UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO

COLEGIO DE CIENCIA Y TECNOLOGÍA

POSGRADO EN CIENCIAS GENÓMICAS

“Evolución molecular de los genes sobrelapados en el genoma del Virus de la Hepatitis B”

QUE PARA OPTAR EL GRADO DE:

MAESTRO EN CIENCIAS GENÓMICAS

P R E S E N T A:

Luis Enrique Soto Cortés

DIRECTORA:

Dra. Selene Zárate Guerra

Ciudad de México, 8 de diciembre del 2020

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

INTEGRACIÓN DEL JURADO:

Presidente:	Dra. Rosa Martha Eugenia Yocupicio Monroy, UACM.
Secretario:	Dra Claudia Selene Zarate Guerra, UACM.
Vocal:	Dr. José Alberto Campillo Balderas, UNAM.

Plantel de adscripción:

PLANTEL DEL VALLE, UACM.

DIRECTOR

Dra. Claudia Selene Zarate Guerra
Universidad Autónoma de la Ciudad de México

AGRADECIMIENTOS ACADÉMICOS

A la Universidad Autónoma de la Ciudad de México (UACM), por permitirme seguir formando como científico y por su apoyo para presentar resultados preliminares de este trabajo en el XI Congreso Nacional de Virología.

Al Posgrado en Ciencias Genómicas (PCG) de la Universidad Autónoma de la Ciudad de México (UACM), por la formación académica de calidad durante mis estudios de maestría.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por el otorgamiento de la beca de manutención (928277) para realizar la maestría.

A la Dra. Selene Zárate por aceptarme en su laboratorio, por su paciencia, constante guía y disposición de ayudar cuando lo necesitaba. Su gran conocimiento sobre la evolución molecular ha sido muy estimulante e inspirador para mí.

A la Dra. Martha Yocupicio Monrroy por aceptar ser mi asesora. Su apoyo y conocimientos fueron clave para mejorar este trabajo, por ser un ejemplo de trabajo arduo y constancia.

Al Dr. José Campillo Balderas, por aceptar ser mi asesor y por su apoyo para mejorar el trabajo de tesis. Gracias por las palabras de apoyo y por ser una inspiración de divulgación científica.

A la Dra. Elisa Azuara Liceaga, por aceptar ser mi lectora y sus comentarios que mejoraron este trabajo.

Al cuerpo académico del PCG, por impulsarnos a desarrollar habilidades necesarias para formarnos como científicos.

A mis compañeros del Laboratorio de Bioinformática, al M. en C. Fabricio Martínez, por guiarme cuando recién ingresé al laboratorio para realizar los primeros análisis, y al M. en C. José Caballero, por ser un referente de constancia y dedicación.

AGRADECIMIENTOS A TÍTULO PERSONAL

A mis padres y hermanos, por ser el pilar de mi vida.

A la Lic. Catalina Sánchez y la Dra. Lilia López del PCG por su apoyo durante la última etapa de la maestría.

A la Dra. Alejandra Serrato, porque gracias a ti la evolución es parte importante de mi interés científico. Por ser un ejemplo de pasión por la ciencia.

A la Dra. Rosaura Grether, por su apoyo antes de iniciar la maestría, por echarme porras para continuar con mi formación científica.

A Diego y Abraham, su amistad fue muy importante en este proceso. Gracias.

A Yanin y Carlos, gracias por muchas cosas. Por su apoyo constante, principalmente al final de la maestría

A la bandita universitaria, por las patoaventuras: Maga, Mel, Ivonne, Bruno, Rodrigo.

A mi colega Jose por ser un excelente compañero de laboratorio, tu compromiso y disciplina me motivaron.

A mis compañeros del PCG: Roger, Darío, Iván, Blanca, Fabricio, Zuly, Brisa, etc., por los momentos que compartimos juntos, las clases, las charlas “académicas” post-trabajo.

Gracias al pueblo mexicano, mediante sus impuestos fue posible realizar esta maestría.

