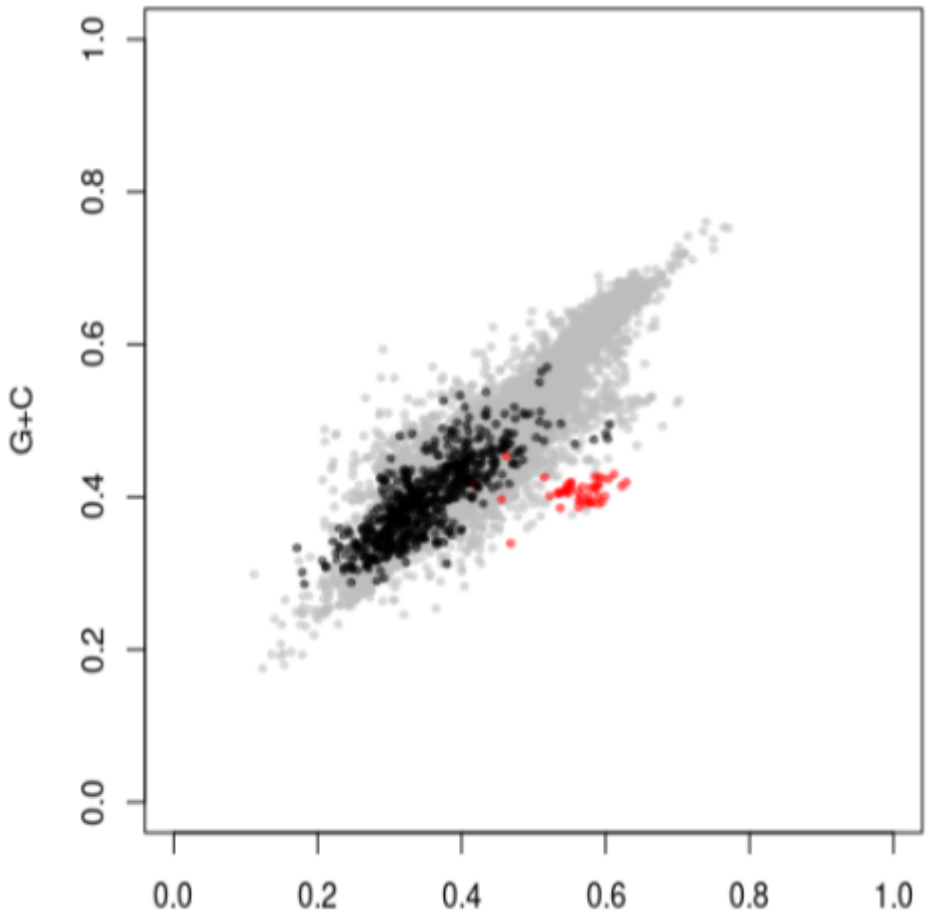



**Figura 14.** Dispersión del G+C no codificante (con largos de 100 o más bases) y el G+C genómico para los virus con genoma de ARN de simple hebra y polaridad negativa (-ssRNA): en **negro** se indican todos los -ssRNA exceptuando a la familia *Arenaviridae*; en **rojo**, familia *Arenaviridae*.

Para intentar aproximarnos a lo que puede estar ocurriendo con los *Arenaviridae*, observamos que estos virus suelen presentar horquillas en sus genomas (segmentos L y S). Dichas horquillas funcionan como mecanismo de terminación de la transcripción (Le Mercier, 2021). En la **Figura 15** puede verse la estructura secundaria de la horquilla en el extremo 3' del segmento S de *Lymphocytic choriomeningitis mammarenavirus*, la especie tipo de los *Mammarenavirus*, el principal género de la familia. El contenido G+C de esta región es 0,63 mientras que el G+C global es 0,42. Si bien es un único par de valores, coinciden con los rangos aproximados de la gran mayoría de los puntos rojos («coordenadas [0,6; 0,4]»).

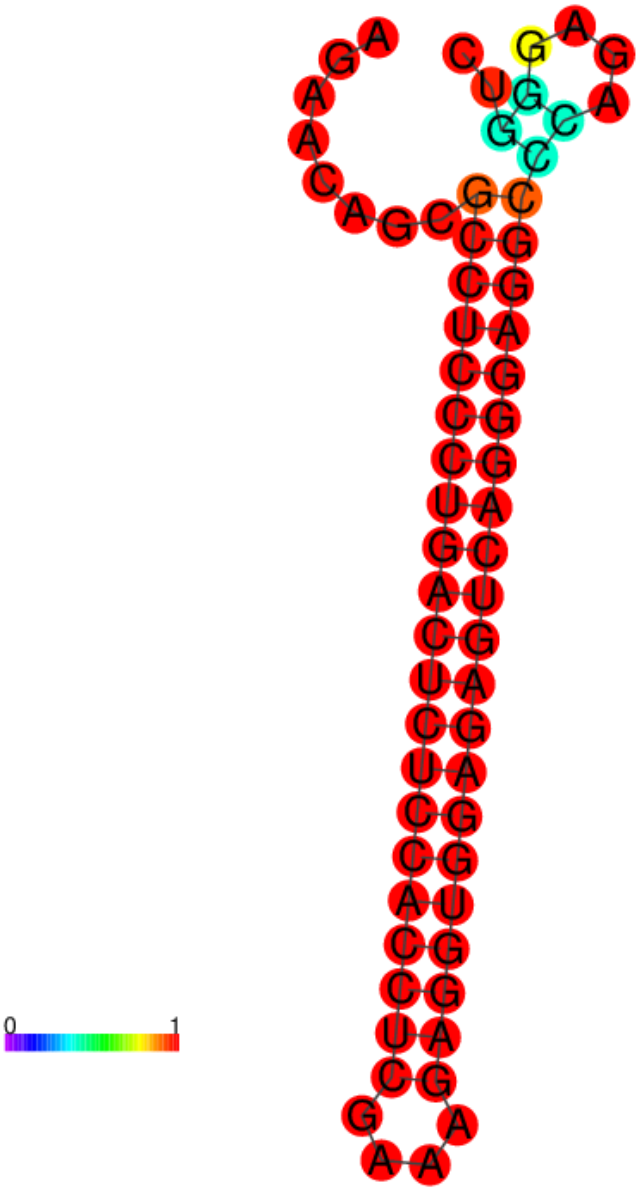



**Figura 15.** Estructura secundaria de la horquilla en el extremo 3’ del segmento S de *Lymphocytic choriomeningitis mammarenavirus* (`mfold` [Zuker, 2013], disponible desde www.unafold.org/).

**Los virus y sus hospederos**

La influencia del G+C del hospedero en los genomas virales es más que evidente. Esto puede darse de manera directa producto del sesgo composicional del hospedero. Pero también de manera indirecta: dado que el G+C del hospedero impacta sobre el uso de oligonucleótidos y de codones (y también de aminoácidos), si el virus es influido por estos factores, lo es también, indirectamente, por el G+C.

Los procariotas (árqueas y bacterias) presentan una distribución bimodal en el contenido de G+C de sus genomas (**Figura 16**; línea continua). Sin embargo, los fagos (sus virus) son trimodales en G+C (**Figura 16**, línea punteada).

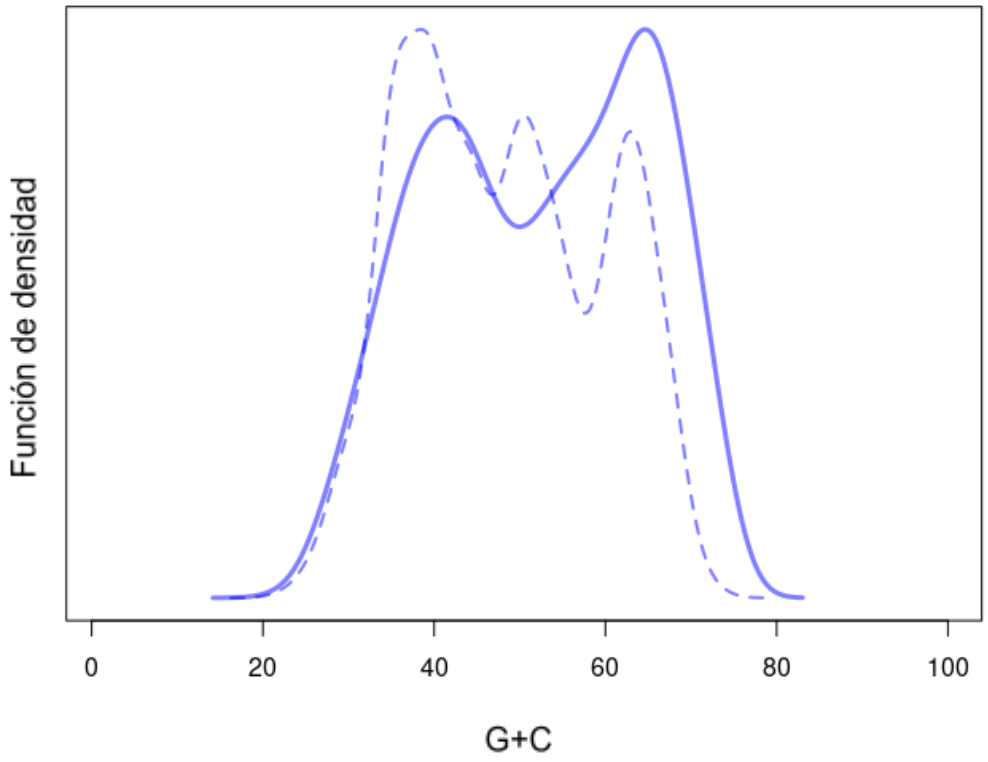



**Figura 16.** Contenido de G+C en los procariotas totalmente secuenciados (tomando un hospedero por género, N = 1169; **línea continua**) y en sus virus (línea punteada).

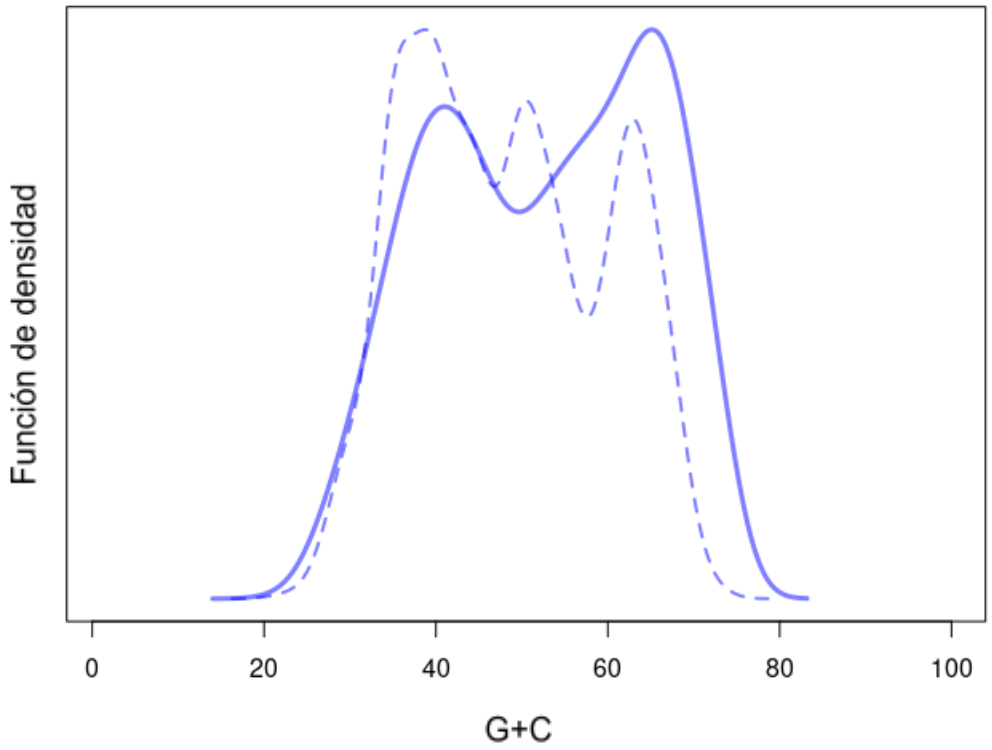



**Figura 17.** G+C en las bacterias totalmente secuenciadas (una especie por género, N = 1076; **línea continua**) y en sus virus (línea punteada).

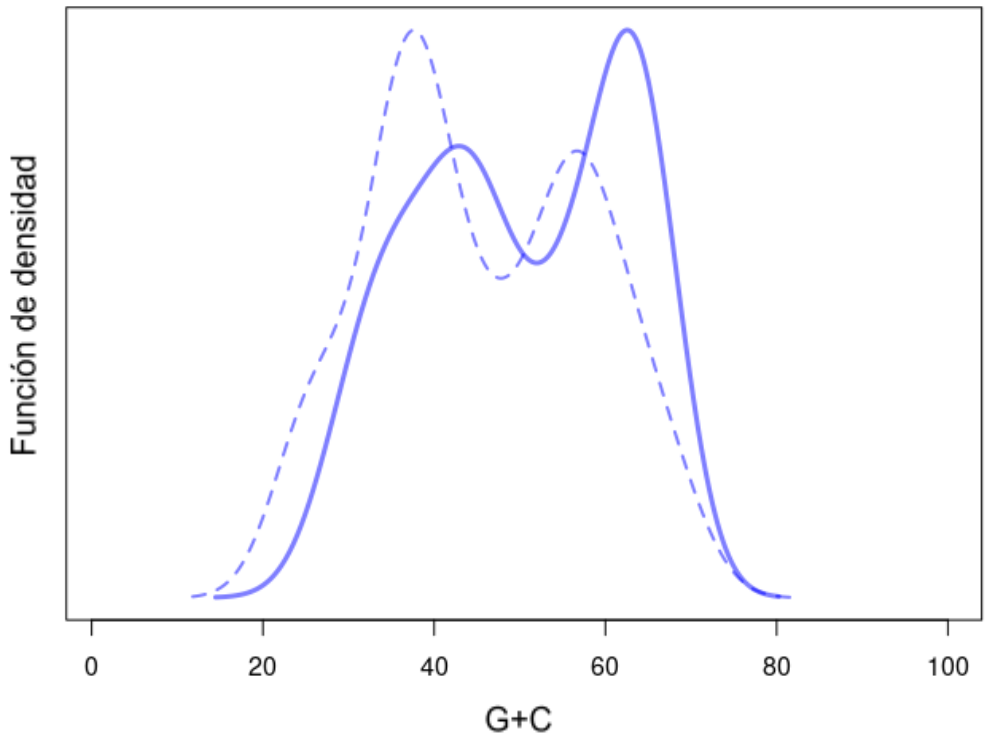



**Figura 18.** G+C en las árqueas totalmente secuenciadas (una especie por género, N = 93; **línea continua**) y en sus virus (línea punteada).

Esta trimodalidad en los fagos se observa solamente en los bacteriófagos (**Figura 17**; línea punteada) y no en los fagos de árqueas (**Figura 18**; línea punteada). Donde sí se observan tres modas en el material genético de los procariotas es en los plásmidos (**Figura 19**; línea continua). Es importante destacar que en este caso, y de manera similar a los fagos, un hospedero puede tener varios plásmidos.

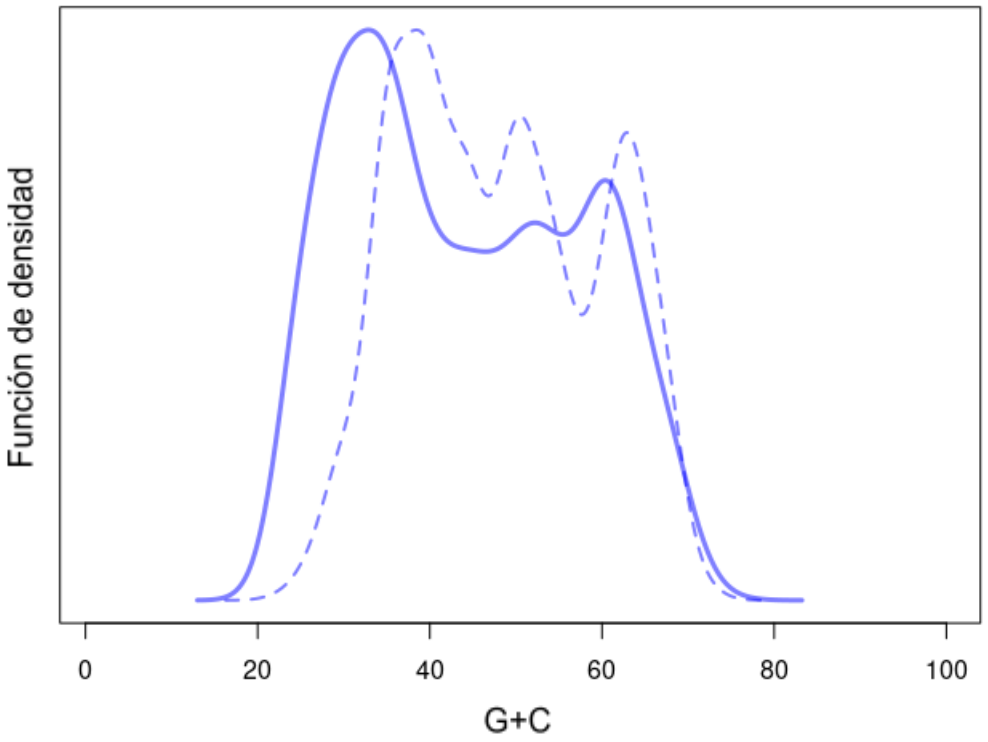



**Figura 19.** G+C de los plásmidos (N = 2353; **línea continua**) y en los fagos (línea punteada).

Esta similitud entre fagos y plásmidos recuerda la apreciación de Rocha y Danchin (2002), quienes consideran que los fagos y los plásmidos (junto con las secuencias de inserción) son elementos genéticos con un comportamiento de patógenos intracelulares. Estos elementos, además, tienden a tener un contenido de G+C inferior al de su hospedero (Rocha y Danchin, 2002).

Las distribuciones obtenidas para los distintos grupos de eucariotas y sus respectivos virus se disponibilizan en el **Apéndice D**.

**Códigos genéticos alternativos**

Los virus, al requerir la maquinaria traduccional de su hospedero, necesitan utilizar el mismo «dialecto» que el ribosoma que parasitan.

El código genético es un carácter ancestral de toda la vida conocida. Este código universal presenta leves variantes, que no son más que adaptaciones de uno o más codones, en general de los codones de terminación (*stop*). También se han descrito codones de inicio o iniciadores alternativos a AUG; algo importante es que pese a estos iniciadores alternativos, la traducción siempre comienza con una metionina en el extremo N-terminal (aunque dicho aminoácido no siempre quede en la proteína madura).

El NCBI se manifiesta cuidadoso al momento de asegurar que la traducción de las cds sea la correcta. Para lograrlo, se realiza una comprobación de taxonomía de cada nuevo registro y luego asocia el código genético correcto para esa taxonomía. Un nuevo virus puede saltarse esta comprobación de la taxonomía y por consiguiente quedar mal asignado al código estándar (1, por defecto).

Esto último puede ocurrir porque muchos grupos de virus no tienen bien establecida su clasificación taxonómica y/o su rango de hospedero. Esto es entendible cuando lo único que se conoce en algunos casos es su genoma, ya sea de manera parcial o completa, con cromosoma «cerrado» o no (i.e., *scaffolds* o *contigs*). Además, NCBI permite más flexibilidad a la hora de someter secuencias de virus (y también de microbios en general: bacterias, arqueas, eucariotas unicelulares).

En eucariotas, si bien lo más extendido son los códigos genéticos alternativos descritos para organelos como mitocondrias o plástidos, también existen variantes en algunos códigos genéticos nucleares. El código genético utilizado por la gran mayoría de los procariotas, fagos y cloroplastos es el número 11. El código es básicamente idéntico al estándar, con la diferencia que posee más y mejor documentados inicios de traducción alternativos. Existen al día de hoy 33 códigos genéticos descritos, de los cuales se mantienen vigentes un total de 25 códigos. Esta diferencia radica principalmente en que algunos códigos genéticos alternativos han sido reagrupados o reorganizados.

Los virus, al igual que la mayoría de los genomas celulares, utilizan en su amplísima mayoría los códigos 1 u 11. Muy pocos representantes tienen asignados códigos genéticos no canónicos. Los virus con códigos genéticos alternativos están asignados a los códigos 4, 5, 6 y 16.

El código genético 4 es el que utilizan los micoplasmas y espiroplasmas y el utilizado también en el genoma mitocondrial de celenterados, mohos y protozoarios; este código es el utilizado en el ADN de los kinetoplastos. La principal diferencia con el código genético estándar es que utilizan UGA como triptófano en vez de como codón de terminación.

El código genético 5 es aquel utilizado por los genomas mitocondriales de la gran mayoría de los invertebrados. Su diferencia principal radica en que el codón UGA codifica para triptófano y no como *stop*; otras diferencias no sinónimas son los codones AGA y AGG, que codifican para serina (y no arginina), y el codón AUA que codifica para metionina (y no isoleucina).

Tanto el código genético 4 como el código genético 5 incorporan triptófano con el codón UGA. Es interesante observar el patrón de distribución de frecuencias de este codón según los códigos genéticos presentes en los genomas virales RefSeq (**Figura 20**).

El código genético 6 es el utilizado en los genomas de algunos eucariotas unicelulares como los ciliados, los diplomonádidos o los trepomonádeos, aunque el código macronuclear ciliado no ha sido determinado completamente (Elzanowski y Ostell, 2019). En los genomas que utilizan este código genético, exclusivamente UGA codifica como codón de terminación. A diferencia de los grupos 4 y 5, los codones de parada co-optados codifican para glutamina.

Los virus que utilizan el código genético alternativo 6 se diferencian claramente del resto de los virus en la frecuencia de uso del codón UAA (**Figura 21**).

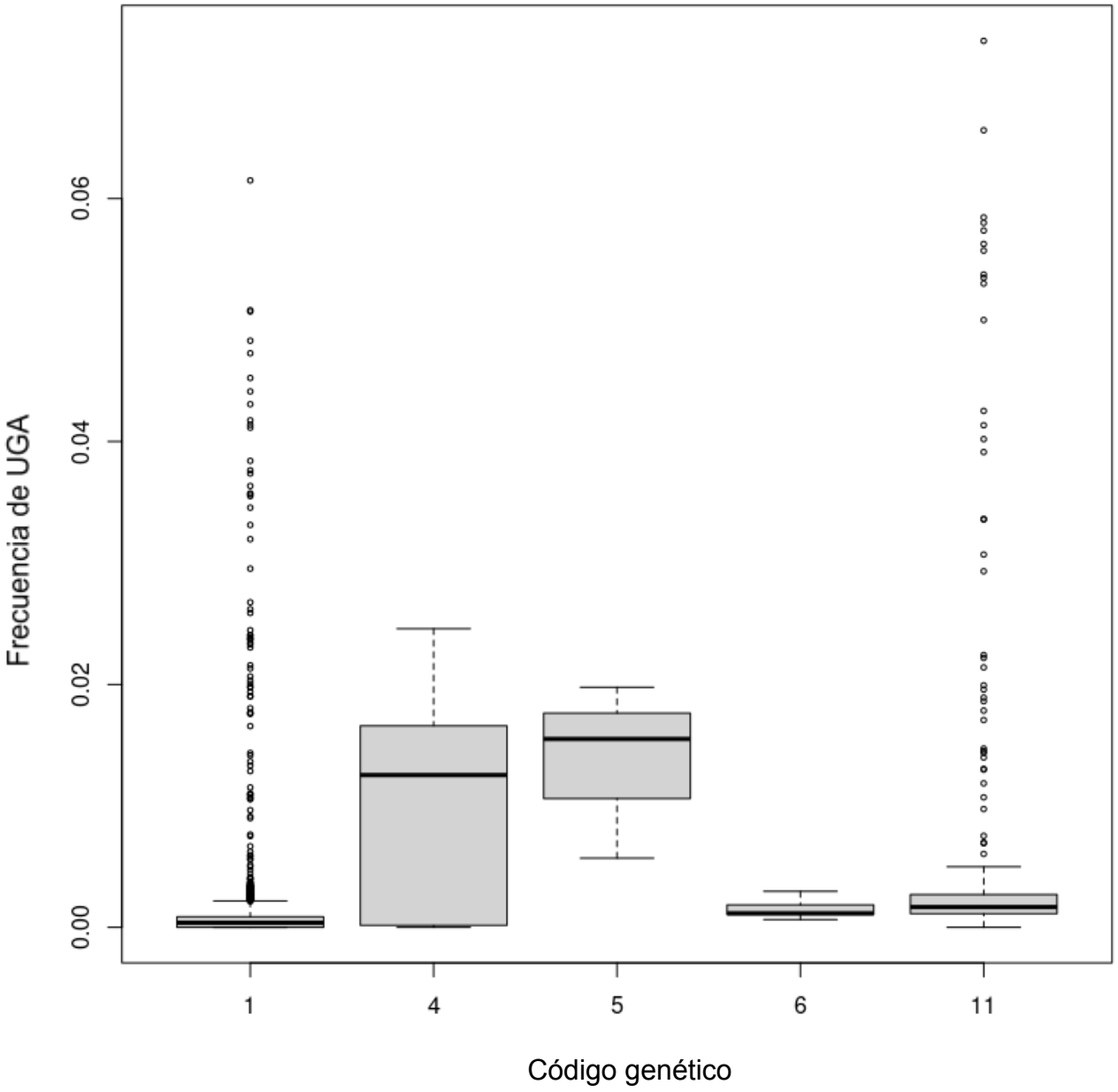



**Figura 20.** Frecuencia del codón UGA por código genético; UGA es uno de los codones *stop* en el código genético estándar (y en 6 y 11), pero codifica para el aminoácido triptófano en los códigos genéticos 4 y 5.

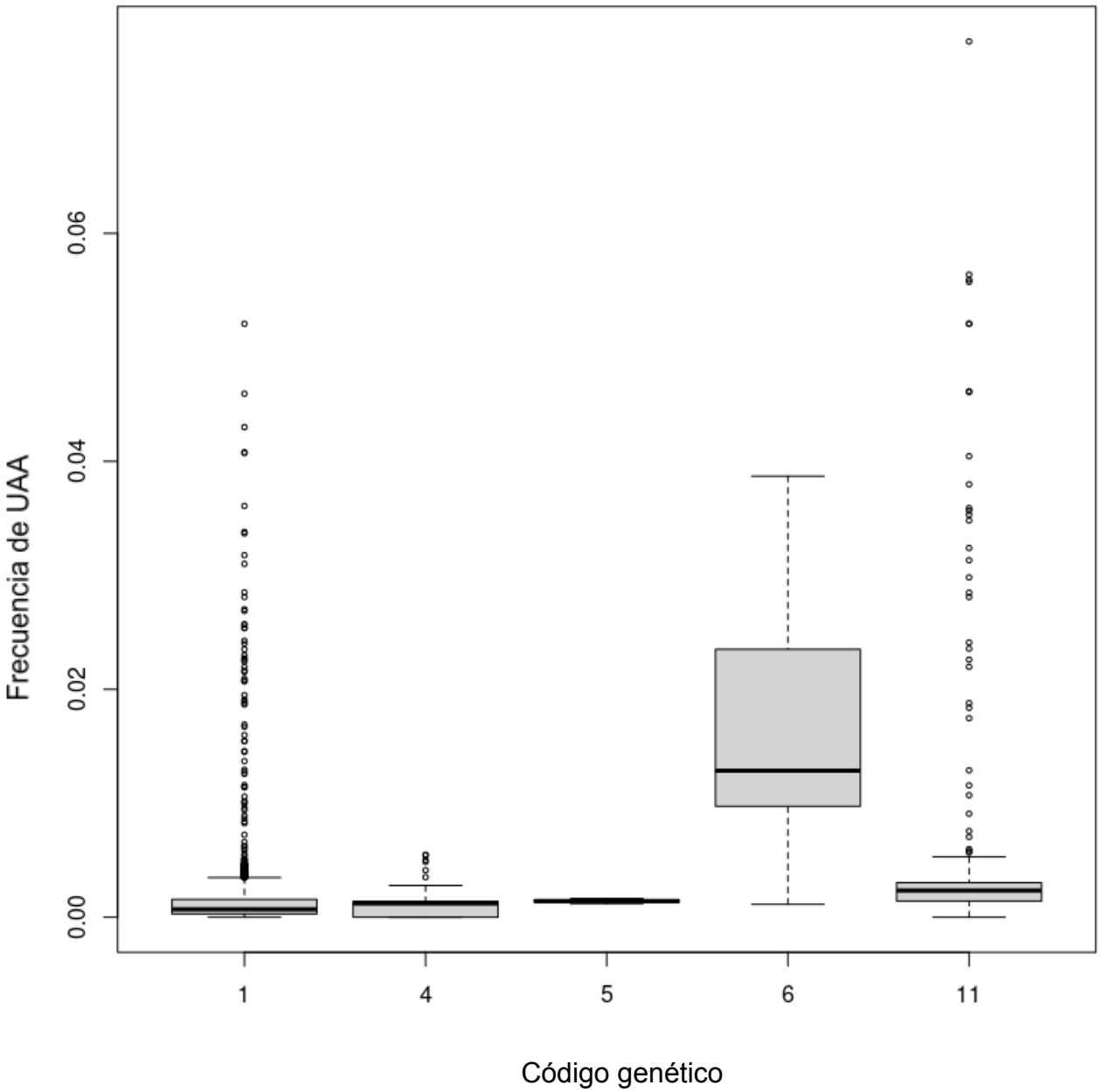



**Figura 21.** Frecuencia del codón UAA por código genético; UAA es uno de los codones *stop* en el código genético estándar (y en 4, 5 y 11), pero codifica para el aminoácido glutamina en el código genético 6.