A Socorro y Enrique, todos mis logros son para y gracias a ustedes

Contenido

ÍNDICE DE TABLAS

INDICE DE FIGURAS

ABREVIATURAS

ORFs= Marcos abiertos de lectura

ORF *P* = Marco abierto de lectura del HBV, codifica para la Polimerasa

ORF *S*= Marco abierto de lectura del HBV, codifica para las proteínas L, M yS.

ORF *C*= Marco abierto de lectura del HBV, codifica para las proteínas C y HBeAg

ORF *X*= Marco abierto de lectura del HBV, codifica para la proteína X.

HBV = Virus de la Hepatitis B

dN= Tasa de sustitución no sinónima

dS= Tasa de sustitución sinónima

HBcAg= Antígeno core de HBV

HBsAg= Antígeno S de superficie del HBV

HBeAg= Antígeno E del HBV

PCA= Análisis de Componentes Principales

RSCU= Uso Relativo de Codones Sinónimos

ENC= Número Efectivo de Codones

GC3= Contenido de GC en el tercer codón

RNA= Ácido ribonucleico

DNA= Ácido desoxirribonucleico

A= Adenina

T= Timina

C= Citocina

G= Guanina

U= Uracilo

RESUMEN

El virus de la hepatitis B (HBV) es miembro de la familia *Hepadnaviridae* y causa enfermedades hepáticas en todo el mundo. El genoma del HBV es circular, de DNA y parcialmente de doble cadena. Presenta cuatro marcos de lectura abiertos (ORFs): *S* codifica para las proteínas de superficie; *C*, codifica para la proteína Core y el Antígeno e de la hepatitis B (HBeAg); *P*, codifica para la polimerasa; *X*, codifica para la proteína regulatoria X. Todos los ORFs se encuentran sobrelapados en algún grado con el ORF *P*. A diferencia de las regiones no sobrelapadas, las regiones sobrelapadas presentan una mayor restricción al cambio evolutivo debido a que una mutación en un sitio nucleotídico afecta simultáneamente dos marcos de lectura. En el presente trabajo de investigación se caracterizaron y compararon las regiones sobrelapadas y no sobrelapadas del genoma del HBV a nivel de composición y uso de codones, así como la variabilidad genética y sitios candidatos bajo selección natural. Los resultados indican que las regiones sobrelapadas presentan un mayor contenido de GC cuya variación dentro del genoma influye en el uso de codones a lo largo del genoma de manera importante. La presión mutacional es el principal mecanismo que determina el sesgo en el uso de codones. La variabilidad genética es menor en estas regiones; sin embargo, las tasas de sustitución no sinónimas son más altas que las de regiones no sobrelapadas. Aparentemente, se debe a que las regiones sobrelapadas codifican para proteínas o dominios que están bajo constante presión selectiva por mecanismos del sistema inmune o por el efecto farmacológico. Es posible que la selección natural actúe a través de la eliminación de mutaciones deletéreas o fije aquéllas que aumenten la capacidad de replicación del HBV (adecuación biológica).

La evolución molecular de las regiones sobrelapadas en el HBV depende de tres aspectos principales: 1) tipo de sobrelape entre los dos ORFs; 2) rol funcional de los dominios de las proteínas codificados por las regiones sobrelapadas; y 3) presión mutacional sobre los ORFs con menor presión selectiva.

ABSTRACT

Hepatitis B virus (HBV) is a member of the *Hepadnaviridae* family and causes liver disease throughout the world. The HBV genome is circular, partially double-stranded DNA and has four open reading frames (ORFs): the *S* gene encodes for surface proteins; the *C* gene encodes for the Core protein and hepatitis B e Antigen (HBeAg); *P*, encodes for polymerase; *X*, encodes for the regulatory protein X. All the ORFs are overlapped with ORF P to some extent. The overlapping regions present a greater restriction to evolutionary change because a mutation in a nucleotide site simultaneously affects two reading frames. Consequently, the action of evolutionary processes acts differently compared to the non-overlapping regions.

In the present research work, the overlapping regions of HBV genome were characterized at the level of composition and codon use, as well as the genetic variability and candidate sites under natural selection, making a special comparison with the non-overlapping regions. The results indicate that the overlapping regions present a higher content of GC which its variation within the genome influences the use of codons throughout the genome in an important way, mutational pressure is the main mechanism that determines the bias in the use of codons. Genetic variability is lower in these regions. however, non-synonymous substitution rates are higher compared to its non-overlapping counterpart, this is because the overlapping regions encode for proteins or regions of them that are under constant selective pressure by elements of the immune system or pharmacological where the Natural selection acts to either eliminate deleterious mutations or fix those that increase the replication capacity of HBV (fitness).

The molecular evolution of the overlapping regions in HBV depends on three aspects: 1) Type of overlap between the two ORFs 2) Functional role of the domains of the proteins encoded by the overlapping regions. 3) Mutational pressure could play an important role in ORFs with lower selective pressure.

1. INTRODUCCIÓN

1.1. **Características generales de los genomas virales**

Los virus son agentes infecciosos obligados intracelulares en todos los seres vivos. Se estima que existen al menos 10^{31} partículas virales, por lo tanto, son las entidades biológicas más abundantes del planeta (Breitbart & Rohwer, 2005). Con base en el tipo de genoma y el modo de replicación (Baltimore, 1971), existen 7 grupos de virus: Grupo I en donde se encentran virus de DNA de doble cadena; Grupo II, virus de DNA de una sola cadena; Grupo III, virus de RNA de doble cadena; Grupo IV, virus de RNA de una sola cadena de polaridad positiva; Grupo V, virus RNA de una sola cadena de polaridad negativa; Grupo VI, virus de RNA de una sola cadena con actividad de retrotranscriptasa y el Grupo VII, virus de DNA parcialmente de DNA con actividad de retrotranscriptasa.

De manera general, los virus contienen genomas cuyo material genético puede ser de DNA o RNA. En ambos casos, la organización y estructura de los genomas varía ampliamente, pueden ser genomas de una o dos cadenas de ácido nucleico, circulares, lineales o segmentados. Los virus que presentan genomas de DNA de doble cadena son no segmentados, mientras que algunos genomas de DNA de una sola hebra se encuentran en múltiples segmentos. Por otro lado, los genomas virales de RNA son generalmente segmentados, principalmente los virus de RNA de una sola cadena (Chaitanya, 2019).

El tamaño de los genomas virales varía ampliamente, en el caso de los virus de DNA, el rango del tamaño de los genomas abarca alrededor de 4 órdenes de magnitud, desde 0.859 nt (circovirus porcino) hasta 2,473,870 nt (*Pandovirus*). Los virus de RNA poseen genomas universalmente pequeños, van desde un rango de 1,682 nt (virus de la hepatitis delta) hasta 31,526 nt (virus de la hepatitis murina) (Cui et al., 2014). Lo que es claro es que la distribución del tamaño de los genomas en el árbol de la vida no es al azar, existe mayor variación de tamaños genómicos en virus de DNA que en virus de RNA, desde virus de DNA de ~ 2000 kbp que infectan protistas y algas hasta de ~3 kbp que infectan a mamíferos. Por otro lado, los virus de RNA presentan tamaños de genoma mucho más pequeños y con menor diversidad de tamaños, interesantemente, los virus de RNA infectan principalmente eucariontes, esto sugiere que la distribución del tamaño de los genomas virales no

correlaciona con la antigüedad de los linajes de sus hospederos (Campillo-Balderas et al., 2015).

El tamaño de los genomas virales está restringido al tamaño de la cápside del virión (Cui et al., 2014). Generalmente, los virus de RNA y los virus de DNA de una sola cadena tienen genomas pequeños y, por lo tanto, cápsides pequeñas. Una propiedad intrínseca de gran parte de los virus es carecer de capacidad corrección de error en el proceso de la replicación, cada grupo de virus tiene una tasa mutacional por nucleótido, entre más grande es el genoma más probabilidad existe de generar mutaciones deletéreas, es decir, mutaciones que provoquen disminución en la capacidad replicativa y de propagación del virus. Esto último aplica principalmente para los genomas de RNA, ya que en general los virus de DNA presentan una menor tasa mutacional que les permite alcanzar tamaños de genomas más grandes (Holmes, 2009).