Por último, el código genético 16 es el utilizado en los genomas mitocondriales de las clorofíceas, un gran grupo de algas verdes de agua dulce. La variante que presentan es que el codón *stop* UAG es incorporado como leucina, disponiendo entonces de siete codones que codifican leucina. Expresado de otro modo, más de un 11% de sus codones posibles (sin contar los dos *stop* restantes).

Por detalles completos de estos códigos genéticos, incluyendo sus sitios de inicio alternativos, ver **Anexo C**.

Fagos de ensamblaje cruzado

Los crAss (de *cross assembly*) son un grupo de fagos descritos recientemente producto de ensamblajes de metagenomas (Dutilh et al., 2014), que en muestras intestinales o fecales suelen alcanzar unas abundancias de hasta el 90% del viroma de dichas muestras (Yutin et al., 2018).

En la versión consultada de la base RefSeq (*release* 203) no se habían incorporado aún virus asignados con el código genético 16. Sin embargo, sí existen genomas disponibles en el GenBank que utilizan este código genético. En octubre de 2020 se publicaron 17 genomas que fueron asignados provisoriamente como "*UAG-readthrough crAss clade*", nombre que resalta el uso alternativo del codón UAG. Estos 17 genomas completos disponibilizados en octubre son dsDNA y cuentan con un tamaño aproximado de 100kpb. Resta aún determinar si estos virus asociados a algas verdes están relacionados o no con los fagos crAss. Resulta interesante, aunque nada sorprendente, que se proponga que estos «fagos» infectan mitocondrias, dado que estos organelos que derivan de *α*-proteobacterias (Sagan, 1967; Thrash et al., 2011).

Para abarcar también virus con un código genético alternativo más, el 16, analizamos las secuencias codificantes de estos 17 nuevos genomas virales y comparamos la frecuencia de UAG con el resto de los virus estudiados (**Figura 22**).

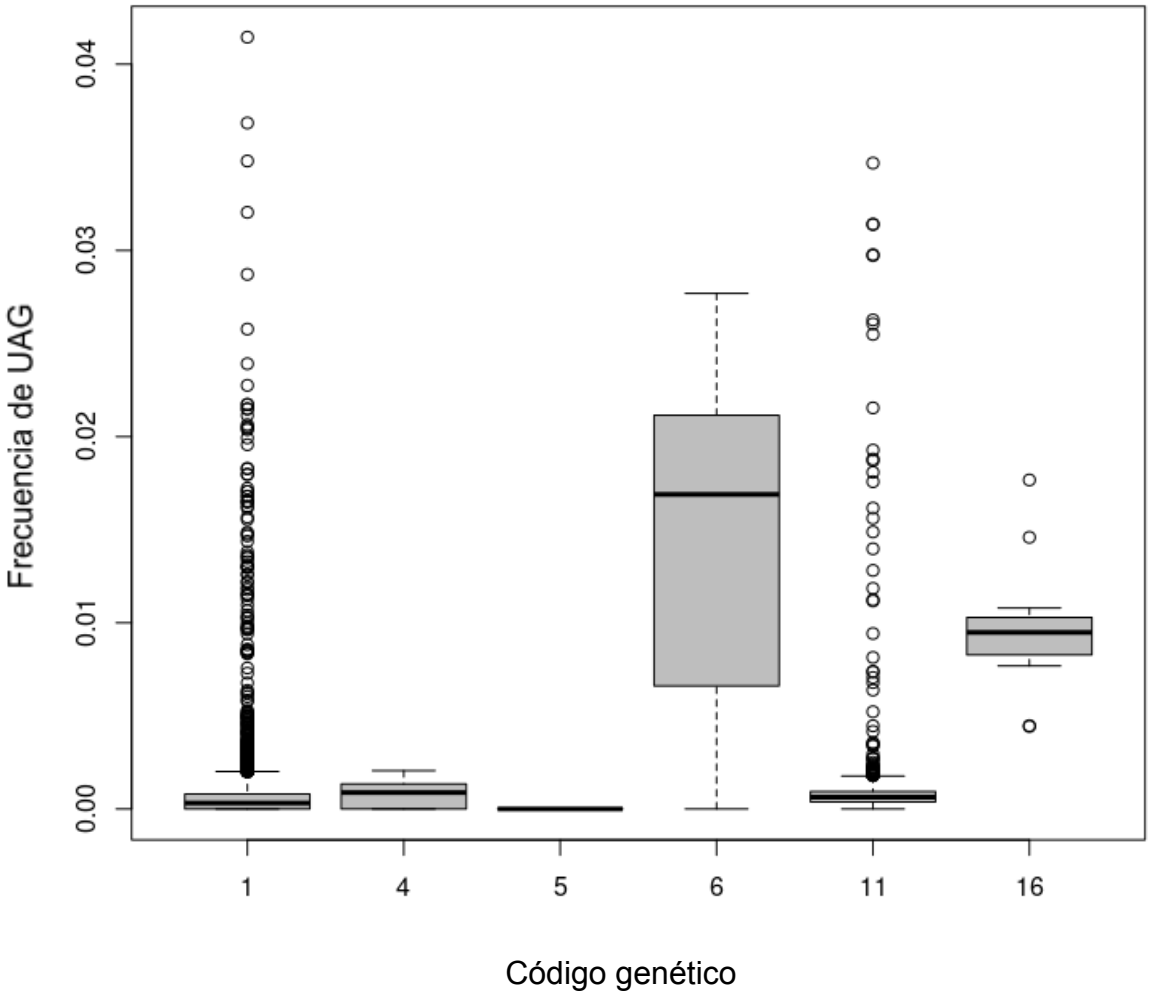



**Figura 22.** Frecuencia del codón UAG por código genético; UAG es uno de los codones *stop* en el código genético estándar (y en 4, 5 y 11), pero codifica para el aminoácido glutamina en el código genético 6 y para leucina en el código genético 16.

Los fagos crAss reasignaron el codón UAG pero mantienen a los codones UAA y UGA como *stop*. Este grupo prefiere significativamente al codón UAA (**Figura 23**).

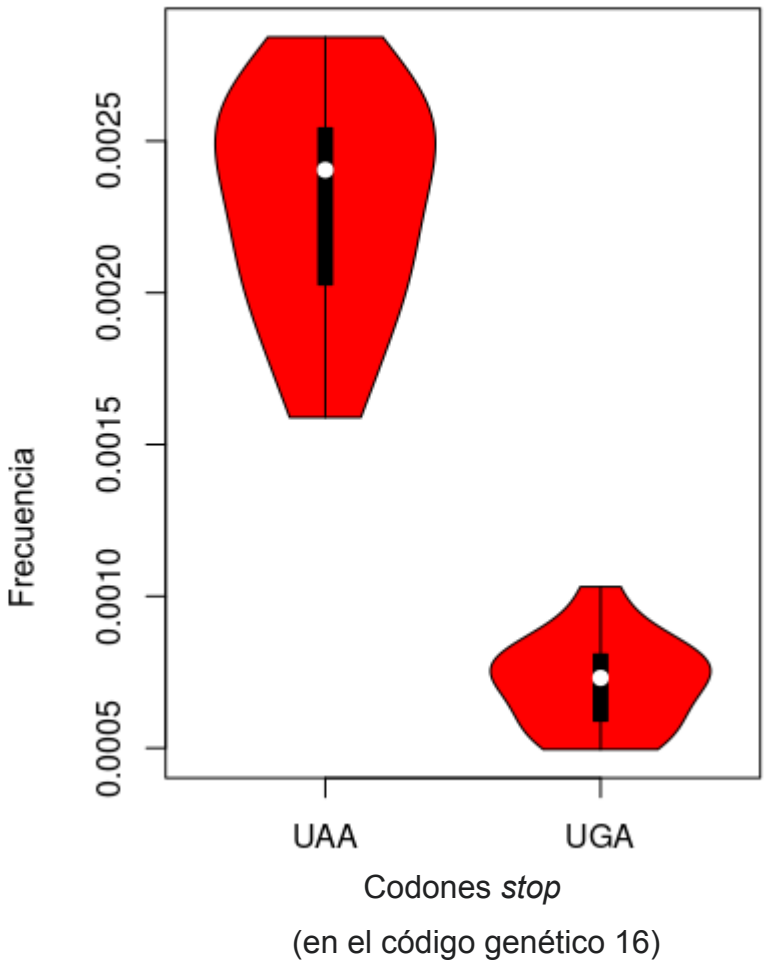



**Figura 23.** Frecuencia de uso de los codones *stop* UAA y UGA en los fagos crAss.

Los fagos crAss han reasignado el codón UAG como leucina. Procedimos, entonces, a analizar el uso relativo de la leucina en los genomas de los fagos crAss. Para ello se utilizó el valor de RSCU (**Figura 24**). Sin ser el codón más usado entre las 7 posibilidades, es utilizado con una frecuencia moderada. Puede ser interesante evaluar si su uso se distribuye de igual forma en toda extensión de sus proteínas o si es preferido hacia el extremo C-terminal. De todas maneras, dado su amplio uso, un mRNA de estos virus carecería de sentido en una célula con otro código genético (sin ir más lejos, en el citoplasma de la célula algal hospedera; es decir, fuera de la mitocondria).

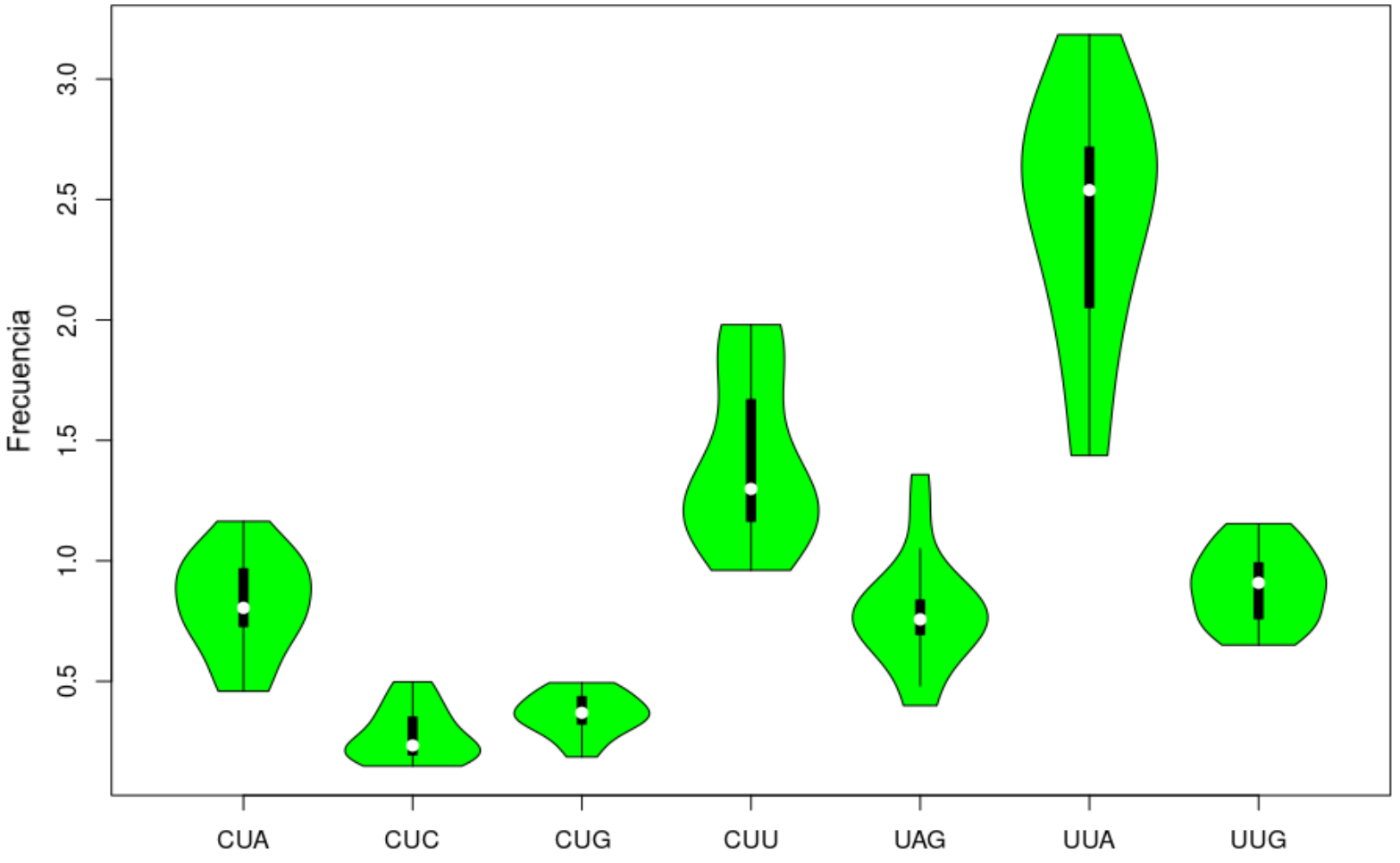



**Figura 24.** Distribución de los valores de RSCU para el aminoácido leucina en las regiones codificantes de los fagos crAss; el codón no estándar es el UAG (código genético 16).

Pese a tener siete codones para el aminoácido leucina, los fagos crAss no utilizan significativamente más este aminoácido en relación a toda la diversidad viral; la mediana de la frecuencia de leucina es 0.085 para los fagos crAss, mientras que para el resto de los virus analizados es 0.083 (**Figura 25**).

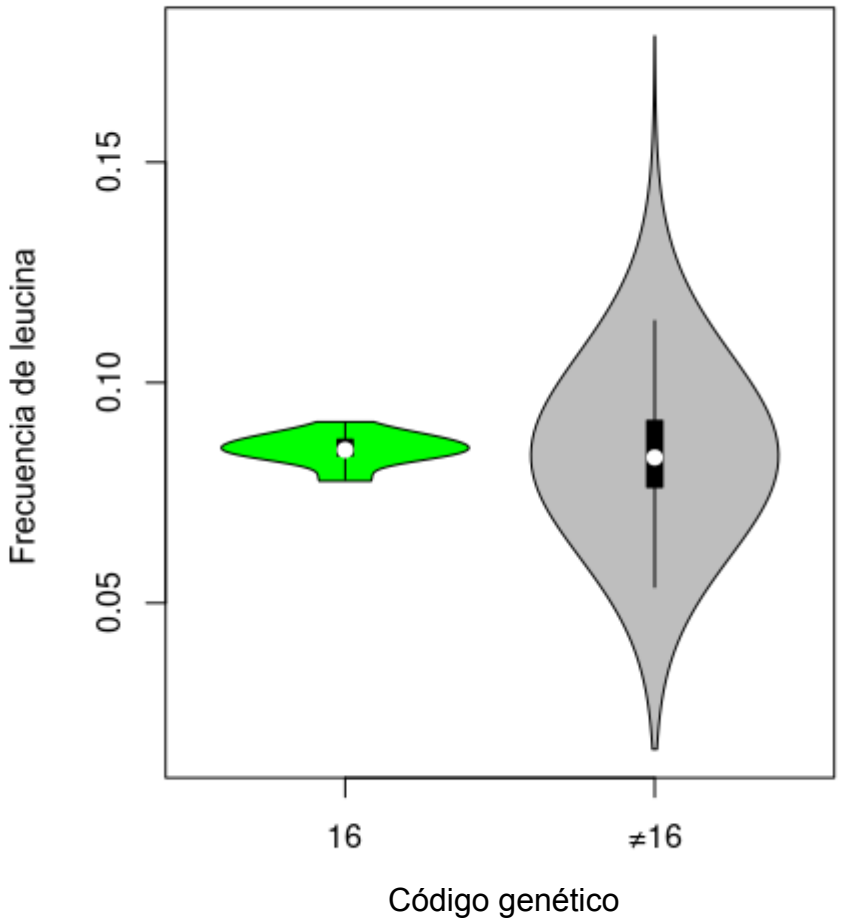



**Figura 25.** Distribución de los valores de frecuencias de uso del aminoácido leucina en los fagos crAss (código genético 16) y en los virus con otros códigos genéticos.

Recientemente, se describieron cientos de virus parecidos a los fagos crAss (*crAss-like phages*) que fueron ensamblados a partir de lecturas (*reads*) de secuencias metagenómicas de muestras intestinales humanas (Yutin et al., 2021). Una gran proporción de estos nuevos genomas (243/673 [36%]) presentaron códigos genéticos alternativos. El mecanismo propuesto es que existen ARNt supresores que permiten un *read-through* de los codones stop UAG y/o UGA. Al parecer, esta reasignación de codones sucedería en algunos genes de expresión tardía, asignando UAG como glutamina y UGA como triptófano (Yutin et al., 2021).

Muy probablemente, grupos de virus aún desconocidos o muy poco estudiados escondan nuevos códigos genéticos alternativos por descubrir.

## Trabajos publicados

Hasta el momento de la presentación de esta tesis, y producto de los análisis hechos durante la misma, se publicaron dos artículos en revistas arbitradas:

**(i)** *Host influence in the genomic composition of flaviviruses: A multivariate approach* (disponible en https://doi.org/10.1016/j.bbrc.2017.06.088 o en https://www.researchgate.net/publication/317690149_Host_influence_in_the_genomic_composition_of_flaviviruses_A_multivariate_approach)

**(ii)** *Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts* (disponible en https://doi.org/10.3389/fmicb.2021.646300)

Se presentan a continuación ambos manuscritos y sus respectivas figuras suplementarias, acompañados de los abordajes utilizados y sus contribuciones.

**Host influence in the genomic composition of flaviviruses: A multivariate approach**

«Influencia del hospedero en la composición genómica de los flavivirus: un enfoque multivariado»

Abordaje

Según el ICTV, al momento de escribir este manuscrito, el género *Flavivirus* comprendía 53 especies con una amplia distribución geográfica, además de un número creciente de especies no clasificadas o tentativas. Sus genomas se traducen en una única poliproteína que se divide en tres proteínas estructurales (C, prM y E) y siete proteínas no estructurales (NS1, NS2A, NS2B, NS3, NS4A, NS4B y NS5). A pesar de la similitud en su organización genómica, existen diferencias sustanciales en el rango de huéspedes y la transmisibilidad entre ellos.

La mayoría de las especies conocidas son arbovirus, que se transmiten horizontalmente entre artrópodos hematófagos y hospederos vertebrados susceptibles, y se clasifican en flavivirus transmitidos por mosquitos (MBFV, de *mosquito-borne flaviviruses*) y flavivirus transmitidos por garrapatas (TBFV, de *tick-borne flaviviruses*). Sin embargo, algunas especies solo se replican en murciélagos o roedores sin conocerse vectores asignados a estos virus (NKV, de *not-known vector*). Además, varias especies exclusivamente infectan a mosquitos, los cuales se denominan flavivirus específicos de insectos (ISFV, de *insect-specific flavivirus*).

Fruto de una colaboración con Álvaro Fajardo del Laboratorio de Virología Molecular, y a partir de análisis filogenéticos presentados en su tesis doctoral (Fajardo, 2016), publicamos este trabajo que examinó, en dicho contexto filogenético, la influencia de los diferentes hospederos sobre la composición genómica y sobre el uso de codones y de aminoácidos de los flavivirus.

El abordaje fue analizar las propiedades composicionales de cada virus, como la composición de bases, los sesgos de dinucleótidos, el uso de codones y las frecuencias de aminoácidos.

Aprovechando el gran número de secuencias disponibles, se analizaron en total noventa y ocho genomas de flavivirus y seis genomas representativos de los hospederos: tres artrópodos (*Aedes aegypti*, *Culex pipiens* e *Ixodes scapularis*) y tres vertebrados (*Gallus gallus*, *Homo sapiens* y *Mus musculus*).

Utilizando estadística multivariada (en concreto, análisis de componentes principales o PCA), logramos describir la influencia en los genomas virales tanto de la historia evolutiva de los flavivirus (sus relaciones filogenéticas) como de sus respectivos hospederos y/o vectores. Dimos cuenta de los diferentes patrones presentes en los flavivirus asociados a los mosquitos de los géneros *Aedes* o *Culex*, o en aquellos transmitidos por garrapatas.

Biochemical and Biophysical Research Communications 492 (2017) 572–578

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

# Host influence in the genomic composition of flaviviruses: A multivariate approach

Diego Simón [a], Alvaro Fajardo [b], Martín Sóñora [b], Adriana Delfraro [c], Héctor Musto [a, *]

[a] Laboratorio de Organización y Evolución del Genoma, Unidad de Genómica Evolutiva, Facultad de Ciencias (FC), Universidad de la República (UDELAR), Iguá 4225, Montevideo 11400, Uruguay
[b] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, FC, UDELAR, Uruguay
[c] Sección Virología, FC, UDELAR, Uruguay

ARTICLE INFO

*Article history:*
Received 16 May 2017
Received in revised form
9 June 2017
Accepted 15 June 2017
Available online 17 June 2017

*Keywords:*
Flavivirus
Base composition
Dinucleotides
Codon usage
Amino acids

ABSTRACT

Flaviviruses present substantial differences in their host range and transmissibility. We studied the evolution of base composition, dinucleotide biases, codon usage and amino acid frequencies in the genus *Flavivirus* within a phylogenetic framework by principal components analysis. There is a mutual interplay between the evolutionary history of flaviviruses and their respective vectors and/or hosts. Hosts associated to distinct phylogenetic groups may be driving flaviviruses at different pace and through various sequence landscapes, as can be seen for viruses associated with *Aedes* or *Culex* spp., although phylogenetic inertia cannot be ruled out. In some cases, viruses face even opposite forces. For instance, in tick-borne flaviviruses, while vertebrate hosts exert pressure to deplete their CpG, tick vectors drive them to exhibit GC-rich codons. Within a vertebrate environment, natural selection appears to be acting on the viral genome to overcome the immune system. On the other side, within an arthropod environment, mutational biases seem to be the dominant forces.



## 1. Introduction

The genus *Flavivirus* belongs to the family *Flaviviridae*, together with *Hepacivirus*, *Pegivirus* and *Pestivirus*. According to the International Committee of Virus Taxonomy, the genus comprises 53 species with wide global distribution, as well as an increasing number of unclassified or tentative species [1]. They are positive-sense single-stranded RNA viruses of about 11 kb, with a 5′ type I cap structure and lacking a poly(A) tail at the 3′ end. Their genome is translated in a single polyprotein which is cleaved in three structural proteins (C, prM and E) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) [2].

Despite the similarity in their genomic organization, there are substantial differences in the host range and transmissibility among them. Most known species are arboviruses, which are transmitted horizontally between hematophagous arthropods and susceptible vertebrate hosts, and are classified in mosquito-borne flaviviruses (MBFV) and tick-borne flaviviruses (TBFV). However, some species only replicate in bats or rodents with not-known vector associated to them (NKV). Furthermore, several species only infect mosquitoes, which are referred to as insect-specific flaviviruses (ISFV) [3–9].

The taxonomic relationship among flaviviruses has been extensively investigated through different approaches, originally based on antigenic cross-reactivity in neutralization, complement fixation and hemagglutination inhibition assays [10,11]. Lately, phylogenetic reconstructions based on nucleotide and amino acid sequences allowed a deeper understanding of the diversity of the genus. Several methodological approaches were followed to analyze different genes and complete coding regions [3–8,12–14]. As a result of these efforts it was found that the general pattern of the inferred phylogenetic relationships correlates with the main epidemiological aspects as host range, vectors and related diseases. Nevertheless, the comparison of different analyses evidences phylogenetic incongruences that difficult a proper definition of the taxonomic relationships.

Another approach to understand both the evolution and the phylogenetic relationships, is to analyze compositional properties of each virus, such as base composition, dinucleotide biases, codon

* Corresponding author.
*E-mail addresses:* dsimon@fcien.edu.uy (D. Simón), afajardo@cin.edu.uy (A. Fajardo), msonora@cin.edu.uy (M. Sóñora), adelfraro@gmail.com (A. Delfraro), hmusto@gmail.com (H. Musto).

http://dx.doi.org/10.1016/j.bbrc.2017.06.088
0006-291X/

usage and amino acid frequencies. These features can be defined as molecular signatures. Taking advantage of the great number of sequences available, in the present communication we update the analyses of these genomic compositional properties in the genus *Flavivirus* within a phylogenetic framework.

## 2. Materials y methods

### 2.1. Dataset construction

Coding sequences (CDS) available from all viral types belonging to the genus *Flavivirus* were retrieved from the ViPR database (Virus Pathogen Database and Analysis Resource) of the National Institute of Allergy and Infectious Diseases, available at http://www.viprbrc.org [15]. Tamana bat virus, an unclassified *Flavivirus*, was excluded from these analyses for being highly divergent [16]. Information about hosts (arthropod and/or vertebrate), status as arbovirus and human pathogenicity were obtained from Arbovirus Catalog, Virus-Host DB and/or ViPR. For details, see Supplementary Table 1.

### 2.2. Phylogenetic analyses

Polyprotein sequences of each species were aligned with MUSCLE [17]. Regions of ambiguous alignment were excised with GBlocks v.0.91b [18,19]. The optimal amino acid substitution model was inferred with ProtTest 3 [20]. With the selected evolutionary model (WAG + F + G + I), maximum likelihood phylogenetic trees were constructed through PhyML v.3.0 [21]. A bootstrap test with 500 replicates was used to evaluate the robustness of each node. FigTree v1.4.2 (available at http://tree.bio.ed.ac.uk/software/figtree) was used to edit the final trees.

### 2.3. Hosts genomes

*Aedes*, *Culex* and *Ixodes*, as the main arthropod hosts, and *Gallus*, *Homo* and *Mus*, representing the three major groups of vertebrate hosts in the sample (fowl, primates and rodents, respectively), were analyzed in a bigger extent. The CDS were obtained from Ensembl repositories [22]: *H. sapiens* and *M. musculus*, from CCDS project; *G. gallus*, from Ensembl genome browser 88; *A. aegypti*, *C. pipiens* and *I. scapularis*, from Ensembl Metazoa.

### 2.4. Compositional analysis

For each viral genome, different compositional properties were obtained with R package seqinr [23]. Only the CDS of the whole polyproteins were considered, discarding some other putative proteins. The same properties were obtained also for the set of hosts.

### 2.5. Principal component analysis

Principal component analysis (PCA) is a statistical procedure to reduce the multidimensionality of the data. PCA allows the observation of patterns, either forming clusters or gradients. Interpreting the weights given by the original numerical variables is also informative. In order to infer associations between the axes and some compositional features, Pearson correlations were tested. PCA were performed in R with function pr.comp [24]. The proportion of variance explained by the ten principle components are displayed as Supplementary Fig. 2. Graphical representations were constructed with scatter3dplot package [25].

## 3. Results and discussion

### 3.1. Frequency of bases

Fig. 1 shows the inferred phylogenetic relationships between the different flaviviruses analyzed in this study. The sequences cluster in four monophyletic groups: MBFV (blue), NKV (green), TBFV (red) and ISFV (purple) as previously described [7,17,25].

The complete compositional analyses are presented as Supplementary Table 2. Overviewing some descriptive statistics of these analyses, the four main groups display heterogeneity in their base composition. In relation to the other groups, ISFV are A/G-poor and U-rich, MBFV are G/U-rich and A-poor, while NKV are G/C-poor and A/U-rich. Finally, TBFV are A/U-poor and G-rich. ISFV are characterized by their relatively high pyrimidines content, opposite to the rest of the groups which tend to be enriched in purines. TBFV are clearly the richest in GC%, which stands for all codon positions while the contrary is true for NKV. These patterns agree with previous studies [26,27] and also coincide with the observed values in their hosts (see Table S3). For instance, *Ixodes*, the predominant vector within TBFV group, has a mean coding GC (cGC) of 58% (and a GC3 of 72%). *Culex* spp. also exhibit high GC% (cGC of 55%; GC3 of 69%), while *Aedes* spp. presents a mean cGC slightly lower than vertebrate hosts analyzed (i.e.; *Gallus*, *Homo* and *Mus*).

The differences in base composition between mosquitoes are also extensive to those flaviviruses that replicates in these insects. The mean cGC values for all flaviviruses which have either *Aedes* or *Culex* spp. as their main hosts are 49.3% ± 2.3 and 50.5% ± 1.3, respectively. ISFV includes two distinct clades: ISFV-*Aedes* (49.1% ± 2.0) and ISFV-*Culex* (51.1% ± 2.1); both subgroups are mainly associated with each genus. Although these intervals overlap, the trend is consistent and suggests an effect of the GC% of the mosquito host in the viral base composition.