Finalmente, una consecuencia de la compactación genómica en los virus, es la abundancia de marcos abiertos de lectura (ORFs) sobrelapados, se ha reportado que existe una relación negativa entre el tamaño del genoma y la proporción de genes sobrelapados en virus de RNA (Figura 1). El sobrelapamiento ocurre como una estrategia evolutiva para aumentar la diversidad de proteínas codificadas en genomas pequeños. En las regiones sobrelapadas puede ocurrir un cambio mutacional que sea neutral en un marco de lectura y no neutral en el otro. Asimismo, se puede presentar un caso de pleiotropía ya que un sitio nucleotídico puede tener un impacto importante en la capacidad de los virus para replicarse, generando un costo para la capacidad adaptativa. Así pues, el sobrelapamiento de genes favorece la diversidad funcional ante las altas restricciones de tamaño en los genomas virales (Belshaw et al., 2007; Holmes, 2009).

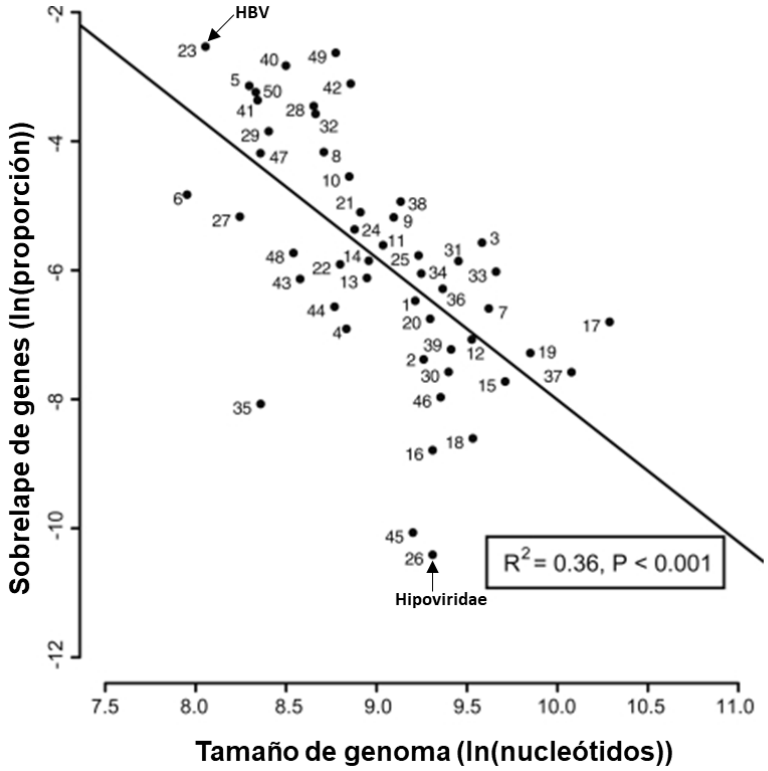


Figura 1 Relación negativa entre la proporción de sobrelapes de genes en genomas virales y el tamaño de su genoma, ambos expresados en logaritmo natural. Los puntos representan diferentes familias virales, principalmente de RNA; los virus con mayor y menor porcentaje de genes sobrelapados son señalados (Hipoviridae y HBV, respectivamente). Modificado de Belshaw et al. (2007)

1.2. **Procesos evolutivos en los genomas virales**

El cambio evolutivo es un proceso que resulta del éxito reproductivo diferencial de los individuos cuyas diferencias fenotípicas son heredables y en consecuencia se alteran las frecuencias de los genotipos correspondientes entre generaciones. La evolución es, por lo tanto, inevitable en poblaciones de entidades que (a) se autorreplican y (b) exhiben una variación hereditaria en los fenotipos que afectan la reproducción (Knipe et al. 2001). Los virus, al ser entidades replicativas autónomas que pueden presentar gran variación genética a nivel intrapoblacional, están también sujetos a los diferentes procesos evolutivos (Morley, & Turner, 2017). Consecuencia de la dinámica de los diferentes procesos evolutivos sobre las poblaciones virales es lo que permite las diferentes estrategias de sobrevivencia a corto y largo plazo de los virus ante las diferentes presiones selectivas. Esto ha generado la inmensa diversidad conocida en los virus en cuanto a su morfología,

estrategia de replicación, tipo de hospederos, tolerancias ambientales, así como en las estrategias de transmisión (Knipe et al. 2001).

1.2.1. Mutación y recombinación

La mutación y la recombinación se caracterizan por ser los únicos procesos evolutivos que generan diversidad genética en las poblaciones. Particularmente en los genomas virales, las tasas mutacionales y de recombinación son considerablemente altas, por lo que son elementos muy importantes a considerar en la dinámica evolutiva de las poblaciones de virus (Moya et al., 2004).

Las mutaciones surgen mediante tres mecanismos: (1) por efecto de mutágenos (luz UV, rayos X) sobre los ácidos nucleicos; (2) por daño simultáneo de las bases nitrogenadas de los ácidos nucleicos y (3) a través del fallo de las enzimas que replican y reparan los ácidos nucleicos. Los primeros dos mecanismos actúan de manera similar en los genomas virales, sin embargo, las variaciones en las tasas de mutación entre virus se deben principalmente a las diferencias en la fidelidad de las enzimas replicadoras del material genético. Así, los virus con polimerasas de alta fidelidad tienen tasas de mutación relativamente bajas y viceversa (Fleischmann WR Jr., 1996).

La tasa de mutación de un organismo se define como la probabilidad de que un cambio en la información genética se transmita a la siguiente generación. En los virus, una generación comúnmente se define como un ciclo de infección celular; esto incluye, la unión a la superficie celular, la entrada, la expresión génica, la replicación, la encapsidación y la liberación de partículas virales. El concepto de tasa de mutacional no es lo mismo que la frecuencia observada de mutaciones en una población viral determinada. Esta última es el producto de otros procesos como la selección natural, la deriva genética, la recombinación, etc., a lo que se conoce como diversidad genética (Sanjuán & Domingo-Calap, 2016)

Por otro lado, las tasas de mutación de los virus varían según la composición del genoma, el tamaño y la estructura. En general, los virus de RNA producen de 10^{-6} a 10^{-4} nuevas sustituciones de bases por nucleótido por ciclo replicativo, mientras que estas tasas varían de 10^{-8} a 10^{-6} en los virus de DNA. Además, los genomas virales de una sola cadena parecen mutar más rápido que los de doble

cadena; además, se ha demostrado una correlación inversa entre el tamaño del genoma y la tasa de mutación. (Sanjuán & Domingo-Calap, 2019)

La recombinación ocurre cuando al menos dos virus coinfectan a la misma célula hospedera e intercambian material genético. Existen diferentes tipos de recombinación viral basados en el sitio en donde se realiza el proceso de entrecruzamiento genético (crossing over). La recombinación homóloga ocurre en el mismo sitio en ambas hebras parentales; mientras que en la no homóloga, ocurre en diferentes sitios de los fragmentos genéticos involucrados, También existe un tipo de recombinación de reordenamiento llamado cambio antigénico (antigenic shift) que ocurre en virus con genomas segmentados, como el de la influenza, que pueden intercambiar segmentos genómicos completos dando lugar a nuevas combinaciones de segmentos genéticos en cada ciclo replicativo. La recombinación es un proceso común en los virus e influye de manera importante en su evolución. Por ejemplo, la recombinación se ha asociado con la expansión del rango de hospederos virales, la aparición de nuevos virus, el aumento de la virulencia y la patogenia, la modificación de tropismos tisulares, la evasión de la inmunidad del huésped y la evolución de resistencia a antiviral (Simon-Loriere & Holmes, 2011; Pérez-Losada et al., 2015).

1.2.2. Selección natural y deriva génica

La mutación genera múltiples variantes genéticas en una población llamadas polimorfismos. Los polimorfismos genéticos pueden llegar a aumentar su frecuencia y eventualmente fijarse o eliminarse en la población. La probabilidad de que una mutación generada recientemente sea fijada o eliminada depende de dos factores: (1) el grado en que la mutación aumenta o disminuye la capacidad de un individuo para sobrevivir y reproducirse (adecuación biológica) en el entorno actual, y (2) el tamaño de la población en cuestión, comúnmente denominada como “N”. Las mutaciones beneficiosas o ventajosas son aquéllas que aumentan la adecuación biológica de los individuos de la población en relación con aquellos individuos que no la tienen, mientras que las mutaciones deletéreas lo disminuyen y las mutaciones neutrales no tienen un efecto apreciable en éste (Pybus & Shapiro, 2012).