Other groups of viruses that are believed to be insect-specific have been described [3]. Two distinct clades, phylogenetically related to MBFV, seem to circulate strictly in mosquitoes since none have been isolated from vertebrates or cell lines derived from them [28,29]. The species within MBFV that shows an ISFV-like phenotype could be separated in two clades; ISFV-2: CHAOV, DONV, ILOV, and LAMV; and ISFV-3: BARKV, NHUV, and NOUV. The topology observed suggests that the emergence of these clades occurred in two independent events, as was previously suggested [3,30]. Moreover, these groups are also primarily associated with aedine (ISFV-2) or culicine (ISFV-3) mosquitoes (see Supplementary Table 1). ISFV-2 (48.9% ± 0.6) have lower mean cGC values than ISFV-3 (50.9% ± 1.1).

### 3.2. Dinucleotides

Concerning dinucleotide observed/expected values (see Supplementary Table 2), our results confirm previous reports [31], which can be summarized as follows: UpA is consistently underrepresented in all flaviviruses; TBFV have the lowest mean values (0.46 ± 0.03), while the other groups present similar values among them. The three groups associated with a vertebrate host (i.e.; MBFV, NKV and TBFV) also present biases in CpG, being NKV the most biased (0.34 ± 0.05); the mean values of ISFV-2 (0.66 ± 0.03) and ISFV-3 (0.65 ± 0.05) were intermediate between MBFV (0.52 ± 0.09) and ISFV (0.86 ± 0.09). CpA and UpG are the dinucleotides with more over-represented values.

The biases in CpG and UpA, which are almost universal for vertebrate RNA viruses, have been associated with an antiviral effect [27,31–34], although the mechanisms remain unclarified.

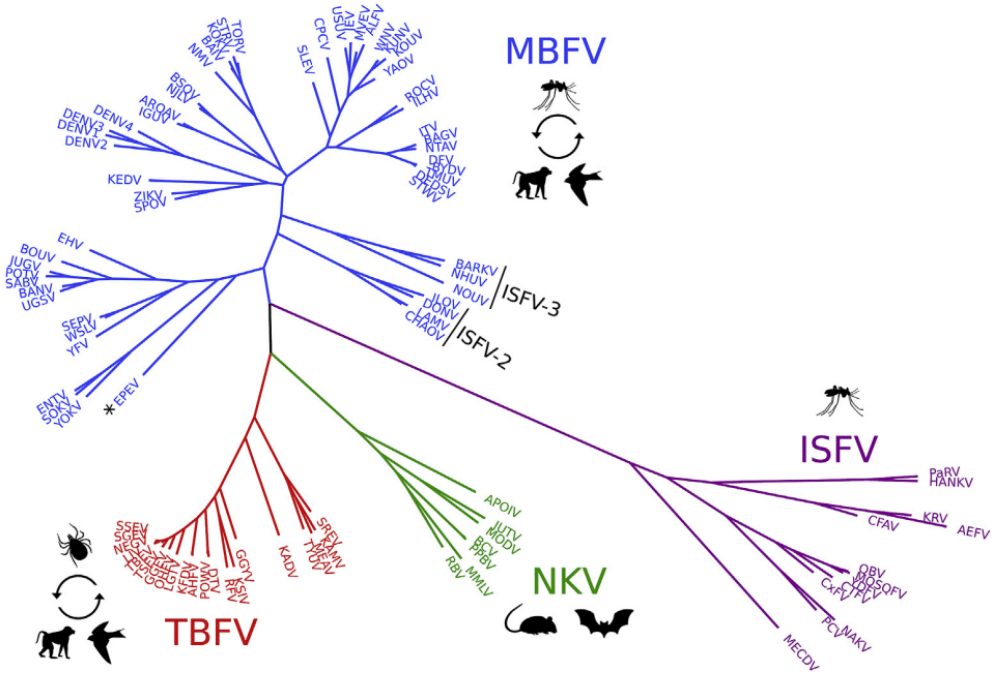



**Fig. 1.** Phylogenetic tree of flaviviruses analyzed in this communication.

Therefore, the CpG methylation/deamination process can explain biases in DNA viruses (especially those with small genomes), but not in RNA viruses.

To go further into the deviations from expected values, and considering that dinucleotide biases by its very nature are multivariate, it is necessary to analyze the data using multivariate statistical techniques, like PCA. The result of this analysis is displayed in a three-dimensional representation (3D) (Fig. 2a). In Fig. 2b, the plane determined by the first and second axes is displayed. These axes account for 54% and 13% of the total variability, respectively. It can be clearly seen that the first axis places the biggest weight on CpG. As stated earlier, CpG is suggested to be recognized by some uncharacterized molecular mechanism in vertebrates as a target for the innate immune response. This is coherent with the distribution of viruses within axis 1: ISFV display the most negatives values (i.e.; highest levels of CpG) and those that infect vertebrates show positive values (Fig. 2b and c); almost all arboviruses distribute on the two right quadrants. The second axis places almost equal weight to UpU, CpU and UpA on one side, and CpU at the other end (Fig. 2d).

If we correlate the position of each dinucleotide with the

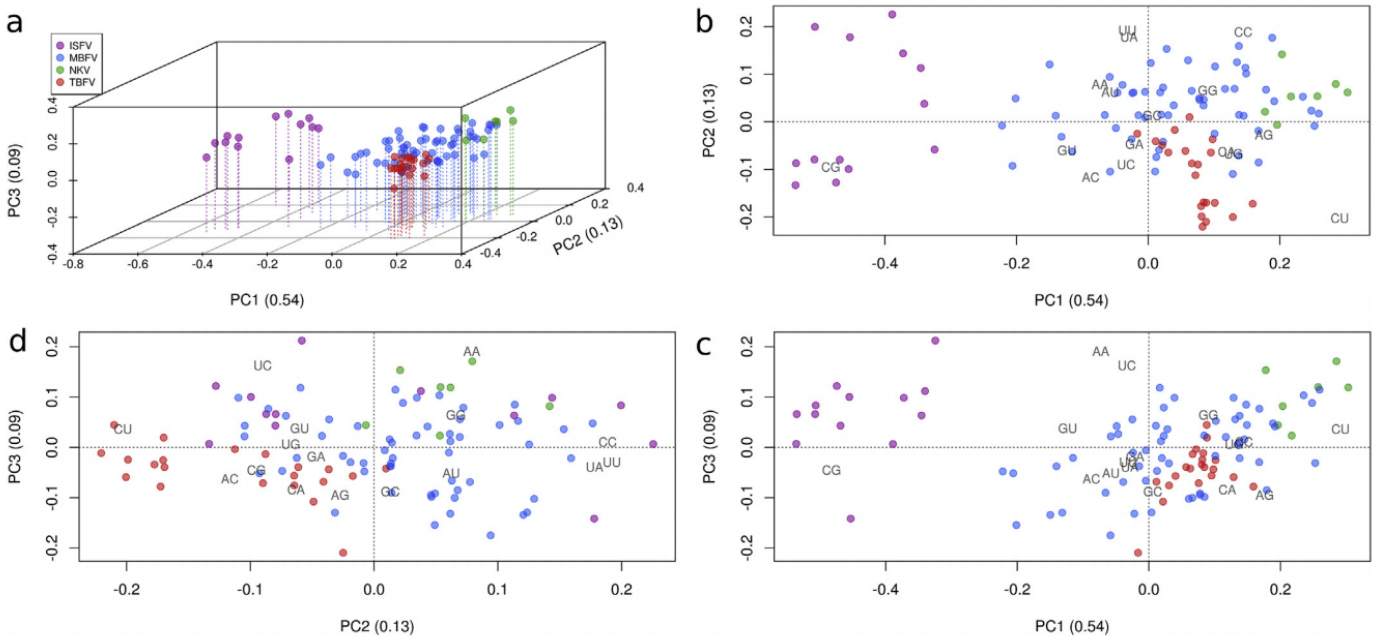



**Fig. 2.** (a) 3D plot of viruses according to the three main axes of PCA using dinucleotide bias (observed/expected ratio) as input. Position of the viruses and dinucleotides according to the plane defined by the first two axes (b); axes one and three (c); and axes two and three (d).

respective position on the three main axes we find the following. The first axis is strongly correlated with the CpG bias (r = −0.98). Also, there is a highly negative correlation (r = −0.78) with GpU. As pointed before, CpG is considered as an antiviral activator as is the case for GpU, being targeted by TLR7 and/or TLR8 [34].

Axis 1 also correlates (but positively) with CpU, ApG, UpG, CpC, GpG and CpA (with r values of 0.84, 0.80, 0.65, 0.62, 0.61 and 0.59, respectively); this appears like a non-random set of dinucleotides, since all of them are one mutation away from CpG. This reinforces the idea of CpG depletion in vertebrate-host flavivirus, while ISFV are unbiased to this dinucleotide, suggesting not such antiviral effect (and no such bias in their mosquito-host genomes also). The over-representation of CpA and UpG, that resembles the pattern obtained by CpG methylation/deamination, could also be explained with the same hypothesis invoked before: CpA and UpG are one transition away from UpA.

The main correlations with the second axis are with UpU (r = 0.75) and to a minor degree with UpA (r = 0.69). Furthermore, GC% correlates with this axis too. Indeed, cGC is negatively correlated with the second axis (r = −0.72), as are GC1 (r = −0.74), GC2 (r = −0.47) and GC3 (r = −0.70). The lower correlation is found with GC2, since this position is the less variable given that any change at these sites will have consequently an aminoacidic substitution. Finally, the third axis (9% of the variability) is mainly correlated with ApA (r = 0.78) and UpC (r = 0.70). Interestingly, this axis is more strongly correlated with GC2 (r = −0.58).

Considering the position of each virus analyzed, the 3D plot tends to discriminate among the main phylogenetical groups (Fig. 2a). ISFV display the most negative values in relation to axis 1. In axes 2 and 3, ISFV behave differently according to its main host (Fig. 2d). TBFV tends to cluster together with positive values for axis 1 and negatives for axis 2, as expected based on their high GC%. MBFV are more scattered in the space; the over-representation of MBFV viruses and their diversity could explain these ranges. Finally, NKV have the most positive values for axis 1 and slightly positive values to axis 2. ISFV-2 and ISFV-3 present intermediates positions between ISFV and MBFV, suggesting that they are tending to the CpG values of ISFV (see Supplementary Fig. 1a).

The patterns displayed in the space defined by the first three principal components, are the result not only of phylogenetic inertia but, more important, are related to the influence of the GC content of their respective hosts. For instance, viruses located at the most positive values for axis 1 replicate in a vertebrate, which leads to a known decrease in CpG. Similar situations will be described below both for codon usage and amino acids frequencies.

### 3.3. Codon usage

The input for PCA in codon usage was the relative synonymous codon usage (RSCU). This value is defined as the ratio of the observed frequency of codons to the expected frequency if all the synonymous triplets for the same amino acids are used equally. It has no relation to the amino acids usage or protein length [35]. The 3D plot of this analysis is displayed in Fig. 3a. Axis 1, which accounts for 37% of the total variability, places the biggest weight on AGA (Arg) at negative values; the viruses more negative in relation to axis 1 are those belonging to NKV. The general pattern is actually determined by axes 1 and 2 (Fig. 3b). The second axis captures a proportion of variance of 19%. At the quadrant corresponding to positive values of both axes 1 and 2, ISFV splits from arboviruses and NKV. Similar to the observed for dinucleotide biases, ISFV-2 and ISFV-3 tend towards the ISFV quadrant (see Supplementary Fig. 1b). This quadrant is dominated by CGN, all these triplets coding for Arg. At the opposite quadrant (negatives values of axes 1 and 2), AGG and AGA (both also coding for Arg) are among the dominant triplets. This is a clear indication that the usage of triplets coding for Arg are a main force driving codon usage among these viruses, as has been noted previously [31]. Several compositional features are strongly correlated with the first axis. This is the case of GC3 (r = 0.87) and cGC (r = 0.82). Moreover, the codons which display the most positive correlations are CCG (Pro, r = 0.82), CGG (Arg, r = 0.78), CGC (Arg, r = 0.78), AUC (Ile, r = 0.77). Between the negative correlated variables are AGA (Arg, r = −0.90), CCA (Pro, r = −0.81, AAU (Asn, r = −0.68), UCA (Ser, r = −0.68).

Finally, the third main axis (9% of the variability) strongly discriminates between U/G-ending codons against A/C-ending ones

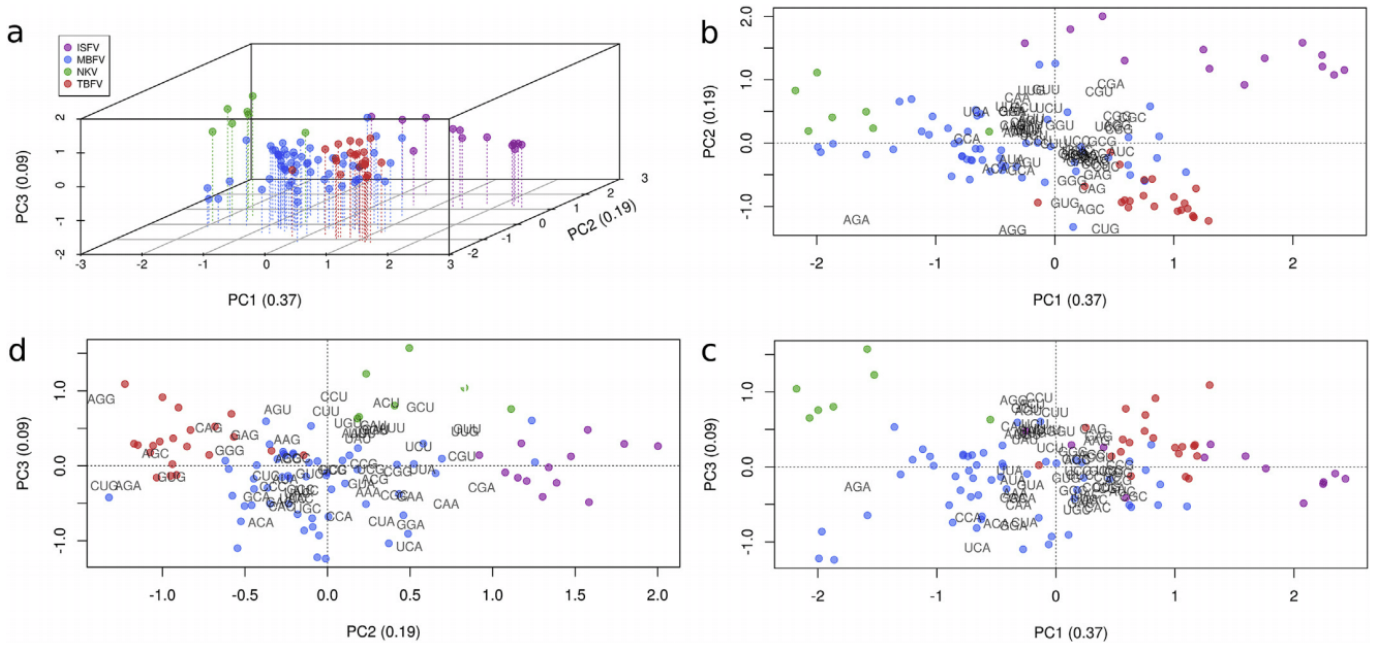



**Fig. 3.** Same as Fig. 2 but using RSCU as input.

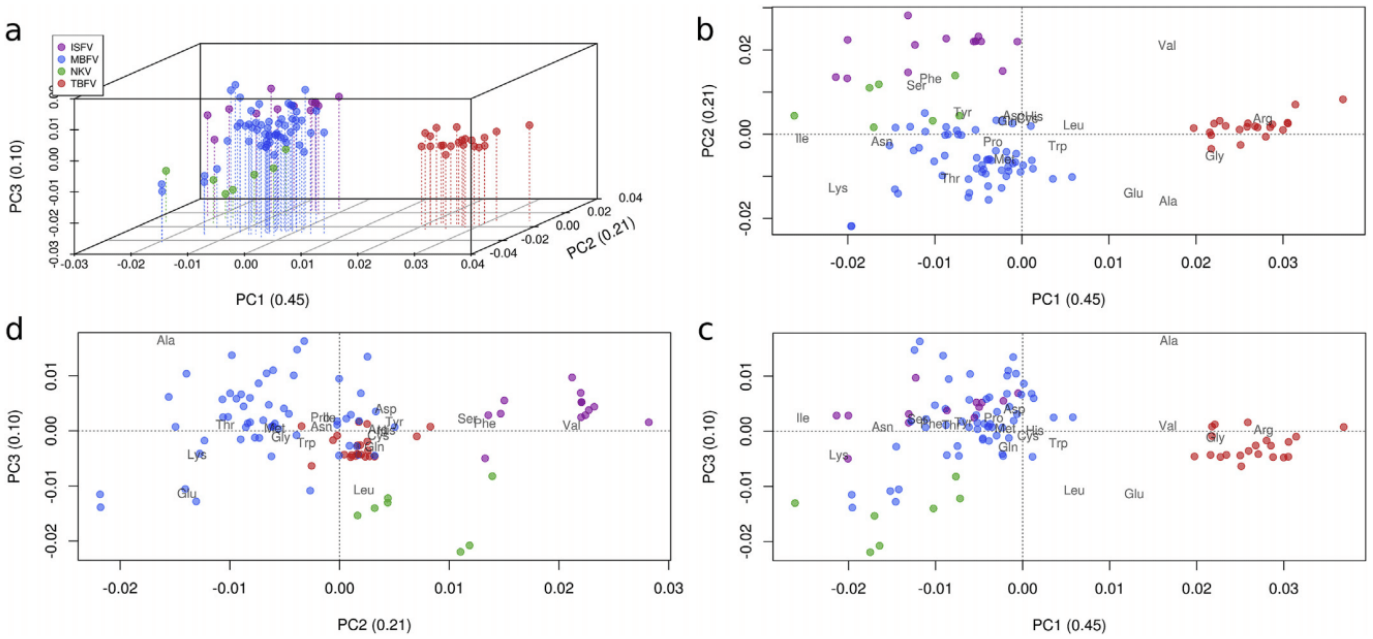



**Fig. 4.** Same as Fig. 2 but using amino acid frequencies as input.

(Fig. 3c and d). Indeed, the first 21 triplets with positive values according to this axis are U/G-ending while the first 20 with negative values display an A or a C at the third codon position. The location of viruses in the space defined by the three axes of the PCA tends, as happens with dinucleotides, to discriminate among the main phylogenetical groups. Indeed, ISFV cluster at positive values both for axis 1 and 2, given by a preferential usage of CGA and CGU for coding Arg, and a highest frequency of UUG in relation to CUG for coding Leu. TBFV clusters at not so extreme positive values for axis 1 and negative values for axis 2, which indicates a preference for AGA and AGG as Arg codons. The position of NKV is indicative of a relative low GC content, preference for AGA and AGG in relation to CGN codons, and a high usage of U/G ending codons. Again, both phylogenetic inertia and factors related to the host are probably at the basis of this clustering of groups. The position of each variable (i.e., codons) in relation to the three main axes is displayed in Supplementary Table 4.

### 3.4. Amino acids frequencies

The input for PCA in this case was the frequency of each amino acid in each virus. The three main axes represented 45%, 21% and 10% of the total variability, respectively. In Fig. 4a is displayed the 3D plot, while in Fig. 4b is shown the position of each amino acid in relation to the plane defined by the first two main axes. The first axis (Fig. 4b and c) is influenced principally by the high and low usage of Arg and Ile, respectively. Secondarily, high usages of Gly, Ala and Val, and low usages of Lys and Asn are also influencing axis 1. TBFV are clearly clustered by this axis. The proteome of *Ixodes* (see Supplementary Table 3), in comparison to the other hosts/vectors analyzed, is the richest in Arg (6.7%), Gly (6.7%), Ala (8.0%) and Val (7.1%), and the poorest in Ile (3.8%), Lys (4.9%) and Asn (3.3%). This axis also strongly correlates with the composition of each sequence. Indeed, the correlation coefficient with cGC is 0.85, while for GC1, GC2 and GC3 the r values are 0.90, 0.81 and 0.74, respectively. This again suggests the influence of tick genomics (and proteomics) composition over these same characteristics in TBFV. On the other hand, NKV shows a relatively low mean value for Arg and appears at the lower end of axis 1 (Fig. 4b and c). In this analysis, ISFV-2 and ISFV-3 cluster with MBFV (see Supplementary Fig. 1c).

The second axis of this analysis is influenced mainly by Val (Fig. 4d). ISFV gets apart from MBFV along this axis, having ISFV and TBFV the highest values for this amino acid (frequencies of 9.0% and 9.1%, respectively) (see Supplementary Table 2). These high frequencies of Val are not similar to the mean values in their vector, since the frequencies are: *Ixodes*, 7.1%; *Culex*, 6.5%; *Aedes*, 6.2%. The vertebrates present even lower percentages. MBFV and NKV, while lower than TBFV and ISFV, have means values higher than their vector and/or host. Hence, hydrophobicity of the encoded polyprotein could be a factor in the evolution of this genus. Finally, the third axis is mainly related to the high usage of Ala; NKV, with a lower frequency of this amino acid, and a high usage of Leu, is relatively apart along this axis.

Given these considerable differences in the use of Arg between TBFV and the other groups, it is indeed interesting to map their distribution across the genome. To address this subject, we divided each polyprotein into 100 partially overlapping fragments. The results presented in Fig. 5 show that most notorius differences between TBFV and the rest of the groups are observed in NS2A, NS2B and both NS5 ends. Moreover, proteins C, NS5 and the C-terminus end of NS3 are relatively rich in Arg, while E and NS4A and NS4B present a low use of this amino acid. The latter observation could be related to the hydrophobicity of these proteins, since NS4A/B are highly hydrophobic [36]. On the other hand, E is embedded in the phospholipid bilayer and therefore displays hydrophobic domains.

### 3.5. Concluding remarks

In this communication, we have shown that there is a mutual interplay between the evolutive history of the flaviviruses and their respective vectors and/or hosts. When the diversity of this genus was less known, the phylogeny and coevolutionary processes with their hosts were considered lineage-specific [31]. It seems that with the finding of both ISFV-2 and ISFV-3, some of these processes

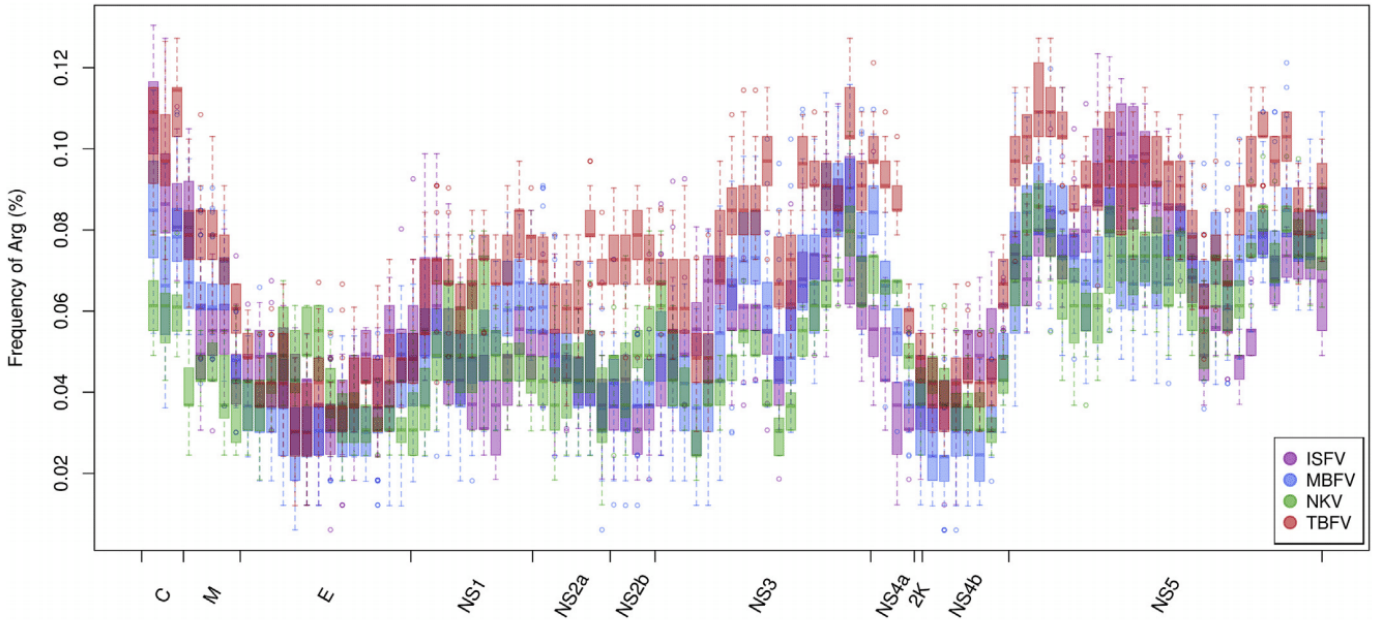



**Fig. 5.** Frequency of Arg by position across the polyprotein in overlapping windows.

should be considered as parallel events. However, the different hosts/vectors associated to these subgroups, may be driving them at different pace and through a different sequence landscape.

Recently, a new *Flavivirus* termed EPEV (Ecuador Paraiso Escondido Virus) was isolated from the phlebotomine *Psathyromyia abonnenci* [37]. EPEV is phylogenetically a MBFV, less divergent than ISFV-2 or ISFV-3 (see Fig. 1). Although Saboya virus (SABV) was previously isolated also from a phlebotomine in West Africa, its status as an arbovirus is well documented [38]. In contrast, the ability of EPEV to replicate in mosquito cells (C6/36) but not in vertebrate cell lines, suggests a vertebrate incompetent phenotype [37]. Coincidentally, EPEV appears near the ISFV-2 and/or ISFV-3 by many axes in the PCAs performed (see Supplementary Fig. 1).

To make the history of the genus *Flavivirus* even more complex, another group of viruses seems to have followed a different evolutionary path. This subgroup (denoted here as NKV-2) includes ENTV (Entebbe bat virus), SOKV (Sokoluk virus) and YOKV (Yokose virus) and is phylogenetically associated with MBFV (see Fig. 1). These viruses have been isolated uniquely from bats. The fact that unlike NKV species, ENTV and SOKV can replicate *in vitro* in C6/36 cells [39], as well as their phylogenetic location, supports the hypothesis that NKV-2 effectively infects a not yet known vector, or used to do it and eventually lost this capacity [6]. Moreover, SOKV has been isolated from ticks and birds [40], so it is clear that deeper studies should be performed in order to unravel the natural cycle of these species. For this subgroup, PCA failed to show a clear pattern, grouping them together with the MBFV. This could be another clue of the actual existence of an arthropod vector for this group, or their relatively recent lost.

TBFV is a singular group. All PCA performed show different sides of the forces that seem to shape their genomes. While the vertebrate immune system “pressures” them to deplete their CpG frequency, the tick environment facilitates their enrichment in GC-rich codons. Their high usage of Arg is in accordance with the amino acid frequencies found in the tick proteome.

In vertebrate viruses, the patterns described suggest the action of natural selection acting at the level of genomic composition to overcome immune system of the host. However, in the insect-only viruses the relative roles of natural selection and neutral forces like mutational biases remain an open question.

## Conflict of interest

Héctor Musto declare that we have no conflict of interest.

## Acknowledgments

Diego Simón thanks PEDECIBA and ANII (POS_NAC_2016_1_130463) for financial support. We thank the editor and reviewers for their constructive comments, which helped us to improve the manuscript.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.bbrc.2017.06.088.

## References

[1] M.J. Adams, E.J. Lefkowitz, A.M.Q. King, B. Harrach, R.L. Harrison, N.J. Knowles, A.M. Kropinski, M. Krupovic, J.H. Kuhn, A.R. Mushegian, M. Nibert, S. Sabanadzovic, H. Sanfaçon, S.G. Siddell, P. Simmonds, A. Varsani, F.M. Zerbini, A.E. Gorbalenya, A.J. Davison, Changes to Taxonomy and the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses, vol. 2017, 2017, http://dx.doi.org/10.1007/s00705-017-3358-5.

[2] B.D. Lindenbach, C.M. Rice, Molecular biology of flaviviruses, Adv. Virus Res. 59 (2003) 23–61.

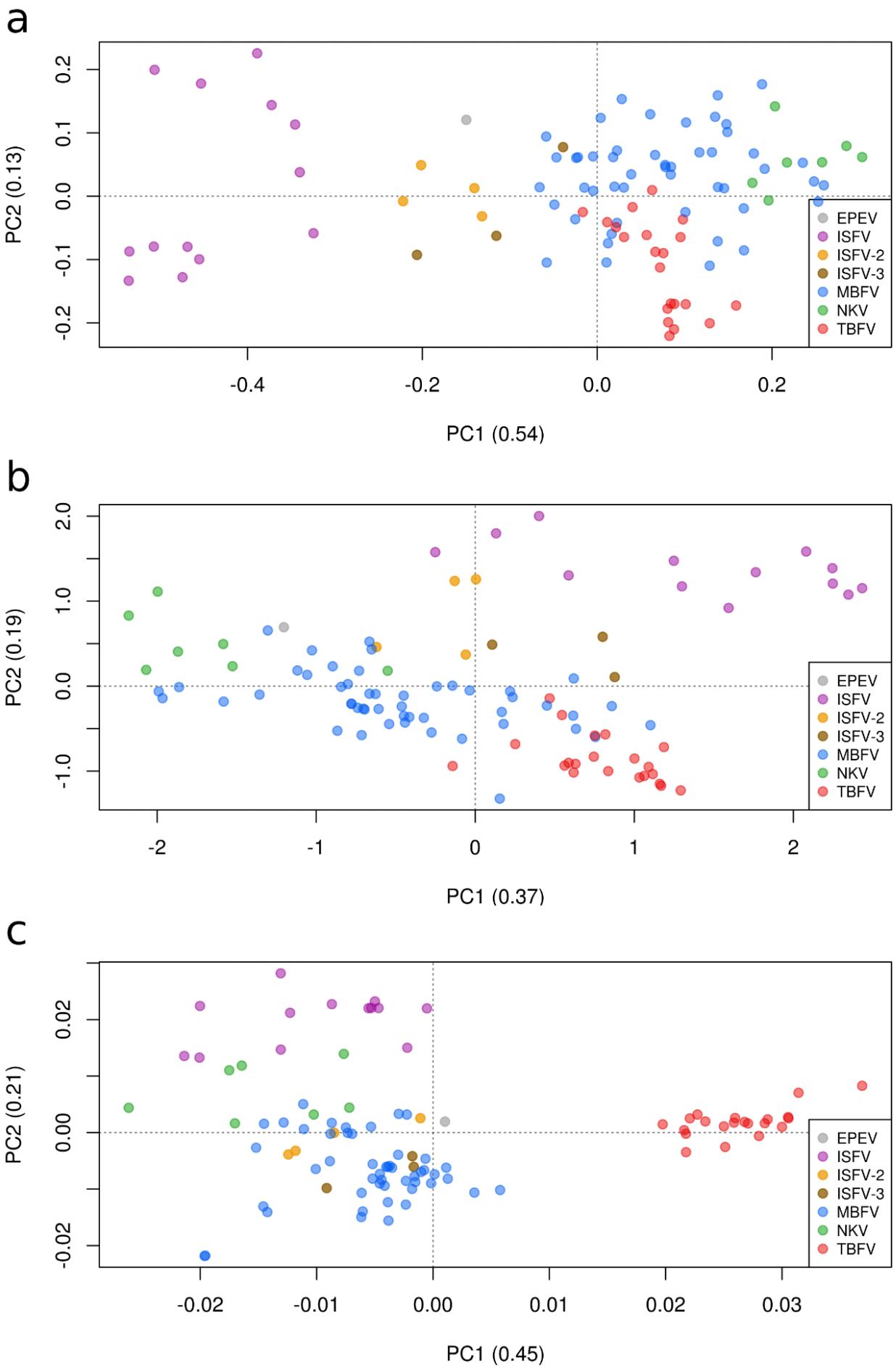
[3] B.J. Blitvich, A.E. Firth, Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization, Viruses 7 (2015), http://dx.doi.org/10.3390/v7041927.

[4] G. Moureau, S. Cook, P. Lemey, A. Nougairede, N.L. Forrester, M. Khasnatinov, R.N. Charrel, A.E. Firth, E.A. Gould, X. de Lamballerie, New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences, PLoS One 10 (2015) e0117849, http://dx.doi.org/10.1371/journal.pone.0117849.

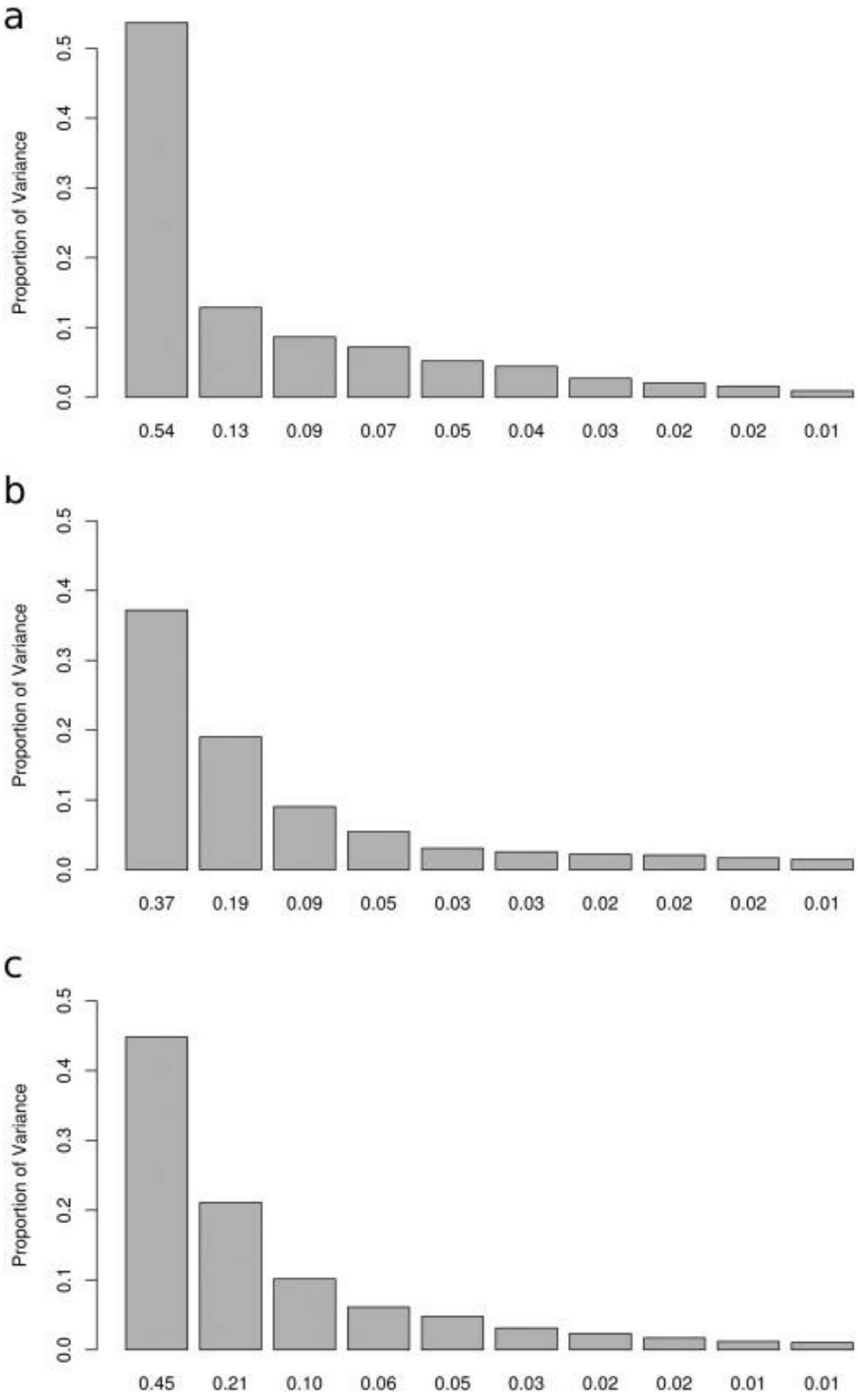
[5] G. Grard, G. Moureau, R.N. Charrel, E.C. Holmes, E.A. Gould, X. de Lamballerie, Genomics and evolution of Aedes-borne flaviviruses, J. Gen. Virol. 91 (2010) 87–94, http://dx.doi.org/10.1099/vir.0.014506-0.

[6] S. Cook, E.C. Holmes, A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: Flaviviridae) and the evolution of vector transmission, Arch. Virol. 151 (2006) 309–325, http://dx.doi.org/10.1007/

s00705-005-0626-6.
[7] M.W. Gaunt, A.A. Sall, X. de Lamballerie, A.K. Falconar, T.I. Dzhivanian, E.A. Gould, Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography, J. Gen. Virol. 82 (2001) 1867–1876, http://dx.doi.org/10.1099/0022-1317-82-8-1867.
[8] F. Billoir, R. de Chesse, H. Tolou, P. de Micco, E.A. Gould, X. de Lamballerie, Phylogeny of the genus flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector, J. Gen. Virol. 81 (2000) 781–790, http://dx.doi.org/10.1099/0022-1317-81-3-781.
[9] G. Kuno, G.-J.J. Chang, Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends, Clin. Microbiol. Rev. 18 (2005) 608–637, http://dx.doi.org/10.1128/CMR.18.4.608-637.2005.
[10] A.T. De Madrid, J.S. Porterfield, The flaviviruses (group B arboviruses): a cross-neutralization study, J. Gen. Virol. 23 (1974) 91–96, http://dx.doi.org/10.1099/0022-1317-23-1-91.
[11] C.H. Calisher, N. Karabatsos, J.M. Dalrymple, R.E. Shope, J.S. Porterfield, E.G. Westaway, W.E. Brandt, Antigenic relationships between flaviviruses as determined by cross-neutralization tests with polyclonal antisera, J. Gen. Virol. 70 (Pt 1) (1989) 37–43, http://dx.doi.org/10.1099/0022-1317-70-1-37.
[12] E.A. Gould, X. de Lamballerie, P.M. Zanotto, E.C. Holmes, Origins, evolution, and vector/host coadaptations within the genus Flavivirus, Adv. Virus Res. 59 (2003) 277–314.
[13] M.S. Marin, P.M. Zanotto, T.S. Gritsun, E.A. Gould, Phylogeny of TYU, SRE, and CFA virus: different evolutionary rates in the genus Flavivirus, Virology 206 (1995) 1133–1139.
[14] G. Kuno, G.J. Chang, K.R. Tsuchiya, N. Karabatsos, C.B. Cropp, Phylogeny of the genus flavivirus, J. Virol. 72 (1998) 73–83.
[15] B.E. Pickett, E.L. Sadat, Y. Zhang, J.M. Noronha, R.B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, L. Zhou, C.N. Larson, J. Dietrich, E.B. Klem, R.H. Scheuermann, ViPR: an open bioinformatics database and analysis resource for virology research, Nucleic Acids Res. 40 (2012) D593–D598, http://dx.doi.org/10.1093/nar/gkr859.
[16] X. de Lamballerie, S. Crochu, F. Billoir, J. Neyts, P. de Micco, E.C. Holmes, E.A. Gould, Genome sequence analysis of Tamana bat virus and its relationship with the genus Flavivirus, J. Gen. Virol. 83 (2002) 2443–2454, http://dx.doi.org/10.1099/0022-1317-83-10-2443.
[17] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797, http://dx.doi.org/10.1093/nar/gkh340.
[18] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Mol. Biol. Evol. 17 (2000) 540–552.
[19] G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, Syst. Biol. 56 (2007) 564–577, http://dx.doi.org/10.1080/10635150701472164.
[20] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, ProtTest 3: fast selection of best-fit models of protein evolution, Bioinformatics 27 (2011) 1164–1165, http://dx.doi.org/10.1093/bioinformatics/btr088.
[21] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Syst. Biol. 59 (2010) 307–321, http://dx.doi.org/10.1093/sysbio/syq010.
[22] A. Yates, W. Akanni, M.R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C.G. Girón, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F.J. Martin, T. Maurel, W. McLaren, D.N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H.S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S.P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S.J. Trevanion, F. Cunningham, B.L. Aken, D.R. Zerbino, P. Flicek, Ensembl 2016, Nucleic Acids Res. 44 (2016) D710–D716, http://dx.doi.org/10.1093/nar/gkv1157.
[23] D. Charif, L. Humblot, J.R. Lobry, A. Necsulea, L. Palmeira, S. Penel, SeqinR 2.0-1: a Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrievel and Analysis, vol. 268, 2008.
[24] R Core Team, R: a Language and Environment for Statistical Computing, 2016.
[25] U. Ligges, M. Mächler, scatterplot3d - an R package for visualizing multivariate data, J. Stat. Softw. 8 (2003) 1–20, http://dx.doi.org/10.18637/jss.v008.i11.
[26] G.M. Jenkins, M. Pagel, E. Gould, P.M.A. Zanotto, E.C. Holmes, Evolution of base composition and codon usage bias in the genus Flavivirus, J. Mol. Evol. 52 (2001) 383–390, http://dx.doi.org/10.1007/s002390010168.
[27] A.M. Schubert, C. Putonti, Evolution of the sequence composition of Flaviviruses, Infect. Genet. Evol. 10 (2010) 129–136, http://dx.doi.org/10.1016/j.meegid.2009.11.004.
[28] E. Huhtamo, N. Putkuri, S. Kurkela, T. Manni, A. Vaheri, O. Vapalahti, N.Y. Uzcategui, Characterization of a novel flavivirus from mosquitoes in northern Europe that is related to mosquito-borne flaviviruses of the tropics, J. Virol. 83 (2009) 9532–9540, http://dx.doi.org/10.1128/JVI.00529-09.
[29] S. Junglen, A. Kopp, A. Kurth, G. Pauli, H. Ellerbrok, F.H. Leendertz, A new flavivirus and a new vector: characterization of a novel flavivirus isolated from uranotaenia mosquitoes from a tropical rain forest, J. Virol. 83 (2009) 4462–4468, http://dx.doi.org/10.1128/JVI.00014-09.
[30] J.L. Kenney, O.D. Solberg, S.A. Langevin, A.C. Brault, Characterization of a novel insect-specific flavivirus from Brazil: potential for inhibition of infection of arthropod cells with medically important flaviviruses, J. Gen. Virol. 95 (2014), http://dx.doi.org/10.1099/vir.0.068031-0.
[31] F.P. Lobo, B.E.F. Mota, S.D.J. Pena, V. Azevedo, A.M. Macedo, A. Tauch, C.R. Machado, G.R. Franco, Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts, PLoS One 4 (2009), http://dx.doi.org/10.1371/journal.pone.0006282.
[32] X. Cheng, N. Virk, W. Chen, S. Ji, S. Ji, Y. Sun, X. Wu, CpG usage in RNA viruses: data and hypotheses, PLoS One 8 (2013), http://dx.doi.org/10.1371/journal.pone.0074109.
[33] N.J. Atkinson, J. Witteveldt, D.J. Evans, P. Simmonds, The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication, Nucleic Acids Res. 42 (2014) 4527–4545, http://dx.doi.org/10.1093/nar/gku075.
[34] M. Schlee, G. Hartmann, Discriminating self from non-self in nucleic acid sensing, Nat. Rev. Immunol. 16 (2016) 566–580, http://dx.doi.org/10.1038/nri.2016.78.
[35] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1986) 28–38, http://dx.doi.org/10.1007/BF02099948.
[36] N.J. da Fonseca, M.Q. Lima Afonso, N.G. Pedersolli, L.C. de Oliveira, D.S. Andrade, L. Bleicher, Sequence, structure and function relationships in flaviviruses as assessed by evolutive aspects of its conserved non-structural protein domains, Biochem. Biophys. Res. Commun. 492 (2017) 565–571, http://dx.doi.org/10.1016/j.bbrc.2017.01.041.
[37] C. Alkan, S. Zapata, L. Bichaud, G. Moureau, P. Lemey, A.E. Firth, T.S. Gritsun, E.A. Gould, X. de Lamballerie, J. Depaquit, R.N. Charrel, Ecuador Paraiso Escondido virus, a new flavivirus isolated from new World sand flies in Ecuador, is the first representative of a novel clade in the genus flavivirus, J. Virol. 89 (2015) 11773–11785, http://dx.doi.org/10.1128/JVI.01543-15.
[38] D. Fontenille, M. Traore-Lamizana, J. Trouillet, A. Leclerc, M. Mondo, Y. Ba, J.P. Digoutte, H.G. Zeller, First isolations of arboviruses from phlebotomine sand flies in West Africa, Am. J. Trop. Med. Hyg. 50 (1994) 570–574.
[39] I. Varelas-Wesley, C.H. Calisher, Antigenic relationships of flaviviruses with undetermined arthropod-borne status, Am. J. Trop. Med. Hyg. 31 (1982) 1273–1284.
[40] D.K. L'vov, S. V Al'khovskiĭ, M.I. Shchelkanov, A.M. Shchetinin, P.G. Deriabin, A.K. Gitel'man, E.I. Samokhvalov, A.G. Botikov, [Taxonomy of the Sokuluk virus (SOKV) (Flaviviridae, Flavivirus, Entebbe bat virus group) isolated from bats (Vespertilio pipistrellus Schreber, 1774), ticks (Argasidae Koch, 1844), and birds in Kyrgyzstan]., Vopr. Virusol. 59 (n.d.) 30–34.




**Supplementary Fig. 1**. Position of the viruses according to PC1 and PC2 for (**a**) dinucleotide biases, (**b**) relative synonymous codon usage and (**c**) amino acid frequencies.




**Supplementary Fig. 2**. Proportion of Variance of the ten first principle components for (**a**) dinucleotide biases, (**b**) relative synonymous codon usage and (**c**) amino acid frequencies.

Contribución

El principal aporte de este trabajo fue actualizar lo observado en la historia evolutiva de los flavivirus con respecto a su composición genómica con un número mucho mayor de especies virales (98). Cuando la diversidad de este género era menos conocida, la filogenia y los procesos coevolutivos con sus hospedadores se consideraban que eran específicos para cada linaje (i.e., MBFV, TBFV, NKV y ISFV). Pero con el hallazgo de los ISFV-2 y los ISFV-3 (grupos de virus específicos de insectos polifiléticos con respecto a los «ISFV-1»), algunos de estos eventos deberían considerarse como paralelos. Los diferentes hospederos y/o vectores asociados a estos subgrupos se enfrentan a «entornos» diferentes, que proporcionan distintos paisajes adaptativos.

Para hacer aún más compleja la historia evolutiva del género, otro grupo de virus parece haber seguido un camino evolutivo diferente. Este subgrupo (denotado aquí como NKV-2) es también polifilético con respecto a los «NKV-1», estando asociados filogenéticamente con los MBFV.

También destacamos el comportamiento particular de los TBFV. Mientras que el sistema inmunitario de los vertebrados les «presiona» para que disminuyan su frecuencia del dinucleótido CpG, el entorno de las garrapatas facilita su enriquecimiento en codones ricos en G+C (dado que son invertebrados con alto contenido de G+C). Además, su elevado uso de arginina en estos virus está en consonancia con las frecuencias de aminoácidos encontradas en el proteoma de las garrapatas.

En los virus de vertebrados, los patrones descritos sugieren la acción de la selección natural que actúa a nivel de la composición genómica para superar al sistema inmunitario del hospedero.

**Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts**

«Composición nucleotídica y uso de codones en los virus y en sus respectivos hospederos»

Abordaje

Se ha propuesto que los virus pueden ser tan antiguos como la vida en la Tierra. Aunque se ha trabajado mucho para comprender el origen y la evolución de los virus y, en particular, de sus diferentes materiales genéticos, todavía está faltando tener una imagen global. Uno de los enfoques más sencillos para estudiar los organismos y la relación entre ellos es el análisis de las respectivas «firmas genómicas», que pueden ir desde la simple composición de bases como el contenido molar de guanina y de citosina (G+C), dinucleótidos, codones y aminoácidos.

Los estudios filogenéticos previos realizados en diferentes virus han destacado a la presión mutacional como el principal factor que determina la evolución de los virus en comparación con la selección natural. Sin embargo, a medida que aumenta nuestra comprensión de la evolución de los virus, parece que aunque la presión mutacional sigue siendo una fuerza importante, no es el único factor cuando se consideran diferentes virus de ADN y de ARN.

Además, la composición del genoma viral también puede estar relacionada con la interacción virus-hospedero. Por ejemplo, al evitar el reconocimiento por parte del sistema inmunitario innato de vertebrados, lo que provoca fuertes presiones selectivas que pueden dejar firmas genómicas típicas de sus hospederos.

Este trabajo incorpora unas 10.000 especies virales y casi 1.200 hospederos de los tres dominios de la vida (*Archaea*, *Bacteria*, y *Eukaryota*). Una visión global centrada en la composición genómica de todos los virus era relevante dado el impresionante aumento de la disponibilidad de secuencias virales en los últimos años.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fmicb.2021.646300

# Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts

*Diego Simón*[1,2,3], *Juan Cristina*[2] *and Héctor Musto*[1]*

[1] *Laboratorio de Genómica Evolutiva, Departamento de Biología Celular y Molecular, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay,* [2] *Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la Republica, Montevideo, Uruguay,* [3] *Laboratorio de Evolución Experimental de Virus, Institut Pasteur de Montevideo, Montevideo, Uruguay*

The genetic material of the three domains of life (Bacteria, Archaea, and Eukaryota) is always double-stranded DNA, and their GC content (molar content of guanine plus cytosine) varies between ≈ 13% and ≈ 75%. Nucleotide composition is the simplest way of characterizing genomes. Despite this simplicity, it has several implications. Indeed, it is the main factor that determines, among other features, dinucleotide frequencies, repeated short DNA sequences, and codon and amino acid usage. Which forces drive this strong variation is still a matter of controversy. For rather obvious reasons, most of the studies concerning this huge variation and its consequences, have been done in free-living organisms. However, no recent comprehensive study of all known viruses has been done (that is, concerning all available sequences). Viruses, by far the most abundant biological entities on Earth, are the causative agents of many diseases. An overview of these entities is important also because their genetic material is not always double-stranded DNA: indeed, certain viruses have as genetic material single-stranded DNA, double-stranded RNA, single-stranded RNA, and/or retro-transcribing. Therefore, one may wonder if what we have learned about the evolution of GC content and its implications in prokaryotes and eukaryotes also applies to viruses. In this contribution, we attempt to describe compositional properties of ∼ 10,000 viral species: base composition (globally and according to Baltimore classification), correlations among non-coding regions and the three codon positions, and the relationship of the nucleotide frequencies and codon usage of viruses with the same feature of their hosts. This allowed us to determine how the base composition of phages strongly correlate with the value of their respective hosts, while eukaryotic viruses do not (with fungi and protists as exceptions). Finally, we discuss some of these results concerning codon usage: reinforcing previous results, we found that phages and hosts exhibit moderate to high correlations, while for eukaryotes and their viruses the correlations are weak or do not exist.

**Keywords: viral diversity, base composition, GC-content, compositional correlations, codon usage**

**Abbreviations:** diNs, dinucleotides; ds, double-stranded; dsDNA, double-stranded DNA; dsDNA-RT, double-stranded DNA retro-transcribing; dsRNA, double-stranded RNA; GC, guanine plus cytosine; GC1, guanine plus cytosine content of first codon position; GC2, guanine plus cytosine content of second codon position; GC3, guanine plus cytosine content of third codon position; ρ (rho), Spearman's correlation coefficient; ss, single-stranded; ssDNA, single-stranded DNA; -ssRNA, negative single-stranded RNA; +ssRNA, positive single-stranded RNA; +ssRNA-RT, positive single-stranded RNA retro-transcribing.

OPEN ACCESS

***Edited by:***
*Rosa Maria Pintó, University of Barcelona, Spain*

***Reviewed by:***
*Vladislav Victorovich Khrustalev, Belarusian State Medical University, Belarus*
*Gwenael Piganeau, UMR 7232 Biologie Intégrative des Organismes Marins (BIOM), France*

***Correspondence:**
*Héctor Musto*
*hmusto@gmail.com*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *26 December 2020*
***Accepted:*** *04 June 2021*
***Published:*** *28 June 2021*

***Citation:***
*Simón D, Cristina J and Musto H (2021) Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. Front. Microbiol. 12:646300. doi: 10.3389/fmicb.2021.646300*

## INTRODUCTION

Viruses are obligate parasites of all free cellular life forms and are, at the same time, the most abundant biological entities on Earth (Cobián Güemes et al., 2016). To understand the relationship among different viruses several distinct approaches have been used (Krupovic et al., 2019), given: (i) the diversity of the architecture of their genetic material, which can be DNA or RNA, double-stranded (ds) or single-stranded (ss), linear or circular, segmented or not; (ii) the huge variation of their size (from very tiny particles of around 10 nm with genomes of only a few kb, to giant viruses that reach 1.5 μm and genomes of up to 2.5 Mb that fall into the genome and particle size ranges typical of Bacteria and Archaea); and (iii) since there are not orthologous genes shared by all viruses, it is universally accepted that these biological entities appeared several times in the course of evolution (Koonin et al., 2006; Holmes, 2011; Durzyńska and Goździcka-Józefiak, 2015; Krupovic et al., 2019). Although a lot of work has been done in order to understand the origin and evolution of viruses, and in particular, of their different genetic materials, a complete picture still lacks. One of the simplest approaches for studying organisms and the relationship among them is analyzing the respective "genomic signatures," which can go from simple base composition as molar content of guanine plus cytosine (GC content), dinucleotides (diNs), and codon and amino acid usage.

Previous phylogenetic studies carried out in different viruses have high-lighted mutational pressure as the major factor in shaping virus evolution in comparison with natural selection (Jenkins and Holmes, 2003; Gu et al., 2004). Nevertheless, as our understanding of virus evolution increases, it appears that although mutational pressure is still a major driving force, it is not the only factor when considering different RNA and DNA viruses (Berkhout and van Hemert, 1994; Chen, 2013; Kustin and Stern, 2021). Moreover, viral genome composition may also be related to virus-host interaction, for instance, by avoiding recognition by the innate immune system (van Hemert et al., 2014). This could provide strong selective pressures, leaving genomic signatures typical of their hosts, both at the nucleotide (Simón et al., 2017) and structural levels (Kindler and Thiel, 2014).

In prokaryotes and eukaryotes, the analyses of these features have led to several conclusions, and perhaps the more relevant for our current purpose can be summarized as follows: (i) base composition is generally more similar within phylogenetically close groups and species living in the same –or very similar– environment (Foerstner et al., 2005; Agashe and Shankar, 2014; Reichenberger et al., 2015), (ii) for prokaryotes, GC content strongly correlates with the mean values for GC1, GC2, and GC3 (that is, the GC content of the three codon positions) for each organism, and also with the global diNs frequencies and amino acid usage (Zhou et al., 2014), (iii) although the variability in genomic GC among prokaryotes is high, within genomes they are remarkably homogeneous (Bohlin and Pettersson, 2019), thought "protoisochores" were found in some Archaea (Khrustalev and Barkovsky, 2011). But on the contrary, (iv) vertebrate genomes (mainly those of mammals and birds) display large contiguous regions characterized by very similar GC content which are termed isochores (Bernardi et al., 1985; Eyre-Walker and Hurst, 2001; Costantini and Musto, 2017), and each of these isochores display a particular and very similar pattern of codon usage (Costantini et al., 2009) and amino acid frequencies (Sabbia et al., 2007), although intragenic GC content heterogeneity has been noted in birds (Khrustalev et al., 2014). Among unicellular eukaryotes, it has been shown that most of them are compositionally heterogeneous (Costantini et al., 2013) as is the case in some flatworms (Lamolle et al., 2016). Therefore, from the study of the genomic composition important features like diNs frequencies and codon usage have been derived, and helped us to understand important biological properties, like patterns of synonymous and non-synonymous substitutions, and the relative effects of neutral and selective forces driving these changes (Pracana et al., 2020).

However, although some recent publications have analyzed several viruses (see, for example, Auewarakul, 2005; Duffy et al., 2008; Mahmoudabadi and Phillips, 2018), an overview focusing on the genomic composition of all viruses is relevant given the impressive increase in viral sequences availability in the last years. In this report, we present the following analyses: (i) base frequencies of all available viruses, (ii) the same feature but sorting viruses according to the Baltimore classification: dsDNA, ssDNA, dsRNA, positive ssRNA (+ssRNA), negative ssRNA (-ssRNA), +ssRNA retro-transcribing (+ssRNA-RT), and dsDNA retro-transcribing (dsDNA-RT), (iii) besides, we analyzed the correlations that hold between the non-coding GC content vs. GC1, GC2 and GC3, (iv) for each group we studied the GC content variation of the viral genomes compared to that of the respective host, and (v) finally, we analyzed codon usage patterns among viruses in relation to the same features of their hosts.

Our main conclusions are that: (i) different viruses (according to the nature and architecture of the respective genetic material), show different properties at their base composition; (ii) there are strong compositional correlations among non-coding regions and the three codon positions; (iii) while GC content of phages strongly correlates with the genomic GC of their hosts, this is not the case for eukaryotic systems; and (iv) in general, the codon usage of phages is dependent of the codon usage of prokaryotes, while the codon usage of animal and plant viruses do not seem to be adapted to the codon usage of their hosts, with the probable exception of fungi and protists.

## MATERIALS AND METHODS

Sequences were retrieved from NCBI RefSeq viral genomes, Release 205, accessed at ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/ (Brister et al., 2015). Each viral species was included only once to avoid the overrepresentation of viruses for which there are multiple sequences. For this purpose, only one representative was considered for each viral species (i.e., one representative per taxonomy identifier, TaxID) in this taxonomic rank ($N$ = 9,994; see **Table 1** and **Supplementary Table 1**). In the case of segmented viruses, we use global compositional values to summarize these genomes.

Compositional features for non-coding regions and coding GC content per codon position (i.e., GC1, GC2, and GC3), were calculated for genomic regions extracted with BEDTools (Quinlan and Hall, 2010). Host GC contents were scrapped from NCBI Genomes website accessed at https://www.ncbi.nlm.nih.gov/genome (Benson et al., 2017). Codon usage tables were retrieved from HiVE's CoCoPUTs database (Alexaki et al., 2019).

Virus-host relationships were obtained from Virus-Host Database, accessed at https://www.genome.jp/virushostdb (Mihara et al., 2016). In **Table 2** is displayed the diversity of hosts represented in this study; it must be taken into account that the same host will have several viruses assigned to it, while the same virus may be assigned, in some cases, to more than one host. In total, this part of the study included 8,411 host-virus pairs (see **Supplementary Material**).

The base composition distributions were drawn using kernel density plots with default bandwidths. To test for unimodality/multimodality, Hartigans' dip tests were performed. The Spearman's rank correlation coefficient (ρ) was chosen to measure the strength of a linear association between variables. The adjusted $R^2$ ($adjR^2$) coefficient was used to access the goodness of fit of linear regression models to the data. All these computations were implemented in R v4.0.[0-5] (R Core Team, 2020). Figures were constructed in RStudio v1.3.1073 (RStudio Team., 2020) using RColorBrewer v1.1-2 (Neuwirth, 2014).

# RESULTS

## Base Composition

In **Table 1** are displayed the number of all the viral sequences we have analyzed, sorted by Baltimore classification. In **Figure 1A** is displayed the genomic GC content of all these sequences. It can be seen that the distribution of the genomic GC ranges from 18% to 77%. Furthermore, it is non-unimodal (Hartigans' dip test, *p*-value = < 0.0001) displaying two modes: a major at a GC of 43% and a minor at 62%. This distribution also presents three shoulders at ≈ 30%, 36%, and 49%, being the latter more evident than the others.

In **Figures 1B–H** are displayed the base composition (i.e., GC content for ds and nucleotide frequencies for ss) of the viruses studied here, sorted by Baltimore classification. In **Figure 1B** it can be seen that the GC distribution of dsDNA viruses exhibit a multimodal distribution (Hartigans' dip test, *p*-value ≈ 0), with three modes at 39%, 51% and 63%. While the value of 39% is representative of the whole sample (see **Figure 1A**), the other two peaks are due to the overrepresentation of *Escherichia* and *Mycobacterium* bacteriophages. Regarding the range of this distribution, minimum and maximum values were the same for this group as for the complete set of viruses. Thus, the extreme GC values occur within this group.

In **Figures 1D,H** are plotted the GC content of the other viruses which display double-stranded genomes: dsRNA (**Figure 1D**) and dsDNA-RT (**Figure 1H**). The former shows a unimodal distribution with a mode at 46% and displays two shoulders located at GC values of 38% and 58%, respectively. In the case of dsDNA-RT, it shows a symmetrical distribution, peaking at a GC of 43% and with two bumps at 37% and 48%.

The other group of retro-transcribing viruses, +ssRNA-RT, tends to present bimodal distributions in all four bases (**Figure 1G**), as is the case for GC content (**Supplementary Figure 1D**; Hartigans' dip test, *p*-value < 0.01). In **Figures 1C,E,F** are plotted the remaining single-stranded genomes. Overall, C is the less frequent base, which reflects the process of cytosine deamination which leads to thymine or uracil. This is reinforced by the fact that in ssDNA viruses, T is the most frequent base. In the case of ssRNA viruses, U is the second base in frequency. Furthermore, in these entities, A is the most abundant nucleotide. Taken globally, for all these cases, A and U(T) are the most frequent bases.

## Compositional Correlations

As happens in prokaryotes and most parasitic or symbiotic unicellular eukaryotes, for viruses protein-coding regions make up the majority of their genomes. In summary, only 9% (median) of a viral genome is not transcribed and translated. However, these regions are usually highly structured and encode *cis*-acting elements. Despite this, non-coding and genomic GC display a very high correlation (ρ = 0.86).

In **Figure 2** are shown the compositional correlations that hold between GC1 (**Figure 2A**), GC2 (**Figure 2B**), and GC3 (**Figure 2C**) with the non-coding GC content of the corresponding virus. These compositional correlations are, in all cases, positive and highly significant (*p*-values ≈ 0). The Spearman correlation coefficients between non-coding GC and GC1, GC2, and GC3 are 0.76, 0.77, and 0.77, respectively. Also, they present big differences in the slopes: 0.57 (GC1), 0.41 (GC2), and 1.37 (GC3). The correlations that hold between non-coding regions and GC1, GC2, and GC3 in viruses sorted by Baltimore classification are displayed in **Table 3**.

Besides these compositional correlations, inherent to each viral genome, it is of great interest to search for putative dependencies with respect to their hosts. This is displayed in **Figure 3A** which shows that there is a linear correlation of viral GC content in relation to their respective host genomic GC, with a Spearman correlation coefficient of 0.61. Furthermore, the GC content of phages strongly correlates to their host values; see **Figure 3B** (ρ = 0.89; *n* = 3,697 host-phage pairs). This holds when

**TABLE 1 |** The total number of viruses analyzed and within each Baltimore classification group.

| Total* | dsDNA | ssDNA | dsRNA | +ssRNA | −ssRNA | +ssRNA-RT | dsDNA-RT |
|---|---|---|---|---|---|---|---|
| 9,994 | 4,165 | 1,951 | 388 | 1,551 | 621 | 78 | 107 |

**The total number (N = 9,994) does not match the sum of Baltimore classification groups (N = 8,861) because for some viruses the nature of the genetic material and/or strandedness remains unknown.*

**TABLE 2 |** The total number of hosts represented in this study and within each taxonomic group considered.

| Total | Animals | Archaea | Bacteria | Fungi | Plants | Protists |
|---|---|---|---|---|---|---|
| 1,170 | 378 | 31 | 486 | 72 | 181 | 22 |

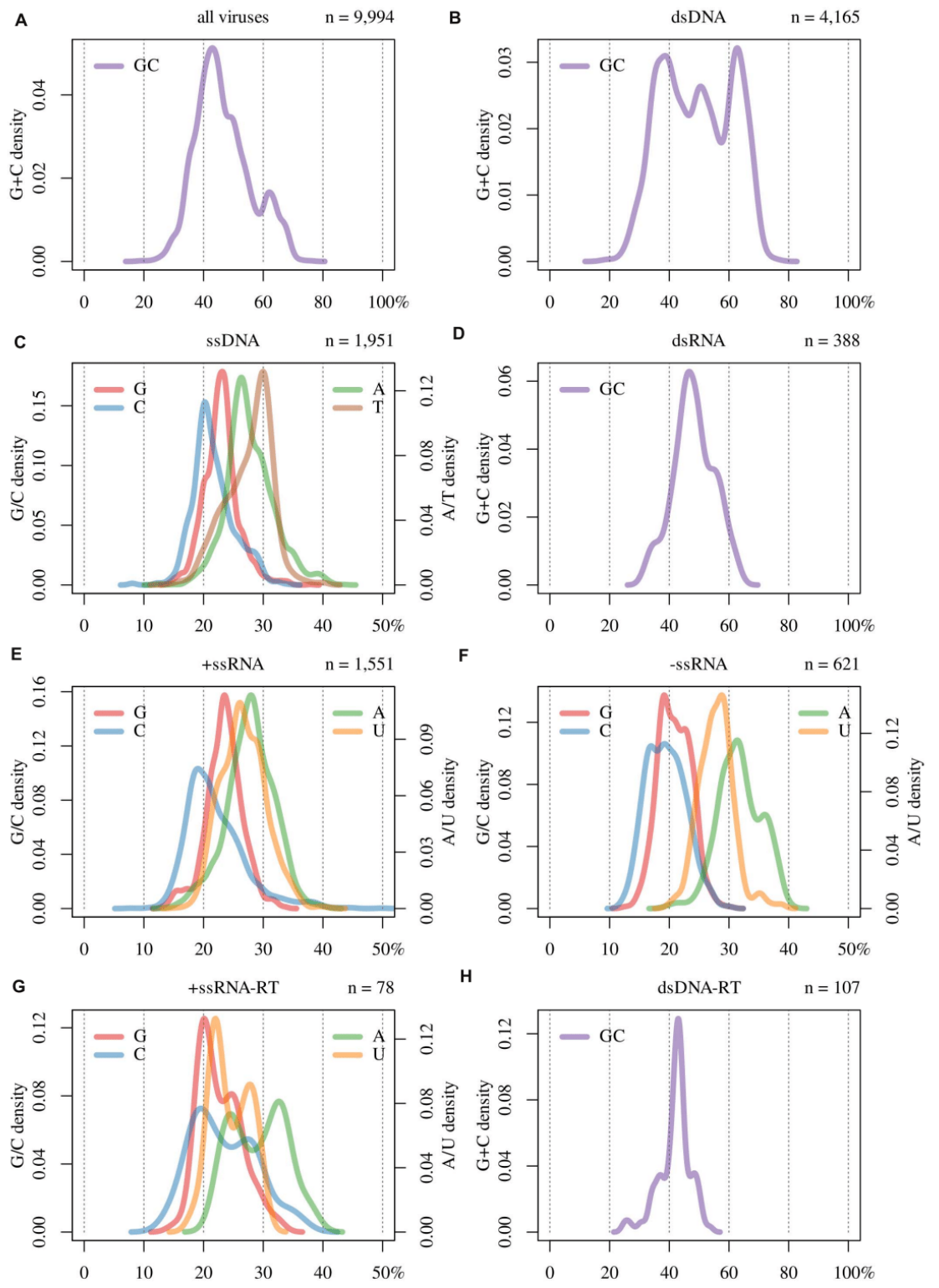



**FIGURE 1 |** Base composition of **(A)** all viruses, and by Baltimore classification groups **(B–H)**; i.e., **(B)** double-stranded DNA (dsDNA), **(C)** single-stranded DNA (ssDNA), **(D)** double-stranded RNA (dsRNA), **(E)** positive single-stranded RNA (+ssRNA), **(F)** negative single-stranded RNA (-ssRNA), **(G)** +ssRNA retro-transcribing (+ssRNA-RT), and **(H)** dsDNA retro-transcribing (dsDNA-RT).

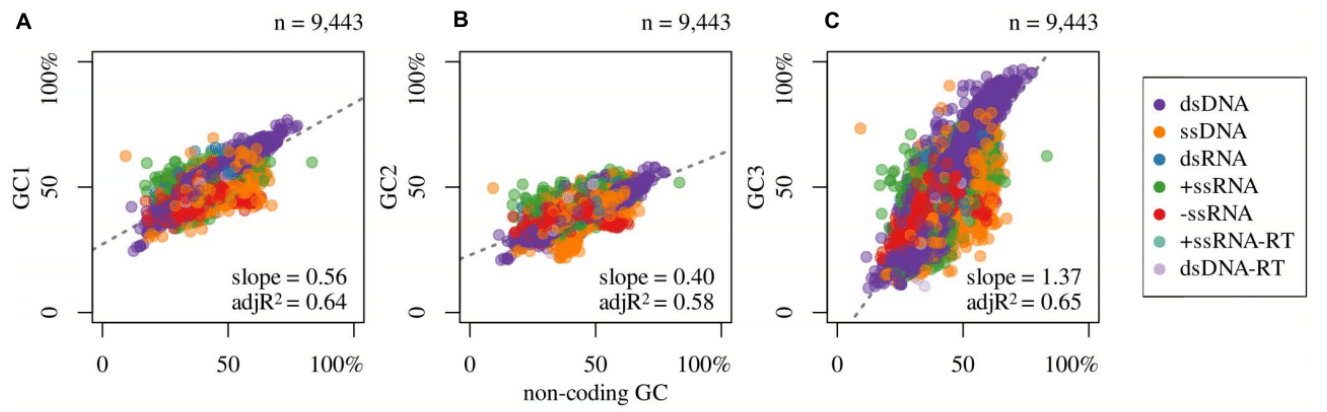



**FIGURE 2 |** Association between GC content of non-coding regions and **(A)** GC1, **(B)** GC2, and **(C)** GC3, and the slope and adjusted $R^2$ ($adjR^2$) of the linear regression model (dashed line) for all colored dots (regardless of color).

considering separately Bacteria ($\rho = 0.90$, $n = 3{,}629$) or Archaea ($\rho = 0.81$, $n = 68$). It is interesting to note that most phages display lower GC values than their hosts. This is noticeable in **Figure 3B**, since a major proportion of blue and purple dots (prokaryotes) are placed below the 1:1 diagonal.

Contrary to what is provided for prokaryotes, eukaryotic viruses show a very weak correlation between their GC values and that of their hosts; see **Figure 3C** ($\rho = 0.19$; $n = 4{,}642$ host-virus pairs). This figure represents the relationship of eukaryotes and their viruses, colored by eukaryotic subgroup (i.e., animals, plants, fungi, and protists). No meaningful correlation exist between viruses and animals ($\rho = 0.14$, $n = 2{,}691$) or plants and their viruses ($\rho = 0.09$, $n = 1{,}672$). Conversely, fungi and mycoviruses (i.e., viruses that infect fungi), do present a moderate positive correlation ($\rho = 0.43$, $n = 218$). Protists and their viruses exhibit a negative correlation ($\rho = -0.48$, $n = 61$), which, although moderate, is a polarizing result.

## Codon Usage

Given the pattern described above regarding GC content, we further analyzed the relationship between codon usage of viruses in relation to that of their hosts. In **Table 4** are displayed the Spearman correlation coefficients for each codon between viruses and hosts. For prokaryotes, all 64 codons show positive correlations between phages and their hosts ($\rho$ values ranging from 0.13 to 0.92) with a median of 0.73, while for eukaryotes the median is 0.08 (ranging from −0.13 to 0.28). Moreover, all but one of the $\rho$ values for phages and their hosts are stronger than any case for eukaryotic system, with the sole exception of the codon CGA ($\rho = 0.13$). The median $adjR^2$ also captures these strong differences between phages and eukaryotic viruses: 0.52, and less than 0.01, respectively.

**TABLE 3 |** Spearman's rank correlation coefficients between non-coding regions and GC1, GC2, and GC3, when available within viral genomes, sorted by Baltimore classification group.

| Baltimore | GC1 | GC2 | GC3 | n |
|---|---|---|---|---|
| dsDNA | 0.97 | 0.94 | 0.94 | 3,884 |
| ssDNA | 0.47 | 0.56 | 0.42 | 1,881 |
| dsRNA | 0.57 | 0.60 | 0.57 | 352 |
| +ssRNA | 0.50 | 0.58 | 0.52 | 1,486 |
| –ssRNA | 0.61 | 0.54 | 0.71 | 597 |
| +ssRNA-RT | 0.51 | 0.60 | 0.69 | 76 |
| dsDNA-RT | 0.48 | 0.48 | 0.65 | 105 |

## DISCUSSION

The most basic approach for characterizing genomes is analyzing the genomic base composition. Although the collective distribution (i.e., utilizing all available viral species fully sequenced), shown in **Figure 1A**, was statistically bimodal, it presents a major mode that is pervasive in the remaining distributions (**Figures 1B–H** and **Supplementary Figure 1**). Certainly, this distribution is biased by dsDNA viruses (**Figure 1B**), which are predominant in the available data set (**Table 1**), as the more evident shoulder at 49% and the minor mode at 62% are due to the overrepresentation of phages infecting *Escherichia* and *Mycobacterium* genera, respectively. Despite the previous points, we hypothesize that the maximum of the distribution (GC content peaking at 43%) will not change significantly, as will not the minimum and maximum values. We postulate this latter point, given the nature of the genetic code and the correlations that hold between the global GC content and GC1, GC2, and GC3 (see below). Indeed, these two factors impose constraints on codon usage and on the frequencies of the amino acids that can be coded by each virus (Li et al., 2015).

In this study, we have shown that when sorting viruses according to Baltimore classification, several differences among them are apparent. A singular behavior is seen in the case of dsDNA viruses. While unimodal distributions are found in dsRNA and dsDNA-RT (**Figure 1D,H**), a trimodal distribution is evident for dsDNA viruses (**Figure 1B**). As shown in **Table 1**, this group is very numerous, and therefore the distribution shown

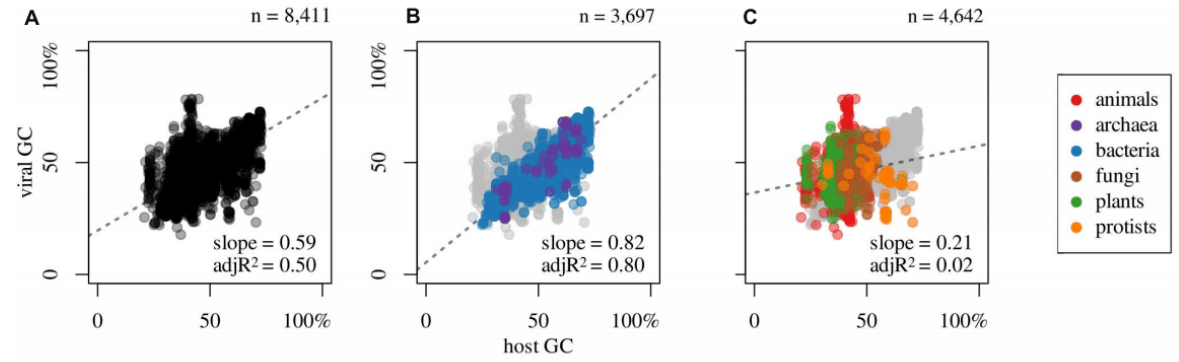



**FIGURE 3 |** Association between GC content of hosts and that of their infecting viruses for **(A)** all host-virus pairs, **(B)** prokaryotic, and **(C)** eukaryotic, and slope and adjusted $R^2$ ($adjR^2$) of the linear regression model (dashed line) for all colored dots (regardless of color), omitting pairs that do not apply; i.e., gray dots at the background: **(B)** eukaryotics or **(C)** prokaryotics.

here is probably robust. This trimodality is due to the adaptation of the GC content between these viruses and their respective hosts. However, we should stress that of the total number of dsDNA viruses studied (4,165), the majority of them (3,778) are phages, which comprises 91% of the total of this group. Therefore, this distribution is directly linked to the adaptation of phages to the GC content of the prokaryotic hosts (see **Figure 3B**).

The bimodal distributions of bases from +ssRNA-RT (**Figure 1G** and **Supplementary Figure 1D**) are intriguing. This was previously observed among members of the family Retroviridae by Berkhout et al. (2002), although with a reduced sample size. This pattern is not due to be single-stranded, since ssDNA, +ssRNA, and −ssRNA viruses (**Figure 1C,E,F**) display unimodal distributions. One possible explanation is that different +ssRNA-RT viruses are replicated by enzymes that introduce dissimilar mutational biases (Berkhout et al., 2002). To fully understand this point, it is necessary to analyze deeply these viruses and their respective life cycles and enzymes.

We expected that single-stranded (i.e., ssDNA and ssRNAs) viruses should display, on average, remarkably lower G and C frequencies in relation to double-stranded, since ss genomes are prone to mutations toward A and T/U (Lynch, 2007; Long et al., 2018). However, we did not see extreme differences among Baltimore classes, with the exception of dsDNA viruses, but we found that in ssRNA viruses (**Figure 1C,E–G** and **Supplementary Figure 1**), always A is the most frequent base followed by U. This is in agreement with a recent study considering a large number of ssRNA viruses (Kustin and Stern, 2021).

Regarding compositional correlations, the main conclusions that can be reached (**Figure 2**) are the following: (a) As has been known from a long time (for the first reports see: Muto and Osawa, 1987; D'Onofrio et al., 1991), strong correlations do hold in prokaryotes and eukaryotes between the GC content and the corresponding values of the three codon positions. To the best of our knowledge, this is the first time that a similar result is found for all viruses. This implies that despite (i) the different life cycles of each virus, including hosts, (ii) the different enzymes that duplicate each genome, and (iii) their different genetic material, the mutational bias operates in the same direction (toward GC or AT/U) in any given genome. In other words, whatever the cycle of the virus or the genetic material (**Table 3**), if the replication and/or repair systems are prone to enrich in either GC or AT/U, it does so in the whole genome, irrespective of the region (coding or non-coding).

(b) In spite of the previous point, as happens with prokaryotes and eukaryotes (Muto and Osawa, 1987; D'Onofrio et al., 1991), the strength of this mutational bias is strongly dependent on the codon position. Although the three codon positions increase (or decrease) with the corresponding non-coding sequences, each position changes with different strength: while GC1 shows a moderate increase (**Figure 2A**), GC3 shows the greatest variation (**Figure 2C**) while GC2 is the most constrained (**Figure 2B**). With no doubt, as it is well documented for prokaryotes, where most compositional studied have been done (Zhou et al., 2014), the different behaviors of the three codon positions reflects the structure of the genetic code. Indeed, while any variation in GC2 leads to an amino acid substitution, GC3 is rather free to change since, with the only exceptions of Trp and Met (which, at least in the universal genetic code, are encoded by only one codon each), most changes in GC3 are synonymous; from this point of view GC1 has an intermediate position.

In summary: (i) these correlations, that hold between non-coding and coding regions and their codon positions are indeed universal. (ii) They are independent of the genetic material: indeed, they can be seen not only in prokaryotes and eukaryotes (with dsDNA as genetic material) but in viruses, which as known, can be ss or dsRNA, ss or dsDNA, retrotranscribed or not. They are independent of the (iii) host and of (iv) the replication enzymes. (v) The structure of the genetic code is the main force that imposes limits to the "degree of freedom" of the correlations with the three codon positions. Hypothetically, a steeper slope between the non-coding sequences of viruses with GC2, similar in magnitude to the one found for GC3 (1.37), could cause that some amino acids would not be used (or used at extremely low frequencies) in viruses displaying extremely high (or low) GC content.

The study of GC content of viruses in relation to the GC content of their hosts (eukaryotes and prokaryotes) displays

**TABLE 4 |** Spearman's correlation coefficients (ρ) and adjusted $R^2$ ($adjR^2$) coefficients between codon frequencies of phages (first and second columns) or eukaryotic viruses (third and fourth columns), and the respective values or their hosts.

| | Phages | | Eukaryotic viruses | |
|---|---|---|---|---|
| **Codon** | **ρ** | **$adjR^2$** | **ρ** | **$adjR^2$** |
| UUU | 0.85 | 0.70 | 0.01 | 0.00 |
| UUC | 0.71 | 0.50 | −0.13 | 0.00 |
| UUA | 0.92 | 0.85 | 0.14 | 0.00 |
| UUG | 0.49 | 0.20 | 0.14 | 0.02 |
| CUU | 0.64 | 0.36 | 0.10 | 0.01 |
| CUC | 0.85 | 0.65 | 0.07 | 0.01 |
| CUA | 0.74 | 0.49 | −0.09 | 0.01 |
| CUG | 0.77 | 0.57 | 0.19 | 0.05 |
| AUU | 0.82 | 0.66 | 0.13 | 0.01 |
| AUC | 0.80 | 0.62 | 0.00 | 0.00 |
| AUA | 0.85 | 0.75 | 0.06 | 0.00 |
| AUG | 0.60 | 0.29 | 0.10 | 0.00 |
| GUU | 0.67 | 0.46 | 0.21 | 0.04 |
| GUC | 0.81 | 0.67 | 0.05 | 0.01 |
| GUA | 0.77 | 0.52 | −0.00 | 0.00 |
| GUG | 0.75 | 0.52 | 0.14 | 0.02 |
| UAU | 0.84 | 0.69 | 0.07 | 0.00 |
| UAC | 0.57 | 0.30 | 0.13 | 0.02 |
| UAA | 0.73 | 0.54 | 0.09 | 0.00 |
| UAG | 0.48 | 0.13 | 0.04 | 0.00 |
| CAU | 0.70 | 0.51 | 0.19 | 0.04 |
| CAC | 0.83 | 0.67 | −0.03 | 0.00 |
| CAA | 0.87 | 0.72 | 0.09 | 0.00 |
| CAG | 0.59 | 0.47 | 0.16 | 0.03 |
| AAU | 0.86 | 0.72 | 0.23 | 0.03 |
| AAC | 0.26 | 0.09 | 0.04 | 0.00 |
| AAA | 0.90 | 0.82 | 0.02 | 0.00 |
| AAG | 0.35 | 0.11 | 0.23 | 0.01 |
| GAU | 0.73 | 0.57 | 0.16 | 0.03 |
| GAC | 0.78 | 0.65 | 0.25 | 0.05 |
| GAA | 0.82 | 0.67 | 0.06 | 0.01 |
| GAG | 0.72 | 0.46 | −0.02 | 0.00 |
| UCU | 0.61 | 0.29 | 0.05 | 0.00 |
| UCC | 0.71 | 0.44 | −0.01 | 0.00 |
| UCA | 0.80 | 0.59 | 0.09 | 0.00 |
| UCG | 0.82 | 0.72 | 0.25 | 0.04 |
| CCU | 0.55 | 0.27 | −0.03 | 0.00 |
| CCC | 0.84 | 0.70 | 0.13 | 0.02 |
| CCA | 0.65 | 0.35 | 0.07 | 0.00 |
| CCG | 0.81 | 0.62 | 0.05 | 0.00 |
| ACU | 0.69 | 0.37 | −0.06 | 0.00 |
| ACC | 0.85 | 0.67 | 0.24 | 0.05 |
| ACA | 0.81 | 0.72 | −0.02 | 0.00 |
| ACG | 0.65 | 0.41 | −0.00 | 0.00 |
| GCU | 0.49 | 0.19 | 0.07 | 0.01 |
| GCC | 0.82 | 0.61 | 0.08 | 0.02 |
| GCA | 0.51 | 0.25 | −0.04 | 0.00 |
| GCG | 0.76 | 0.57 | 0.04 | 0.01 |
| UGU | 0.72 | 0.49 | 0.09 | 0.01 |
| UGC | 0.62 | 0.39 | 0.08 | 0.01 |
| UGA | 0.66 | 0.51 | −0.00 | 0.00 |
| UGG | 0.39 | 0.19 | 0.07 | 0.00 |
| CGU | 0.54 | 0.28 | 0.28 | 0.05 |
| CGC | 0.79 | 0.56 | 0.12 | 0.02 |
| CGA | 0.13 | 0.05 | −0.01 | 0.00 |
| CGG | 0.84 | 0.67 | −0.02 | 0.00 |
| AGU | 0.82 | 0.67 | 0.10 | 0.01 |
| AGC | 0.35 | 0.14 | 0.09 | 0.02 |
| AGA | 0.79 | 0.62 | 0.04 | 0.00 |
| AGG | 0.33 | 0.37 | 0.15 | 0.02 |
| GGU | 0.48 | 0.20 | 0.10 | 0.01 |
| GGC | 0.78 | 0.56 | 0.21 | 0.04 |
| GGA | 0.55 | 0.39 | −0.07 | 0.00 |
| GGG | 0.59 | 0.30 | 0.16 | 0.02 |
| Median | 0.73 | 0.52 | 0.08 | 0.01 |

*At the bottom of each column, the median value is presented.*

two completely different patterns. While in the majority of eukaryotes (animals and plants) there appears to be no relation (**Figure 3B,C**), in prokaryotes does exist a strong positive correlation: as the GC content of the host increases, there is an increment in the genomic GC of the respective phages, which was noted previously by Bahir et al. (2009) and Bohlin and Pettersson (2019), among others. Furthermore, as noted by Rocha and Danchin (2002), the GC content of the phages is, in general, lower than that of the respective hosts. However, it is interesting to note that fungi and their viruses do display a moderate positive correlation. Finally, among protists, we note that there is a negative and significant linear correlation between the two mentioned variables. This latter result needs more data to be more accurately portrayed.

Concerning codon usage, we found a similar pattern as in genomic compositional correlations (displayed in **Figure 3**). Indeed, for a long time, it has been known that in general there is a strong similarity in codon usage between prokaryotes and their phages (Sau et al., 2005; Esposito et al., 2006; Lucks et al., 2008), mainly with dsDNA phages in relation to ssDNA (Chithambaram et al., 2014). The very weak correlation observed for Arg CGA codon ($\rho = 0.13$) is interesting in light of the fact that this codon is involved in ribosome stalling when appear paired with CCG (i.e., CGA-CCG codon pair) and with another CGA (i.e., CGA-CGA) (Samatova et al., 2021).

However, from **Table 4** it is evident that codon usage in eukaryotic viruses is independent of the codon usage of their hosts (see, for instance: Cristina et al., 2015; Castells et al., 2017; Tian et al., 2018; Anwar et al., 2019), although some exceptions do this general rule exist, at least in some unicellular eukaryotes and giant viruses (Michely et al., 2013). This is important given that a codon usage pattern in viruses similar to their hosts could be advantageous for these obligate parasites, since this would

allow them to replicate faster and with a lesser extent of errors (Bahir et al., 2009).

This general lack of adaptation might be due to at least three non-mutually exclusive facts. First, most viruses that infect pluricellular species tend to infect specific tissues, where highly specific expressed genes display in turn different codon frequencies [for example, in the case of humans, see TissueCoCoPUTs database (Kames et al., 2020)]. Second, the concept "adaptation" might imply using the less frequent codons in the infected eukaryote, and thus reduce the competition with the more highly expressed host genes, avoiding placing greater stress on the host cell (Chen et al., 2020). Third, the most predominant force shaping codon usage in some eukaryotic viruses could be the mutational bias intrinsic to the enzymes that replicate their genomes. This would lead to very different GC contents and, consequently, different patterns of codon usage, which might, or might not, coincide with that of the host.

## CONCLUSION

In this study, we have analyzed several compositional properties of nearly 10,000 viral species: genomic base composition (globally and according to Baltimore classification), correlations among non-coding regions and the three codon positions, and the relationship of viral genomic base composition and codon usage with the same feature of their hosts. This allowed us to confirm, with a high number of viruses and hosts, that the genomic base composition and codon usage of phages strongly correlates with the respective values of their hosts. In contrast, as previously but not consensually reported, animal and plant viruses show no correlation between their GC content and that of their hosts. Finally, while all 64 codons show positive correlations between phages and hosts values, in contrast, for eukaryotes and their viruses, overall, the correlations are weak or do not exist.

## DATA AVAILABILITY STATEMENT

The source code and datasets presented here are available on GitHub at: https://github.com/lompa/virushostgc.

## AUTHOR CONTRIBUTIONS

DS and HM conceived and designed the work and drafted the manuscript. DS conducted all bioinformatics analyses and arranged figures and tables. DS, JC, and HM revised the manuscript, participated in the literature search and discussion, and read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

DS gratefully acknowledges ANII for the scholarship awarded (application POS_NAC_2016_1_130463) and PEDECIBA-Bioinformática for student financial aid received. JC and HM are members of PEDECIBA and SNI. The authors would like to thank PEDECIBA for partially covering the publication fees.

## ACKNOWLEDGMENTS

We thank the reviewers for useful suggestions that significantly improved the original manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.646300/full#supplementary-material

## REFERENCES

Agashe, D., and Shankar, N. (2014). The evolution of bacterial DNA base composition. *J. Exp. Zool. Part B Mol. Dev. Evol.* 322, 517–528. doi: 10.1002/jez.b.22565

Alexaki, A., Kames, J., Holcomb, D. D., Athey, J., Santana-Quintero, L. V., Lam, P. V. N., et al. (2019). Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J. Mol. Biol.* 431, 2434–2441. doi: 10.1016/j.jmb.2019.04.021

Anwar, A. M., Soudy, M., and Mohamed, R. (2019). vhcub: virus-host codon usage co-adaptation analysis. *F1000Res* 8:2137. doi: 10.12688/f1000research.21763.1

Auewarakul, P. (2005). Composition bias and genome polarity of RNA viruses. *Virus Res.* 109, 33–37. doi: 10.1016/j.virusres.2004.10.004

Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* 5:311. doi: 10.1038/msb.2009.71

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2017). GenBank. *Nucleic Acids Res.* 45, D37–D42.

Berkhout, B., and van Hemert, F. J. (1994). The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Res.* 22, 1705–1711. doi: 10.1093/nar/22.9.1705

Berkhout, B., Grigoriev, A., Bakker, M., and Lukashov, V. V. (2002). Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. *AIDS Res. Hum. Retroviruses* 18, 133–141. doi: 10.1089/08892220252779674

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., et al. (1985). The mosaic genome of warm-blooded vertebrates. *Science* 80, 953–958. doi: 10.1126/science.4001930

Bohlin, J., and Pettersson, J. H. O. (2019). Evolution of genomic base composition: from single cell microbes to multicellular animals. *Comput. Struct. Biotechnol. J.* 17, 362–370. doi: 10.1016/j.csbj.2019.03.001

Brister, J. R., Ako-adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. *Nucleic Acids Res.* 43, D571–D577.

Castells, M., Victoria, M., Colina, R., Musto, H., and Cristina, J. (2017). Genome-wide analysis of codon usage bias in Bovine Coronavirus. *Virol. J.* 14:115.

Chen, F., Peng, W., Deng, S., Zhang, H., Hou, Y., Zheng, H., et al. (2020). Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat. Ecol. Evol.* 4, 589–600. doi: 10.1038/s41559-020-1124-7

Chen, Y. (2013). A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *Biomed. Res. Int.* 2013:406342.

Chithambaram, S., Prabhakaran, R., and Xia, X. (2014). Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. *Mol. Biol. Evol.* 6, 1606–1617. doi: 10.1093/molbev/msu087

Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as winners in the game of life. *Annu. Rev. Virol.* 3, 197–214. doi: 10.1146/annurev-virology-100114-054952

Costantini, M., Alvarez-Valin, F., Costantini, S., Cammarano, R., and Bernardi, G. (2013). Compositional patterns in the genomes of unicellular eukaryotes. *BMC Genomics* 14:755. doi: 10.1186/1471-2164-14-755

Costantini, M., and Musto, H. (2017). The isochores as a fundamental level of genome structure and organization: a general overview. *J. Mol. Evol.* 84, 93–103. doi: 10.1007/s00239-017-9785-9

Costantini, M., Cammarano, R., and Bernardi, G. (2009). The evolution of isochore patterns in vertebrate genomes. *BMC Genomics* 10:146. doi: 10.1186/1471-2164-10-146

Cristina, J., Moreno, P., Moratorio, G., and Musto, H. (2015). Genome-wide analysis of codon usage bias in Ebolavirus. *Virus Res.* 196, 87–93. doi: 10.1016/j.virusres.2014.11.005

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., and Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 32, 504–510. doi: 10.1007/bf02102652

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276. doi: 10.1038/nrg2323

Durzyńska, J., and Goździcka-Józefiak, A. (2015). Viruses and cells intertwined since the dawn of evolution. *Virol. J.* 12, 169.

Esposito, L. A., Gupta, S., Streiter, F., Prasad, A., and Dennehy, J. J. (2006). Evolutionary interpretations of mycobacteriophage biodiversity and host-range through the analysis of codon usage bias. *Microb. Genom.* 2:10. doi: 10.1099/mgen.0.000079

Eyre-Walker, A., and Hurst, L. D. (2001). The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555. doi: 10.1038/35080577

Foerstner, K. U., von Mering, C., Hooper, S. D., and Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6, 1208–1213. doi: 10.1038/sj.embor.7400538

Gu, W., Zhou, T., Ma, J., Sun, X., and Lu, Z. (2004). Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. *Virus Res.* 101, 155–161. doi: 10.1016/j.virusres.2004.01.006

Holmes, E. C. (2011). What does virus evolution tell us about virus origins? *J. Virol.* 85, 5247–5251. doi: 10.1128/jvi.02203-10

Jenkins, G. M., and Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* 92, 1–7. doi: 10.1016/s0168-1702(02)00309-x

Kames, J., Alexaki, A., Holcomb, D. D., Santana-Quintero, L. V., Athey, J. C., Hamasaki-Katagiri, N., et al. (2020). TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. *J. Mol. Biol.* 432, 3369–3378. doi: 10.1016/j.jmb.2020.01.011

Khrustalev, V. V., and Barkovsky, E. V. (2011). "Protoisochores" in certain archaeal species are formed by replication-associated mutational pressure. *Biochimie* 93, 160–167. doi: 10.1016/j.biochi.2010.09.006

Khrustalev, V. V., Barkovsky, E. V., Khrustaleva, T. A., and Lelevich, S. G. (2014). Intragenic isochores (intrachores) in the platelet phosphofructokinase gene of Passeriform birds. *Gene* 546, 16–24. doi: 10.1016/j.gene.2014.05.045

Kindler, E., and Thiel, V. (2014). To sense or not to sense viral RNA—essentials of coronavirus innate immune evasion. *Curr. Opin. Microbiol.* 20, 69–75. doi: 10.1016/j.mib.2014.05.005

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct* 1:29.

Krupovic, M., Dolja, V. V., and Koonin, E. V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* 17, 449–458. doi: 10.1038/s41579-019-0205-6

Kustin, T., and Stern, A. (2021). Biased mutation and selection in RNA viruses. *Mol. Biol. Evol.* 38, 575–588. doi: 10.1093/molbev/msaa247

Lamolle, G., Protasio, A. V., Iriarte, A., Jara, E., Simón, D., and Musto, H. (2016). An isochore-like structure in the genome of the flatworm *Schistosoma mansoni*. *Genome Biol. Evol.* 8, 2312–2318. doi: 10.1093/gbe/evw170

Li, L., Zhou, J., Wu, Y., Yang, W., and Tian, D. (2015). GC-content of synonymous codons profoundly influences amino acid usage. *G3* 5, 2027–2036. doi: 10.1534/g3.115.019877

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., et al. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* 2, 237–240. doi: 10.1038/s41559-017-0425-y

Lucks, J. B., Nelson, D. R., Kudla, G. R., and Plotkin, J. B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol.* 4:e1000001. doi: 10.1371/journal.pcbi.1000001

Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates, Inc.

Mahmoudabadi, G., and Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. *Elife* 7:e31955.

Michely, S., Toulza, E., Subirana, L., John, U., Cognat, V., Maréchal-Drouard, L., et al. (2013). Evolution of codon usage in the smallest photosynthetic eukaryotes and their giant viruses. *Genome Biol. Evol.* 5, 848–859. doi: 10.1093/gbe/evt053

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016). Linking virus genomes with host taxonomy. *Viruses* 8:66. doi: 10.3390/v8030066

Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci.* 84, 166–169. doi: 10.1073/pnas.84.1.166

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes. R Package Version* 1.1-2.

Pracana, R., Hargreaves, A. D., Mulley, J. F., and Holland, P. W. H. (2020). Runaway GC evolution in gerbil genomes. *Mol. Biol. Evol.* 37, 2197–2210. doi: 10.1093/molbev/msaa072

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

R Core Team (2020). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reichenberger, E. R., Rosen, G., Hershberg, U., and Hershberg, R. (2015). Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol. Evol.* 7, 1380–1389. doi: 10.1093/gbe/evv063

Rocha, E. P. C., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18, 291–294. doi: 10.1016/s0168-9525(02)02690-2

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio.

Sabbia, V., Piovani, R., Naya, H., Rodríguez-Maseda, H., Romero, H., and Musto, H. (2007). Trends of Amino Acid usage in the proteins from the human genome. *J. Biomol. Struct. Dyn.* 25, 55–59. doi: 10.1080/07391102.2007.10507155

Samatova, E., Daberger, J., Liutkute, M., and Rodnina, M. V. (2021). Translational control by ribosome pausing in bacteria: how a non-uniform pace of translation affects protein production and folding. *Front. Microbiol.* 11:3428.

Sau, K., Gupta, S. K., Sau, S., and Ghosh, T. C. (2005). Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. *Virus Res.* 113, 123–131. doi: 10.1016/j.virusres.2005.05.001

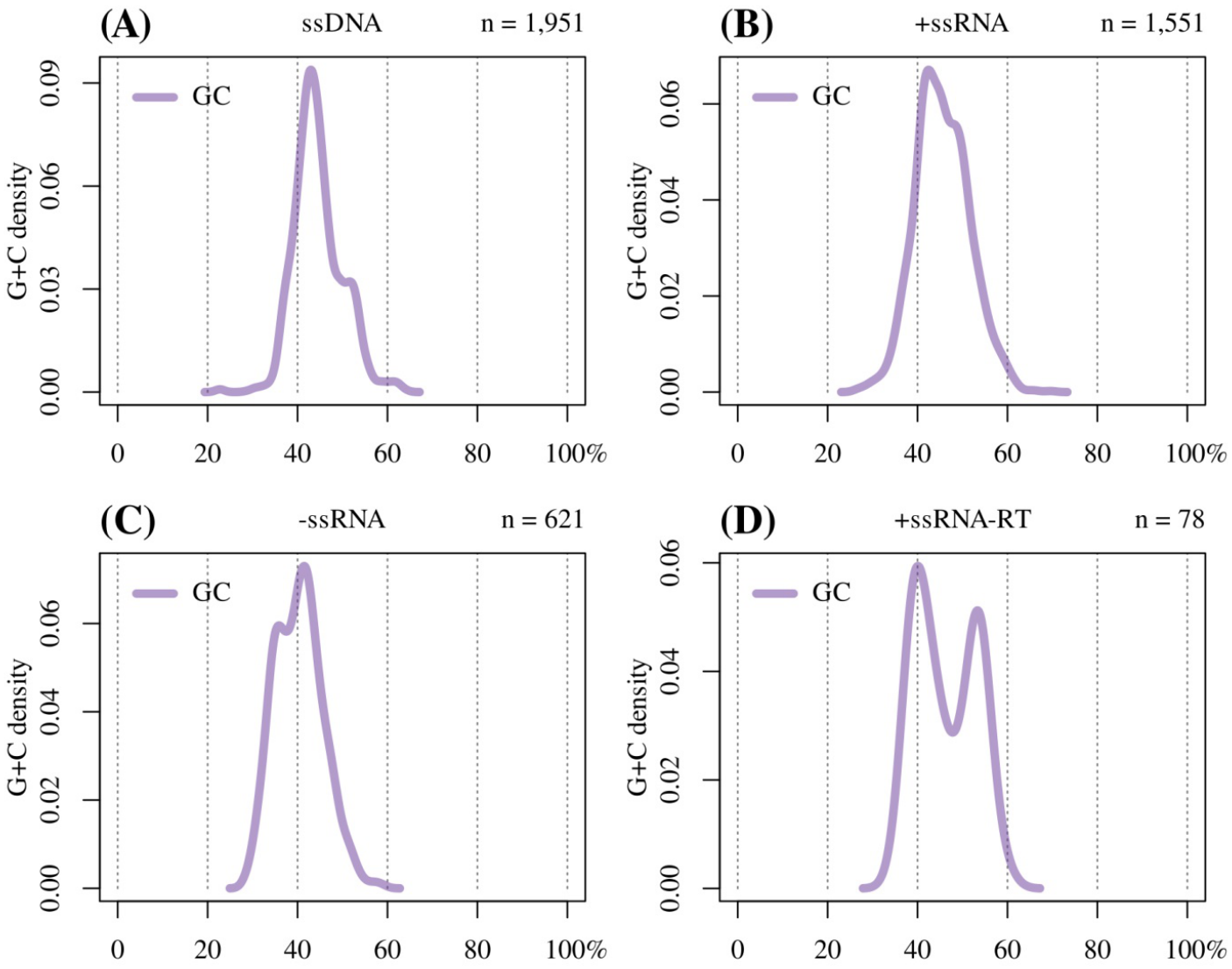
Simón, D., Fajardo, A., Sóñora, M., Delfraro, A., and Musto, H. (2017). Host influence in the genomic composition of flaviviruses: a multivariate approach. *Biochem. Biophys. Res. Commun.* 492, 572–578. doi: 10.1016/j.bbrc.2017.06.088

Tian, L., Shen, X., Murphy, R. W., and Shen, Y. (2018). The adaptation of codon usage of +ssRNA viruses to their hosts. *Infect. Genet. Evol.* 63, 175–179. doi: 10.1016/j.meegid.2018.05.034

van Hemert, F., van der Kuyl, A. C., and Berkhout, B. (2014). On the nucleotide composition and structure of retroviral RNA genomes. *Virus Res.* 193, 16–23. doi: 10.1016/j.virusres.2014.03.019

Zhou, H. Q., Ning, L. W., Zhang, H. X., and Guo, F. B. (2014). Analysis of the relationship between genomic GC content and patterns of base usage, codon usage and amino acid usage in prokaryotes: similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages. *PLoS One* 9:e107319. doi: 10.1371/journal.pone.0107319

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.




**SUPPLEMENTARY FIGURE 1**. GC-content of single-stranded (ss) viruses by Baltimore classification groups: **(A)** ssDNA, **(B)** positive ssRNA (+ssRNA), **(C)** negative ssRNA (-ssRNA), and **(D)** +ssRNA retro-transcribing (+ssRNA-RT).

Contribución

En este trabajo realizamos análisis composicionales incorporando:

**(i)** Frecuencias de bases de todos los virus disponibles.

**(ii)** Frecuencias de bases de los virus según la clasificación de Baltimore: dsDNA, ssDNA, dsRNA, ssRNA, +ssRNA, ssRNA, -ssRNA, +ssRNA-RT y dsDNA-RT.

**(iii)** Correlaciones composicionales entre el contenido de G+C no codificante vs. GC1, GC2 y GC3.

**(iv)** Para cada grupo, estudiamos la variación del contenido de G+C de los genomas virales en comparación con el de sus respectivos hospederos.

**(v)** Analizamos los patrones de uso de codones de los virus en relación con el de sus hospederos.

Nuestras principales conclusiones son que los diferentes virus (según la naturaleza y arquitectura de su material genético), muestran diferentes propiedades en su composición de bases. Además, existen fuertes correlaciones composicionales entre las regiones no codificantes y las tres posiciones de codones.

En cuanto a la relación entre los virus y sus respectivos hospederos, pudimos describir con un elevado número de virus (~10.000) y hospederos (~1.200), que la composición de bases y el uso de codones de los fagos se correlacionan fuertemente con los valores de los procariotas que infectan.

Por el contrario, como ya se había informado anteriormente pero no de forma consensuada, los virus de animales y de plantas no muestran correlación entre su contenido de G+C y el de sus hospederos.

Por último, el uso de codones de los fagos depende del uso de codones de los procariotas, mientras que el uso de codones de los virus animales y vegetales no parece estar adaptado al uso de codones de sus hospederos, con la excepción de hongos y protistas.

## Consideraciones finales

El principal balance ha sido el crecimiento personal y profesional, tomando contacto estrecho con múltiples herramientas, bases de datos y lenguajes de programación. Otro crecimiento a lo largo de esta tesis ha sido el aprendizaje que requiere redactar manuscritos científicos para ser evaluados en revistas arbitradas internacionales.

Además, se ha continuado una colaboración fructífera entre los laboratorios de Héctor Musto y de Juan Cristina que ya existía, principalmente en análisis de codones sinónimos, pero que se ha desarrollado aún más, incorporando análisis a nivel de dinucleótidos y de aminoácidos (Simón et al., 2017; Simón et al., 2018).

En un próximo paso, por una línea de investigación que se abrió luego de mi trabajo con flavivirus (Simón et al., 2017), intentaré responder algunas preguntas evolutivas en arbovirus, en colaboración con el Laboratorio de Virología Molecular de la Facultad de Ciencias de Udelar. Dicho trabajo se enmarcará en proyectos en evaluación y en mi Doctorado en Ciencias Biológicas de Pedeciba.

Los análisis composicionales pueden ser interesantes para complementar cualquier abordaje en genómica evolutiva, tanto en virus como en cualquier organismo celular. No quise tampoco perder la perspectiva histórica de la relevancia que los estudios composicionales han tenido en la historia de la biología molecular.

Actualmente, gozan de buena salud, fundamentalmente por su aplicación en el diseño de vacunas atenuadas (Gonçalves-Carneiro y Bieniasz, 2021; Pereira-Gómez et al., 2021). Pueden además ser útiles en otras situaciones más generales; por ejemplo, quienes hagan expresión heteróloga de proteínas deben tener en cuenta los sesgos de la proteína a expresar y los sesgos del sistema de expresión.

Queda de manifiesto el impacto de la composición nucleotídica del hospedero en el genoma viral de los virus que lo infectan. Esto había sido descrito para los flavivirus (Simón et al., 2017), pero recientemente se intentó abarcar toda la diversidad viral (Simón et al., 2021). Además, los datos generados en esta tesis permitirán comparar virus de interés para cualquier investigador con toda la diversidad viral conocida, o parte de ella (e.g., grupos, familias, géneros, hospederos).

Esta tesis de Maestría en Bioinformática del Programa de Desarrollo de las Ciencias Básicas (PEDECIBA) ha sido conceptualizada para realizar análisis composicionales en genomas virales de manera semiautomatizada. Sin embargo, los scripts aquí presentados son totalmente utilizables en formatos compatibles de cualquier otro origen.

A su vez, los scripts permitirán hacer los mismos análisis para nuevas secuencias. En muchos casos, debido a que la mayor fuente actual de nuevas especies virales es la metagenómica, abordajes de este tipo pueden ayudar a inferir información biológica de virus de los que solamente conocemos su genoma.

## Referencias bibliográficas


Academic Tree (2005) **Renato Dulbecco** [Internet], Chemistry Tree, The Academic Family Tree. Disponible desde: academictree.org/chemistry/peopleinfo.php?pid=1949 [Acceso nov. 2021].

Altmann, R. (1889) Ueber Nucleinsäuren. **Archiv für Anatomie und Physiologie in Leipzig**, 5-6, pp. 524–36.

Argos, P. et al. (1984) Similarity in gene organization and homology between proteins of animal picomaviruses and a plant comovirus suggest common ancestry of these virus families. **Nucleic Acids Research**, 12(18), pp. 7251–67. doi: 10.1093/nar/12.18.7251.

Avery, O. T., Macleod, C. M. y McCarty, M. (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. **The Journal of Experimental Medicine**, 79(2), pp. 137–58. doi: 10.1084/jem.79.2.137.

Baltimore, D. (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. **Nature**, 226(5252), pp. 1209–11. doi: 10.1111/j.2164-0947.1971.tb02600.x.

Baltimore, D. (1971) Expression of animal virus genomes. **Bacteriological reviews**, 35(3), pp. 235–41. doi: 10.1128/br.35.3.235-241.1971.

Baltimore, D. (1971) Viral genetic systems. **Transactions of the New York Academy of Sciences**, 33(3), pp. 327–32. doi: 10.1111/j.2164-0947.1971.tb02600.x.

Beijerinck, M. W. (1898) Ueber ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabakblätter. **Verhandelingen der Koninklijke Akademie vall Wetensellappen te Amsterdam**, 6(5), pp. 3–24. url: dwc.knaw.nl/DL/publications/PU00011860.pdf

Belozersky, A. N. y Spirin, A. S. (1958) A correlation between the compositions of deoxyribonucleic and ribonucleic acids. **Nature**, 182(4628), pp. 111–2. doi: 10.1038/182111a0.

Betz, F. (2011) **Managing Science**. Cham, Springer Nature Switzerland AG. doi: 10.1007/978-1-4419-7488-4

Brocchieri, L. (2014) The GC Content of Bacterial Genomes. **Journal of Phylogenetics & Evolutionary Biology**, 02(01), p. e108. doi: 10.4172/2329-9002.1000e108.

Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage. **Genetics**, 129(3), pp. 897–907. doi: 10.1093/genetics/129.3.897.

Burns, C. C. et al. (2006) Modulation of Poliovirus Replicative Fitness in HeLa Cells by Deoptimization of Synonymous Codon Usage in the Capsid Region. **Journal of Virology**, 80(7), pp. 3259–72. doi: 10.1128/JVI.80.7.3259-3272.2006.

Case, B. C. y Hingorani, M. M. (2017) Polymerase. En: **Reference Module in Life Sciences**. Elsevier. doi: 10.1016/B978-0-12-809633-8.06928-4.

Chan, S. y Conova, S. (2011) **Francis Crick (1916-2004)** [Internet], Cold Spring Harbor, NY, DNA Learning Center, Cold Spring Harbor Laboratory. Disponible desde: dnaftb.org/19/bio-2.html [Acceso nov. 2021].

Chan, S. y Conova, S. (2011) **Friedrich Miescher (1844-1895)** [Internet], Cold Spring Harbor, NY, DNA Learning Center, Cold Spring Harbor Laboratory. Disponible desde: dnaftb.org/15/bio.html [Acceso nov. 2021].

Chan, S. y Conova, S. (2011) **Phoebus Levene (1869-1940)** [Internet], Cold Spring Harbor, NY, DNA Learning Center, Cold Spring Harbor Laboratory. Disponible desde: dnaftb.org/15/bio-2.html [Acceso nov. 2021].

Charif, D. y Lobry, J. R. (2007) SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. En: Bastolla U., Porto, M., Roman, H.E., Vendruscolo, M. (eds.) **Structural Approaches to Sequence Evolution**, pp. 207–32. Cham, Springer Nature Switzerland AG. doi: 10.1007/978-3-540-35306-5_10.

Chargaff, E. et al. (1949) The composition of the desoxypentose nucleic acids of thymus and spleen. **The Journal of Biological Chemistry**, 177(1), pp. 405–16. url: ncbi.nlm.nih.gov/pubmed/18107444.

Chargaff, E., Zamenhof, S. y Green, C. (1950) Composition of human desoxypentose nucleic acid. **Nature**, 165(4202), pp. 756–7. doi: 10.1038/165756b0.

Cock, P. J. A. et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, 25(11), pp. 1422–3. doi: 10.1093/bioinformatics/btp163.

Cohen, S. S. (2004) Erwin Chargaff. **Proceedings of the American Philosophical Society**, 148(2), pp. 221–8. url: ncbi.nlm.nih.gov/pubmed/15338562.

Comfort, N. y Goldstein, W. (1995) **Coming of Phage**. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory. url: virology.ws/phage_brochure_lowres.pdf

Crick, F. H. (1962) **On the Genetic Code** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1962/crick/lecture [Acceso nov. 2021].

Crick, F. H. et al. (1961) General nature of the genetic code for proteins. **Nature**, 192, pp. 1227–32. doi: 10.1038/1921227a0.

d'Herelle, F. (1917) Sur un microbe invisible antagoniste des bacilles dysentérique. **Comptes Rendus de l'Académie des Sciences de Paris**, 165, pp. 373–5.

Dahm, R. (2005) Friedrich Miescher and the discovery of DNA. **Developmental Biology**, 278(2), pp. 274–88. doi: 10.1016/j.ydbio.2004.11.028.

Dahm, R. (2010) From discovering to understanding. Friedrich Miescher's attempts to uncover the function of DNA. **EMBO reports**, 11(3), pp. 153–60. doi: 10.1038/embor.2010.14.

Dutilh, B. E. et al. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. **Nature Communications**, 5, p. 4498. doi: 10.1038/ncomms5498.

Ellis, E. L. y Delbrück, M. (1939) The growth of bacteriophage. **The Journal of General Physiology**, 22(3), pp. 365–84. doi: 10.1085/jgp.22.3.365.

Elzanowski, A. y Ostell, J. (2019) **The Genetic Codes** [Internet], Bethesda, MD, National Center for Biotechnology Information. Disponible desde: ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi [Acceso nov. 2021].

Fajardo, Á. (2016) **Epidemiología molecular de flavivirus emergentes en Uruguay y las Américas: virus dengue y zika** [En línea]. Tesis de doctorado, Universidad de la República, Uruguay. Facultad de Ciencias. doi: 20.500.12008/17196.

Fiers, W. et al. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. **Nature**, 260(5551), pp. 500–7. doi: 10.1038/260500a0.

Franklin, R. E. y Gosling, R. G. (1953) Molecular configuration in sodium thymonucleate. **Nature**, 171(4356), pp. 740–1. doi: 10.1038/171740a0.

Gerstein, M. (1998) **What is Bioinformatics?** [Internet], New Haven, CT, Yale University. url: bioinfo.mbb.yale.edu/what-is-it.html

Golfheman (2010) Operating system placement [Internet], San Petersburgo, FL, Wikimedia Commons. Disponible desde: commons.wikimedia.org/wiki/File:Operating_system_placement.svg [Acceso nov. 2021].

Gonçalves-Carneiro, D. y Bieniasz, P. D. (2021) Mechanisms of Attenuation by Genetic Recoding of Viruses. **mBio**, 12(1). doi: 10.1128/mBio.02238-20.

Gouy, M. y Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. **Nucleic Acids Research**, 10(22), pp. 7055–74. doi: 10.1093/nar/10.22.7055.

Grantham, R. et al. (1980) Codon catalog usage and the genome hypothesis. **Nucleic Acids Research**, 8(1), p. 197. doi: 10.1093/nar/8.1.197-c.

Greenstein, J. P. (1943) Friedrich Miescher, 1844-1895. **The Scientific Monthly**, 57(6), pp. 523–32. url: jstor.org/stable/18231.

Griffith, F. (1928) The Significance of Pneumococcal Types. **The Journal of Hygiene**, 27(2), pp. 113–59. doi: 10.1017/s0022172400031879.

Grunberg-Manago, M. y Ochoa, S. (1955) Enzymic synthesis and breakdown of polynucleotides; Polynucleotide phosphorylase. **Journal of the American Chemical Society**, 77(11), pp. 3165–3166. doi: 10.1021/ja01616a093.

Hershey, A. D. y Chase, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. **The Journal of General Physiology**, 36(1), pp. 39–56. doi: 10.1085/jgp.36.1.39.

Hesper, B. y Hogeweg, P. (1970) Bioinformatica: een werkconcept. **Kameleon**, 1(6), pp. 28–9. Leiden, Leidse Biologen Club.

Holley, R. W. et al. (1965) Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid. **The Journal of Biological Chemistry,** 240, pp. 2122–8. url: ncbi.nlm.nih.gov/pubmed/14299636.

Holley, R. W. et al. (1965) Structure of a Ribonucleic Acid. **Science**, 147(3664), pp. 1462–5. doi: 10.1126/science.147.3664.1462.

Holmes, F. O. (1948) The Filterable Viruses. En: Breed, R. S., ed., Murray, E. G. D., ed., Hitchens, A. P., ed. **Bergey's Manual of Determinative Bacteriology**, 6(2), pp. 1127–86. Baltimore, MD, The Williams & Wilkins Company.

Hurwitz, J., Bresler, A. y Diringer, R. (1960) The enzymic incorporation of ribonucleotides into polyribonucleotides and the effect of DNA. **Biochemical and Biophysical Research Communications**, 3(1), pp. 15–9. doi: 10.1016/0006-291X(60)90094-2.

Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. **Journal of Molecular Biology**, 151(3), pp. 389–409. doi: 10.1016/0022-2836(81)90003-6.

International Committee on Taxonomy of Viruses Executive Committee (2020) The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. **Nature Microbiology**, 5(5), pp. 668–674. doi: 10.1038/s41564-020-0709-x.

Ivanovski, D. (1892) Über die Mosaikkrankheit der Tabakspflanze. **Bulletin Scientifique Publié Par l'Académie Impériale des Sciences de Saint-Pétersbourg**, 35, pp. 67–70.

Ivanovski, D. y Polovtsev, V. V. (1890) Die Pockenkrankheit der Tabakspflanze. **Mémoires de l'Académie Impériale des Sciences de Saint-Pétersbourg**, 7(37), pp. 1–24.

Jeffery, W. R., Adams, D. S. y Noonan, D. (1981) Cytoplasmic processing events in the polyadenylate region of Physarum messenger RNA. **Molecular Biology Reports**, 7(1-3), pp. 63–70. doi: 10.1007/BF00778735.

Jorge, D. M. de M., Mills, R. E. y Lauring, A. S. (2015) CodonShuffle: a tool for generating and analyzing synonymously mutated sequences. **Virus Evolution**, 1(1), p. vev012. doi: 10.1093/ve/vev012.

Kano-Sueoka, T. y Sueoka, N. (1969) Leucine tRNA and cessation of Escherichia coli protein synthesis upon phage T2 infection. **Proceedings of the National Academy of Sciences**, 62(4), pp. 1229–36. doi: 10.1073/pnas.62.4.1229.

Khorana, H. G. et al. (1965) Studies on Polynucleotides. XLII. The Synthesis of Deoxyribopolynucleotides Containing Repeating Nucleotide Sequences. Introduction and General Considerations. **Journal of the American Chemical Society**, 87, pp. 2954–6. doi: 10.1021/ja01091a027.

Komsta, L. y Novomestky, F. (2015) **moments: Moments, cumulants, skewness, kurtosis and related tests** [Internet]. R package version 0.14. url: CRAN.R-project.org/package=moment [Acceso nov. 2021].

Koonin, E. V et al. (2020) Global Organization and Proposed Megataxonomy of the Virus World. **Microbiology and Molecular Biology Reviews**, 84(2). doi: 10.1128/MMBR.00061-19.

Kornberg, A. (1957) Enzymatic synthesis of deoxyribonucleic acid. **The Harvey Society Lectures**, 53, pp. 83–112. url: ncbi.nlm.nih.gov/pubmed/13640421.

Kuhn, J. H. (2021) Virus Taxonomy. En: Bamford, D. H. y Zuckerman, M. **Encyclopedia of Virology**, 4, pp. 28–M37. doi: 10.1016/B978-0-12-809633-8.21231-4.

Le Mercier, P. (2021) **Arenaviridae** [Internet], Lausana, Swiss Institute of Bioinformatics. Disponible desde: viralzone.expasy.org/501 [Acceso nov. 2021].

Levene, P. A. y Jacobs, W. A. (1908) On glycothionic acid. **The Journal of Experimental Medicine**, 10(4), pp. 557–8. doi: 10.1084/jem.10.4.557.

Levene, P. A. y London, E. J. (1928) On The Structure of Thymonucleic Acid. **Science**, 68(1771), pp. 572–3. doi: 10.1126/science.68.1771.572-a.

Lynn, D. J., Singer, G. A. C. y Hickey, D. A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. **Nucleic Acids Research**, 30(19), pp. 4272–7. doi: 10.1093/nar/gkf546.

Lwoff, A., Horne, R. and Tournier, P. (1972) A System of Viruses. **Spring Harbor Symposia on Quantitative Biology**, 27, pp. 51–5.

Maechler, M. (2021) **diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected** [Internet], R package version 0.76-0. url: CRAN.R-project.org/package=diptest [Acceso nov. 2021].

Markham, R. y Smith, J. D. (1949) Chromatographic studies of nucleic acids 1. A technique for the identification and estimation of purine and pyrimidine bases, nucleosides and related substances. **The Biochemical Journal**, 45(3), pp. 294–8. doi: 10.1042/bj0450294.

Markham, R. y Smith, J. D. (1949) The quantitative analysis of ribonucleic acids. **The Biochemical Journal**, 45(5), pp. xxxii–iii [Suplemento]. url: ncbi.nlm.nih.gov/pubmed/15407852.

Markham, R. y Smith, J. D. (1950) Chromatographic studies on nucleic acids. 3. The nucleic acids of five strains of tobacco mosaic virus. **The Biochemical Journal**, 46(5), pp. 513–7. doi: 10.1042/bj0460513.

Mayer, A. (1886) Über die Mosaikkrankheit des Tabaks. **Die Landwirtschaftliche Versuchs-stationen**, 32, pp. 451–67.

McGeoch, D. J. y Davison, A. J., (1986) DNA sequence of the herpes simplex virus type 1 gene encoding glycoprotein gH, and identification of homologues in the genomes of varicella-zoster virus and Epstein-Barr virus. **Nucleic Acids Research**, 14(10), pp. 4281–92. doi: 10.1093/nar/14.10.4281.

Meselson, M. y Stahl, F. W. (1958) The replication of DNA in Escherichia coli. **Proceedings of the National Academy of Sciences**, 44(7), pp. 671–82. doi: 10.1073/pnas.44.7.671.

Mueller, S. et al. (2006) Reduction of the Rate of Poliovirus Protein Synthesis through Large-Scale Codon Deoptimization Causes Attenuation of Viral Virulence by Lowering Specific Infectivity. **Journal of Virology**, 80(19), pp. 9687–96. doi: 10.1128/JVI.00738-06.

Nakamura, M. y Sugiura, M. (2011) Translation efficiencies of synonymous codons for arginine differ dramatically and are not correlated with codon usage in chloroplasts. **Gene**, 472(1–2), pp. 50–4. doi: 10.1016/j.gene.2010.09.008.

Nirenberg, M. W. y Matthaei, J. H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. **Proceedings of the National Academy of Sciences**, 47, pp. 1588–602. doi: 10.1073/pnas.47.10.1588.

Nirenberg, M. W. y Leder, P. (1964) RNA Codewords and Protein Synthesis. The Effect of Trinucleotides upon the Binding of sRNA to Ribosomes. **Science**, 145(3639), pp. 1399–407. doi: 10.1126/science.145.3639.1399.

Nobel Prize (1910) **The Nobel Prize in Physiology or Medicine 1910** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1910 [Acceso nov. 2021].

Nobel Prize (1957) **The Nobel Prize in Chemistry 1957** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/chemistry/1957 [Acceso nov. 2021].

Nobel Prize (1959) **The Nobel Prize in Physiology or Medicine 1959** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1959 [Acceso nov. 2021].

Nobel Prize (1962) **The Nobel Prize in Physiology or Medicine 1962** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1962 [Acceso nov. 2021].

Nobel Prize (1968) **The Nobel Prize in Physiology or Medicine 1968** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1968 [Acceso nov. 2021].

Nobel Prize (1969) **The Nobel Prize in Physiology or Medicine 1969** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1969 [Acceso nov. 2021].

Nobel Prize (1975) **The Nobel Prize in Physiology or Medicine 1975** [Internet], Estocolmo, The Nobel Foundation. Disponible desde: nobelprize.org/prizes/medicine/1975 [Acceso nov. 2021].

Novoa, E. M. et al. (2019) Elucidation of Codon Usage Signatures across the Domains of Life. **Molecular Biology and Evolution**, 36(10), pp. 2328–2339. doi: 10.1093/molbev/msz124.

Pereira-Gómez, M. et al. (2021) Altering Compositional Properties of Viral Genomes to Design Live-Attenuated Vaccines. **Frontiers in Microbiology**, 12, p. 676582. doi: 10.3389/fmicb.2021.676582.

Philippe, N. et al. (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. **Science**, 341(6143), pp. 281–6. doi: 10.1126/science.1239181.

Plotkin, J. B. y Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. **Nature Reviews Genetics**, 12(1), pp. 32–42. doi: 10.1038/nrg2899.

Pruitt, K. et al. (2020) **RefSeq Frequently Asked Questions (FAQ)** [Internet], Bethesda, MD, National Center for Biotechnology Information. Disponible desde: ncbi.nlm.nih.gov/books/NBK50679/ [Acceso nov. 2021].

Rocha, E. P. C. y Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. **Trends in Genetics**, 18(6), pp. 291–4. doi: 10.1016/S0168-9525(02)02690-2.

Romero, H. (2010) El Club de la Corbata ARN. En: Tassino, B. y Silva, A., (eds.) **Biología: Unidad en la diversidad**. Montevideo, DIRAC.

Rudner, R., Karkas, J. D. y Chargaff, E. (1968) Separation of B. subtilis DNA into complementary strands, I. Biological properties. **Proceedings of the National Academy of Sciences**, 60(2), pp. 630–5. doi: 10.1073/pnas.60.2.630.

Sagan, L. (1967) On the origin of mitosing cells. **Journal of Theoretical Biology**, 14(3), pp. 255–74. doi: 10.1016/0022-5193(67)90079-3.

Sanger, F. et al. (1977) Nucleotide sequence of bacteriophage φX174 DNA. **Nature**, 265(5596), pp. 687–75. doi: 10.1038/265687a0.

Schoch, C. L. et al. (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database**, baaa062. doi: 10.1093/database/baaa062.

Simón, D. (2015) **Generación de una base de datos completa y no redundante de virus y su análisis composicional** [En línea]. Tesis de grado. Universidad de la República, Uruguay. Facultad de Ciencias. doi: 20.500.12008/8339.

Simón, D., Cristina, J. y Musto, H. (2021) Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. **Frontiers in Microbiology**, 12, p. 646300. doi: 10.3389/fmicb.2021.646300.

Simón, D. et al. (2018) An Evolutionary Insight into Zika Virus Strains Isolated in the Latin American Region. **Viruses**, 10(12). doi: 10.3390/v10120698.

Simón, D. et al. (2017) Host influence in the genomic composition of flaviviruses: A multivariate approach. **Biochemical and Biophysical Research Communications**, 492(4), pp. 572–8. doi: 10.1016/j.bbrc.2017.06.088.

Smith, J. D. y Wyatt, G. R. (1951) The composition of some microbial deoxypentose nucleic acids. **The Biochemical Journal**, 49(2), pp. 144–8. doi: 10.1042/bj0490144.

Splettstoesser, T. (2012). **Virus Baltimore Classification** [Internet], San Petersburgo, FL, Wikimedia Commons. Disponible desde: commons.wikimedia.org/wiki/File:VirusBaltimoreClassification.svg [Acceso nov. 2021].

Stevens, A. (1960) Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli B. **Biochemical and Biophysical Research Communications**, 3(1), pp. 92–6. doi: 10.1016/0006-291X(60)90110-8.

Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. **Proceedings of the National Academy of Sciences**, 48, pp. 582–92. doi: 10.1073/pnas.48.4.582.

Temin, H. M. y Mizutani, S. (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. **Nature**, 226(5252), pp. 1211–3. doi: 10.1038/2261211a0.

Thrash, J. C. et al. (2011) Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. **Scientific Reports**, 1, p. 13. doi: 10.1038/srep00013.

Twort, F. W. (1915) An investigation on the nature of ultra-microscopic viruses. **The Lancet**, 186(4814), pp. 1241–3. doi: 10.1016/S0140-6736(01)20383-3.

Watson, J. D. (1968) **The Double Helix**. Nueva York, NY, Atheneum.

Watson, J. D. y Crick, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. **Nature**, 171(4356), pp. 737–8. doi: 10.1038/171737a0.

Watson, J. D. y Crick, F. H. (1953) Genetical implications of the structure of deoxyribonucleic acid. **Nature**, 171(4361), pp. 964–7. doi: 10.1038/171964b0.

Weaver, W. (1970) Molecular biology: origin of the term. **Science**, 170(3958), pp. 581–2. doi: 10.1126/science.170.3958.581-a.

Weiner, J. (1999) **Time, Love, Memory**. Nueva York, NY, Alfred A. Knopf.

Winkler, H. (1920) **Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche**. Jena, Verlag von Gustav Fischer. doi: 10.5962/bhl.title.1460.

Wyatt, G. R. (1950) Occurrence of 5-methylcytosine in nucleic acids. **Nature**, 166(4214), pp. 237–8. doi: 10.1038/166237b0.

Wyatt, G. R. y Cohen, S. S. (1953) The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine. **The Biochemical Journal**, 55(5), pp. 774–82. doi: 10.1042/bj0550774.

Yutin, N. et al. (2021) Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. **Nature Communications**, 12(1), p. 1044. doi: 10.1038/s41467-021-21350-w.

Yutin, N. et al. (2018) Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. **Nature Microbiology**, 3(1), pp. 38–46. doi: 10.1038/s41564-017-0053-y.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. **Nucleic Acids Research**, 31(13), pp. 3406–15. doi: 10.1093/nar/gkg595

# Anexos

## (A) Archivos GenBank

Porción de los atributos (FEATURES) del archivo gbk correspondiente a la secuencia de referencia del SARS-Cov-2 (código de acceso número **NC_045512**; disponible desde ncbi.nlm.nih.gov/nuccore/1798174254).

```
(...)
FEATURES             Location/Qualifiers
     source          1..29903
                     /organism="Severe acute respiratory syndrome coronavirus
                     2"
                     /mol_type="genomic RNA"
                     /isolate="Wuhan-Hu-1"
                     /host="Homo sapiens"
                     /db_xref="taxon:2697049"
                     /country="China"
                     /collection_date="Dec-2019"
     5'UTR           1..265
     gene            266..21555
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /db_xref="GeneID:43740578"
     CDS             join(266..13468,13468..21555)
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /ribosomal_slippage
                     /note="pp1ab; translated by -1 ribosomal frameshift"
                     /codon_start=1
                     /product="ORF1ab polyprotein"
                     /protein_id="YP_009724389.1"
                     /db_xref="GeneID:43740578"
(...)
```

## (B) Bases de datos consultadas

| Base de datos | Sitios web y/o FTP |
|---|---|
| CoCoPUTs | https://hive.biochemistry.gwu.edu/review/codon |
| ICTV | https://talk.ictvonline.org/ |
| NCBI Genome | https://www.ncbi.nlm.nih.gov/genome/<br>https://ftp.ncbi.nlm.nih.gov/genomes |
| NCBI PubMed | https://pubmed.ncbi.nlm.nih.gov/<br>https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/ |
| NCBI RefSeq | https://www.ncbi.nlm.nih.gov/refseq/<br>https://ftp.ncbi.nlm.nih.gov/genomes/refseq/ |
| NCBI Taxonomy | https://www.ncbi.nlm.nih.gov/taxonomy<br>https://ftp.ncbi.nih.gov/pub/taxonomy/ |
| NCBI Virus | https://www.ncbi.nlm.nih.gov/labs/virus/ |
| ViralZone | https://viralzone.expasy.org/ |
| Virus-Host | https://www.genome.jp/virushostdb/<br>https://www.genome.jp/ftp/db/virushostdb/ |

## (C) Códigos genéticos presentes en los genomas analizados

Se toman como referencia las tablas (i.e., `transl_table`) de los códigos genéticos compilados en el NCBI (Elzanowski y Ostell, 2019).

**1.** Código estándar

```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
iniciadores
---M------**--*----M---------------M----------------------------
```

**4.** Micoplasmas[9] y mitocondrias de celentéreos, mohos y protozoarios

```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
iniciadores
--MM------**-------M------------MMMM---------------M------------
```

[9] Incluídos los espiroplasmas.

**5.** Mitocondrias de invertebrados

```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGG
iniciadores
---M------**--------------------MMMM--------------M------------
```

**6.** Ciliados, dasicladales y los diplomonádidos del género *Hexamita*

```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYYQQCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
iniciadores
--------------*--------------------M----------------------------
```

**11.** Árqueas, bacterias y plastos

```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
iniciadores
---M------**--*----M------------MMMM--------------M------------
```

**16.** Mitocondrias de clorofíceas

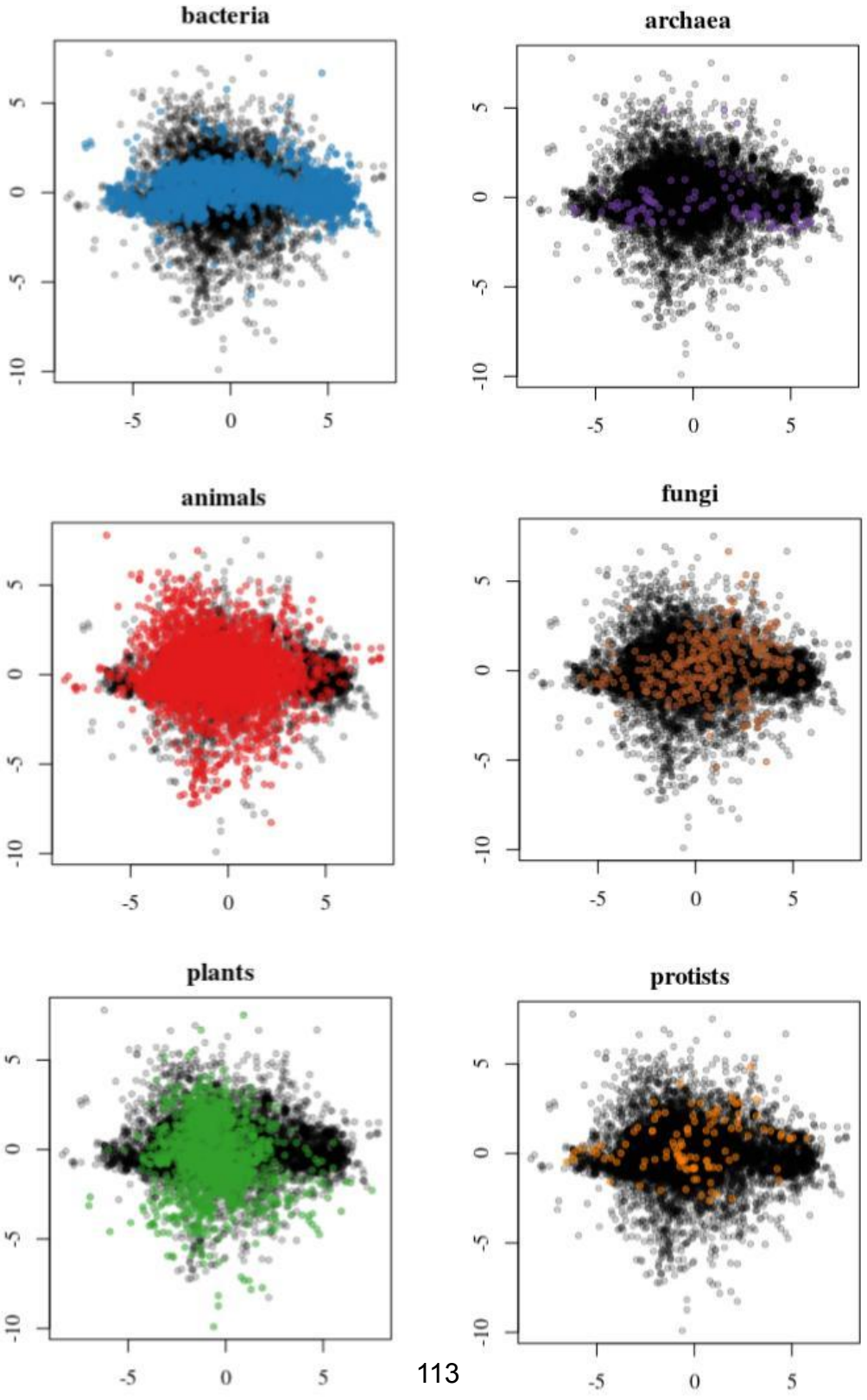
```
n-- (base1)
ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggg
-n- (base2)
ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggg
--n (base3)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
aminoácidos
FFLLSSSSYY*LCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
iniciadores
----------*---*--------------------M---------------------------
```

## Apéndices

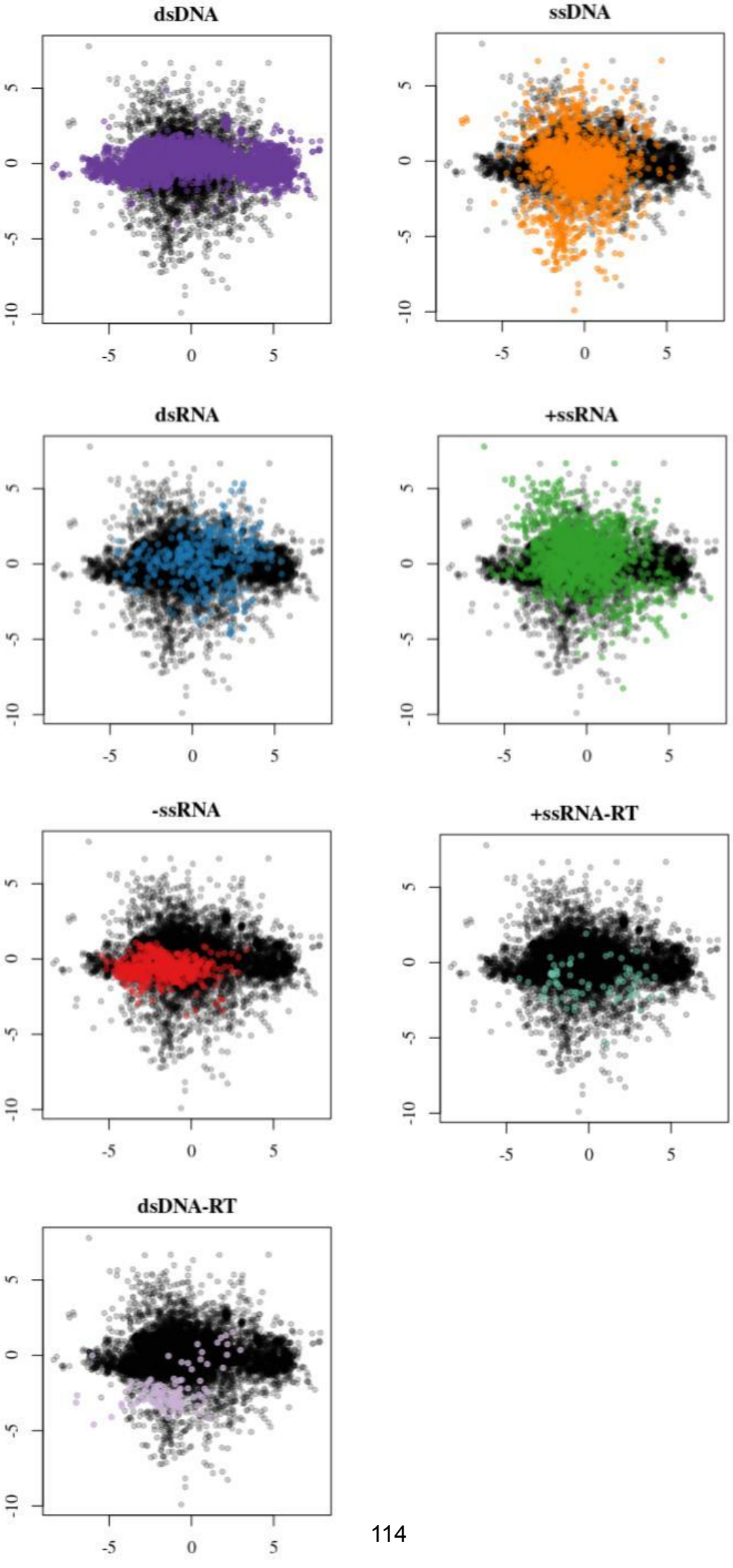
### Análisis multivariados para dinucleótidos, codones y aminoácidos

**A.** Frecuencia de dinucleótidos por hospedero **(1)** y por Baltimore **(2)**

**(A1)**

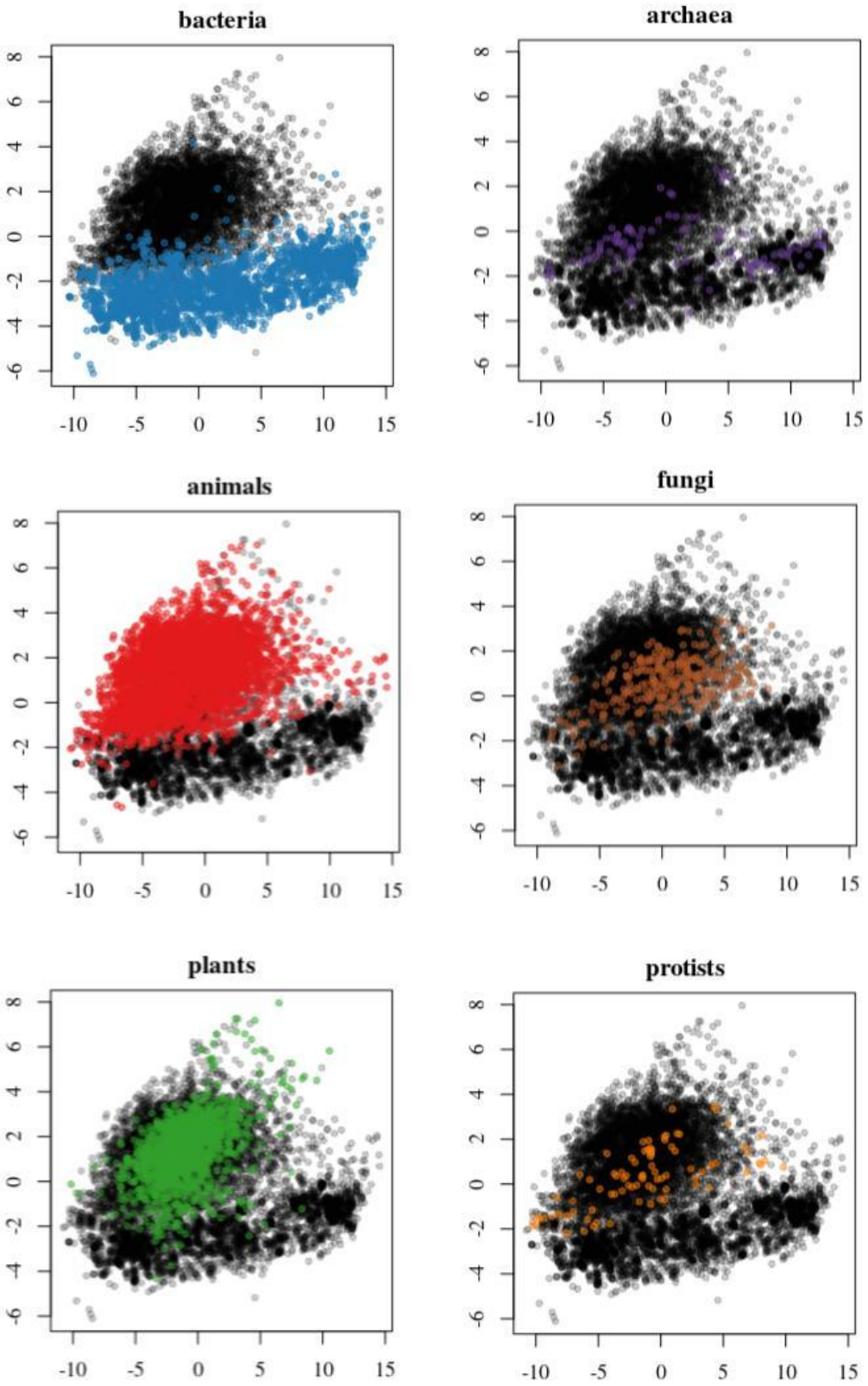


(A2)

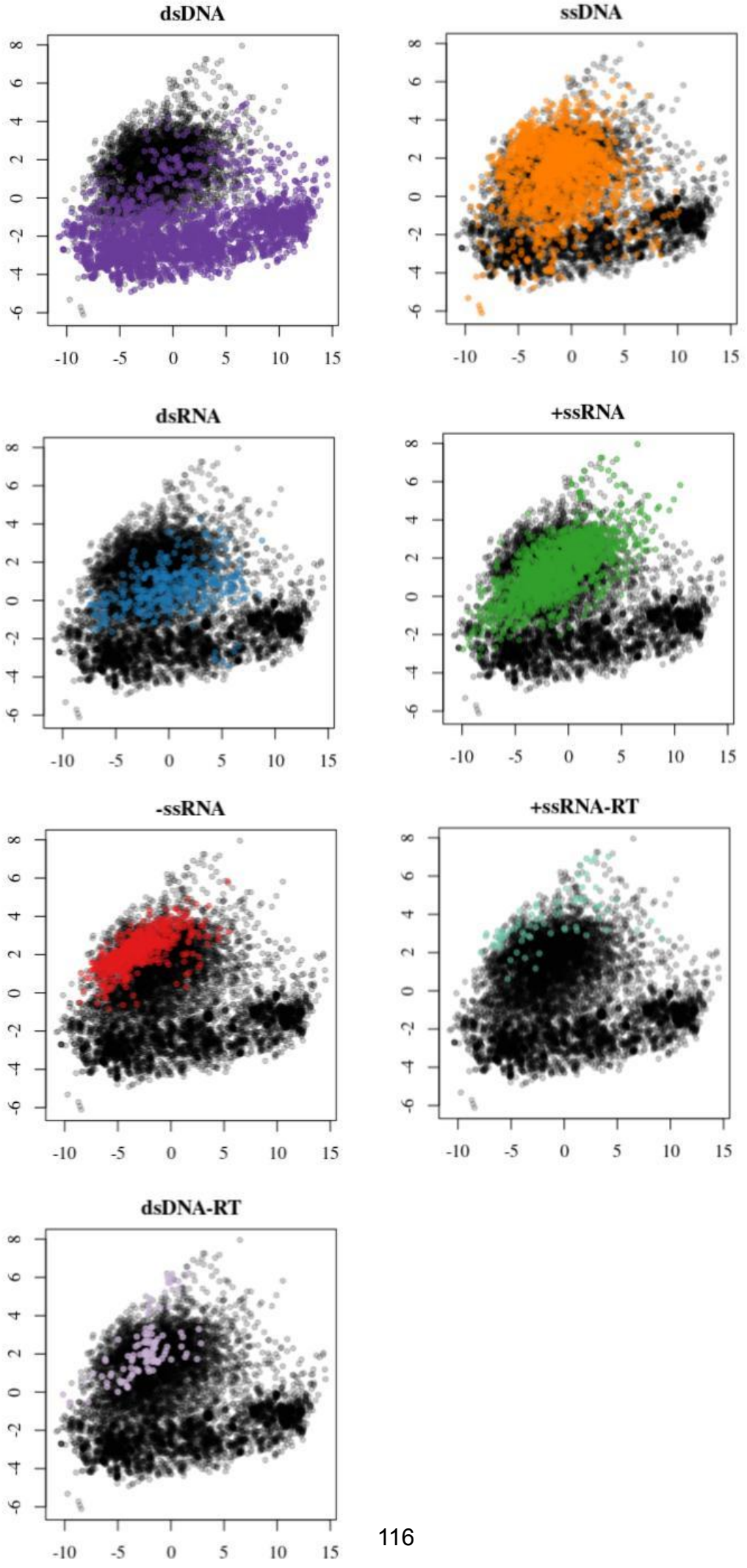


**B.** Frecuencia de codones por hospedero **(1)** y por Baltimore **(2)**

**(B1)**

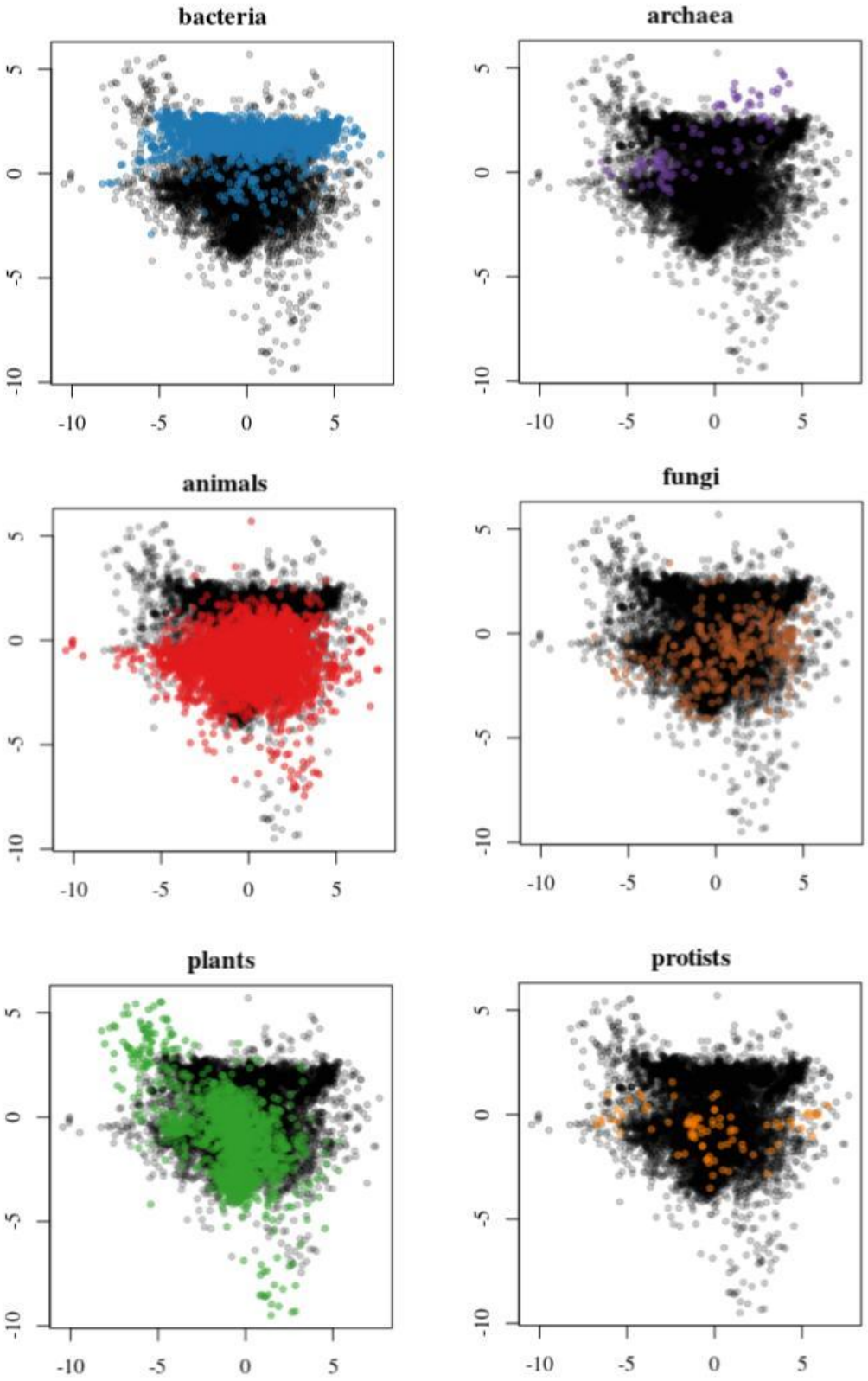


(B2)

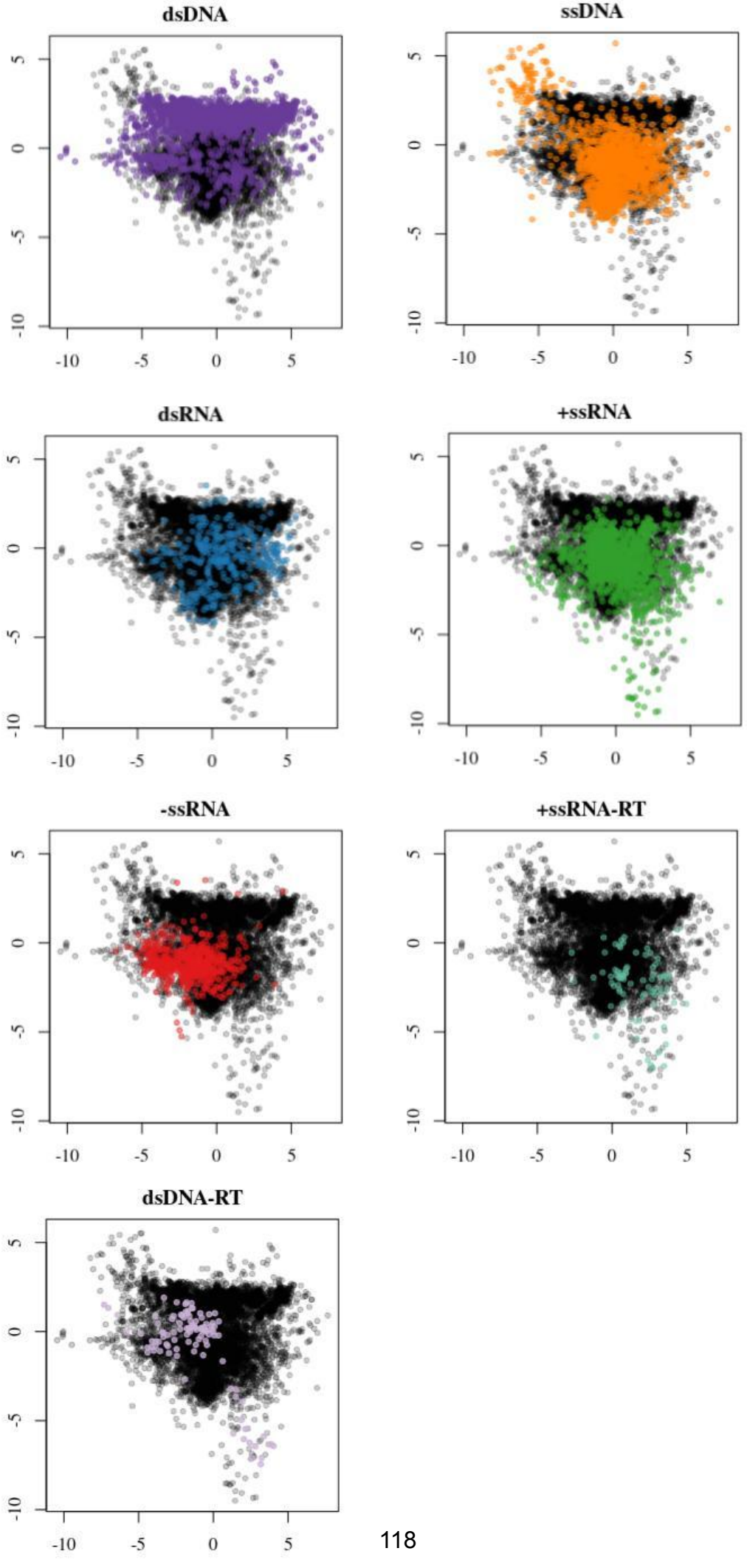

dsDNA
ssDNA
dsRNA
+ssRNA
-ssRNA
+ssRNA-RT
dsDNA-RT
-10 -5 0 5 10 15
-6 -4 -2 0 2 4 6 8

**C.** Frecuencia de aminoácidos por hospedero **(1)** y por Baltimore **(2)**

**(C1)**



(C2)



## (D) Distribuciones del contenido de G+C en sistemas eucariotas

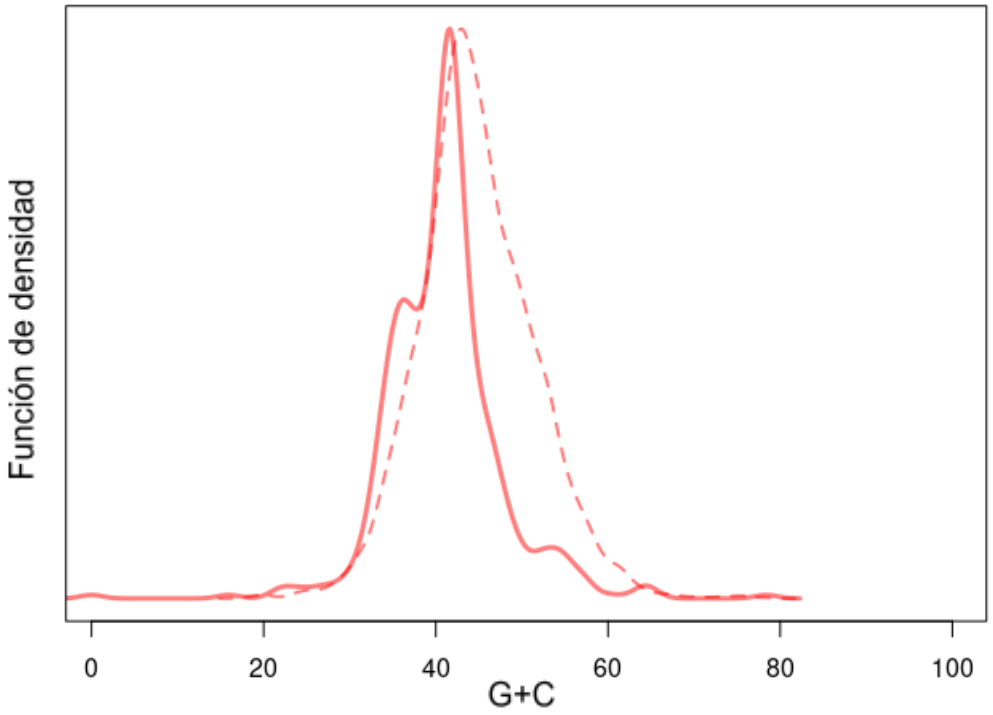



**Figura D1.** G+C en los eucariotas totalmente secuenciados (una especie por género, N = 496; **línea continua**) y en sus virus (línea punteada).

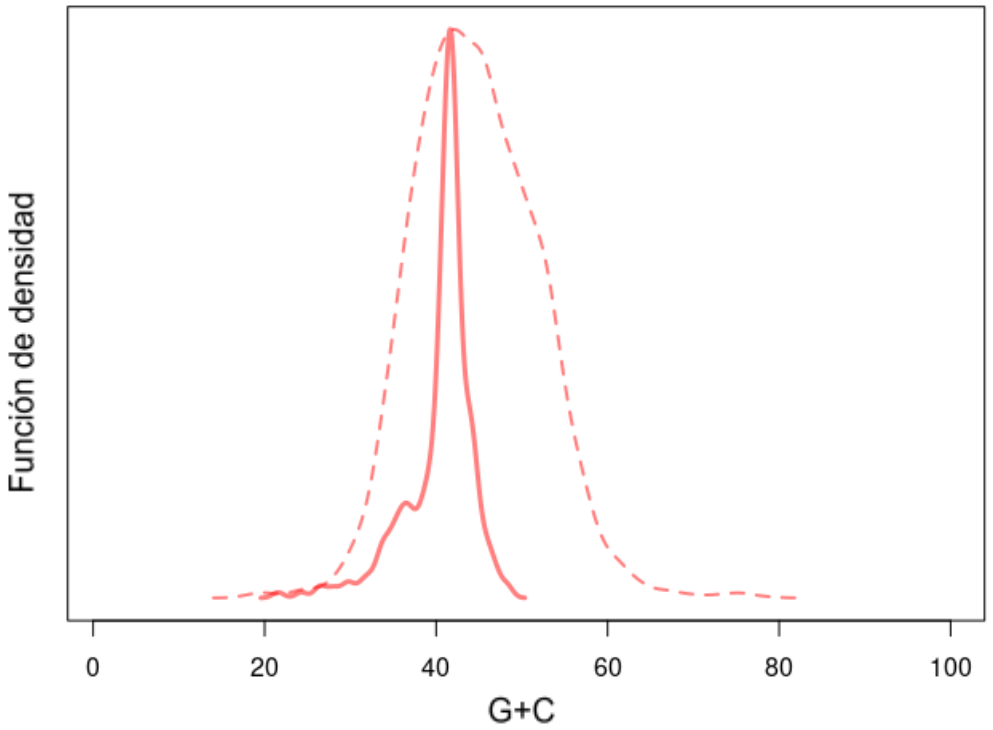



**Figura D2.** G+C en los animales totalmente secuenciados (una especie por género, N = 283; **línea continua**) y en sus virus (línea punteada).

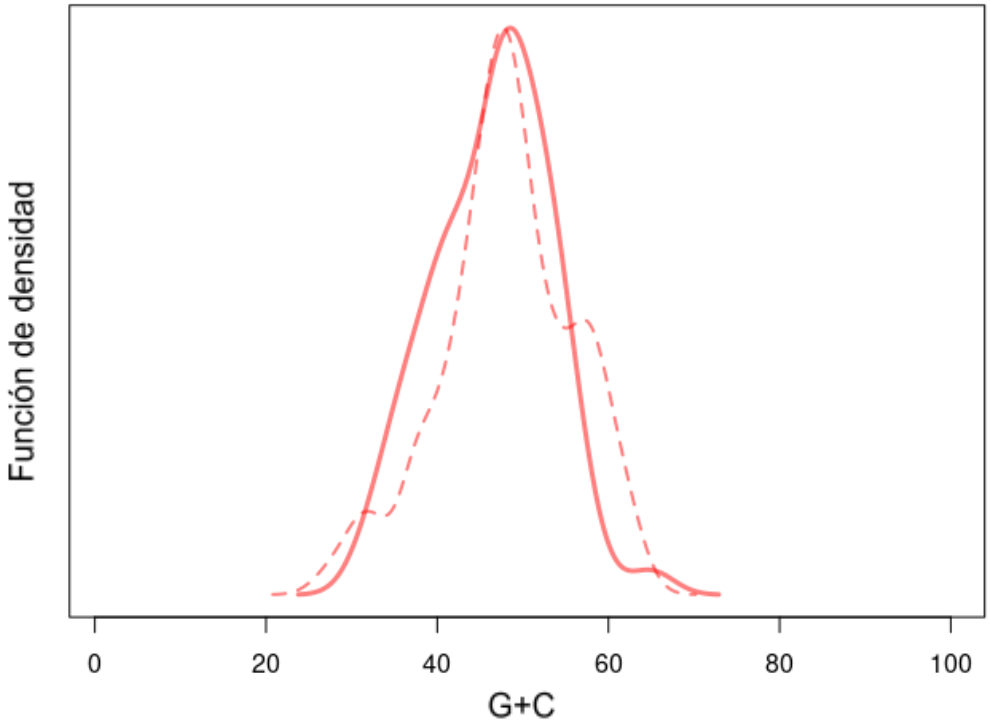



**Figura D3.** G+C en los hongos totalmente secuenciados (una especie por género, N = 62; **línea continua**) y en sus virus (línea punteada).

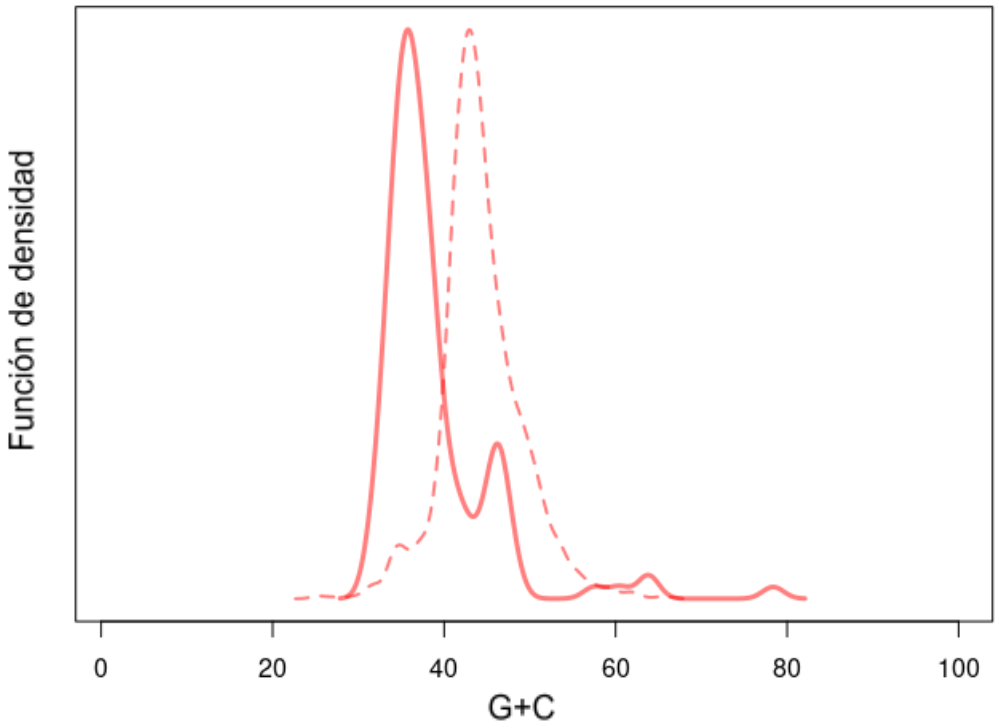



**Figura D4.** G+C en las plantas y algas vegetales totalmente secuenciadas (una especie por género, N = 128; **línea continua**) y en sus virus (línea punteada).

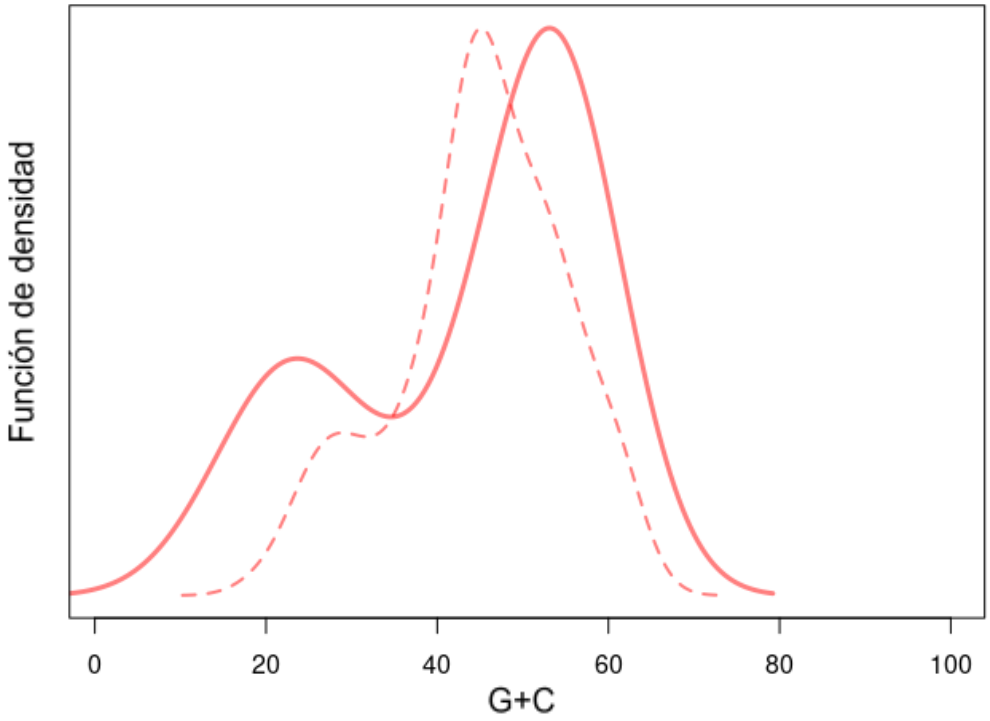



**Figura D5.** G+C en los protistas totalmente secuenciados (una especie por género, N = 17; **línea continua**) y en sus virus (línea punteada).

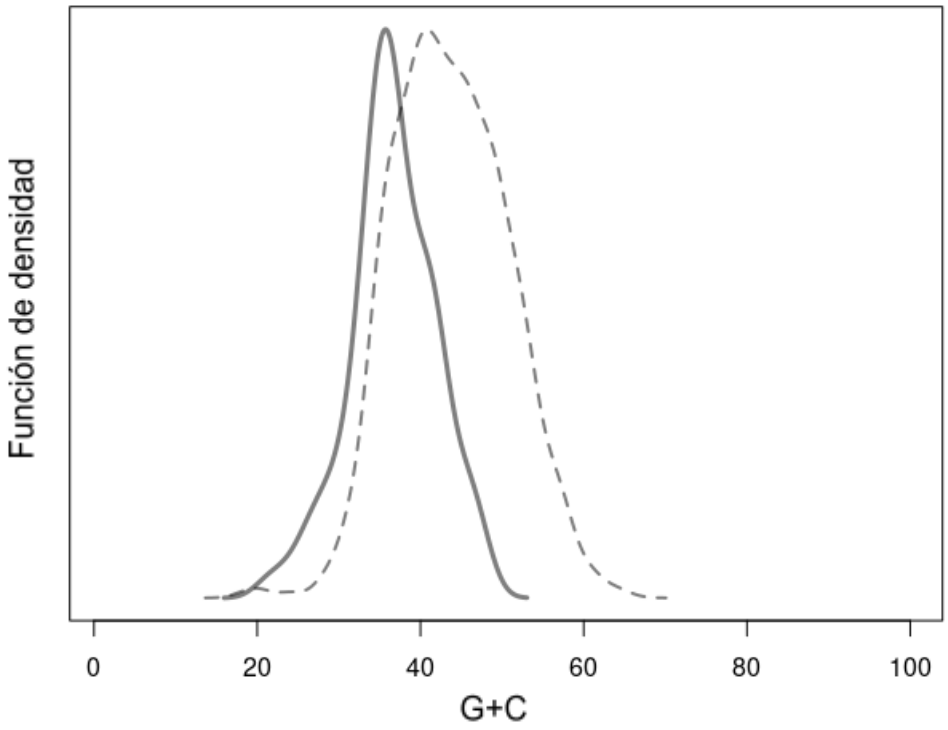



**Figura D6.** G+C en los invertebrados totalmente secuenciados (una especie por género, N = 78; **línea continua**) y en sus virus (línea punteada).

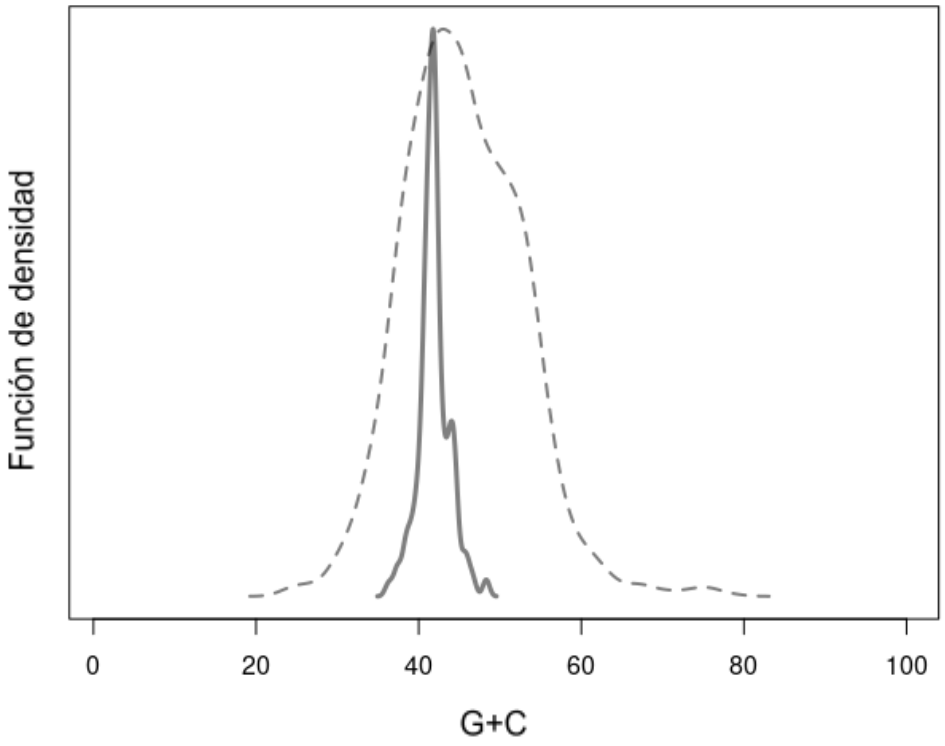



**Figura D7.** G+C en los vertebrados totalmente secuenciados (una especie por género, N = 205; **línea continua**) y en sus virus (línea punteada).

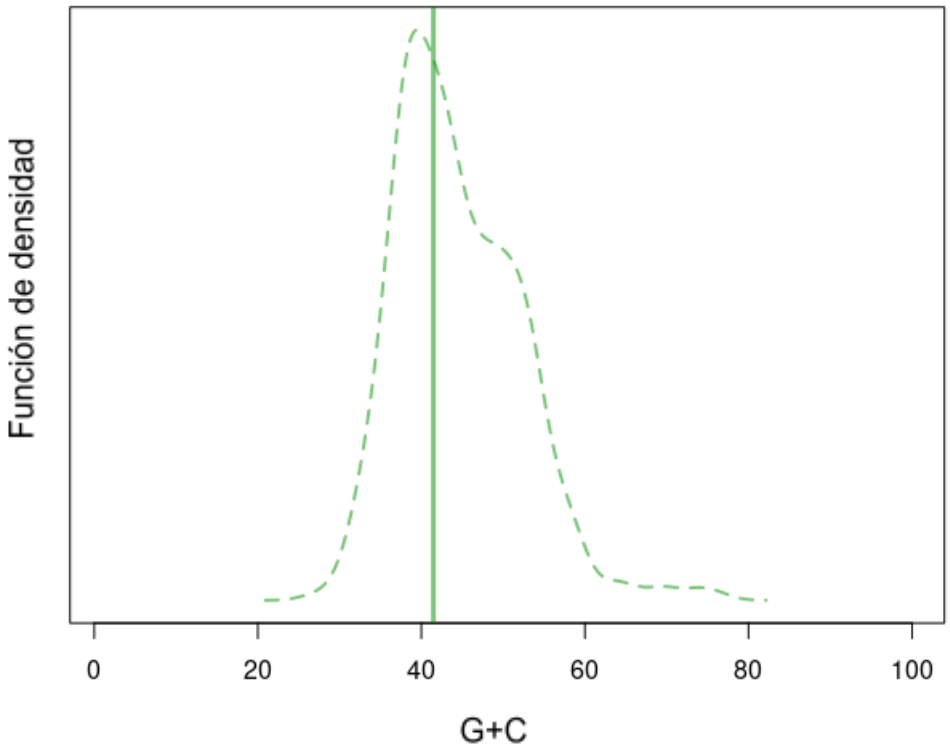



**Figura D8.** G+C (mediana) del genoma humano (N = 1; **línea vertical**) y en sus virus (línea punteada; N = 437)