El efecto de la selección natural es aumentar la frecuencia de una mutación beneficiosa hasta que se fija en la población (selección positiva) o disminuir la frecuencia de una mutación perjudicial hasta que se elimina (selección negativa) . Por el contrario, la deriva genética ocasiona que las frecuencias de todos los tipos de mutaciones fluctúen aleatoriamente a lo largo del tiempo, sin una tendencia a aumentar o disminuir, hasta que las mutaciones se fijen o eliminen. El tamaño de las fluctuaciones aleatorias es mayor si la N es pequeña y menor si la N es grande. En la (Pybus & Shapiro, 2012; Sanjuán & Domingo-Calap, 2019).

1.3. **Composición y uso de codones en los genomas virales**

La composición de los genomas cambia en función de las dinámicas de diferentes procesos evolutivos como mutaciones de bases (por ejemplo, el cambio de C a T durante la replicación), pérdida de fragmentos largos o cortos de nucleótidos (incluidos genes), replicación de nucleótidos, reordenamientos, recombinación, duplicación, transformación, conjugación y/o transducción (Bohlin & Pettersson, 2019).

El contenido de nucleótidos varía considerablemente entre grupos de virus, inclusive dentro de la misma familia viral, por ejemplo, los lentivirus y deltaretrovirus pertenecen la familia *Retroviridae*, mientras que los lentivirus presentan genomas que son ricos en Adenina (A) y bajos en citosina (C), los deltaretrovirus, como el virus de la leucemia de células T humanas (HTLV) poseen genomas pobres en A y ricos en C (van der Kuyl & Berkhout, 2012). Por otro lado, miembros de la familia Adenoviridae, virus de DNA de doble cadena, presentan una predominancia de citocina (C) y guanina (Reddy et al., 1998; Sprengel et al., 1994).

El sesgo de uso de codones se define como a las diferencias en frecuencia de codones sinónimos en una secuencia de DNA codificante. Se han propuesto dos hipótesis para explicar el sesgo en uso de codones, la primera menciona que lo establece la eficiencia de traducción, los genes correspondientes a proteínas cuya expresión sea del tipo constitutiva y/o frecuente, cuyo uso de codones debe ser similar al de la célula hospedera, mientras que los genes de las proteínas que deben expresarse bajo condiciones restrictivas podrían involucrar codones que no son comúnmente utilizados por la célula hospedera. La segunda hipótesis se refiere a la noción de que el sesgo en el uso de codones existe debido a restricciones

genéticas inherentes (por ejemplo, contenido de GC) y a sesgos en los eventos mutacionales (Hershberg & Petrov, 2008) .

Una de las posibles aplicaciones del análisis del contenido de nucleótidos de los genomas virales es para determinar el impacto potencial de las restricciones de composición en el uso de codones. Por ejemplo, en el genoma del virus del Zika (ZIKV), la A y G se presentan en mayor frecuencia que U y C, los contenidos promedio de GC y AU fueron 46.75 y 53.24%, respectivamente. En la tercera posición de los codones, G y A presentaron alta frecuencia, mientras que U fue el menor entre ellos. Esto sugiere que en el genoma del ZIKV el contenido de A y G influye en el uso de los codones por restricción en la composición, debido a que variación de estos nucleótidos modificaría considerablemente la frecuencia de codones sinónimos (Butt et al. 2016)

Existen una gran variedad de índices para medir el sesgo de codones en secuencias codificantes. Aquí se mencionan tres de las métricas más utilizadas en los genomas virales: la frecuencia relativa de uso de codones sinónimos (RSCU, por sus siglas en inglés), el número efectivo de codones (ENC) y la gráfica de relación ENC vs el contenido de GC en la tercera posición (GC3) (Roth et al., 2012).

La frecuencia relativa de uso de codones sinónimos (RSCU) de un codón se define mejor como la relación de su frecuencia observada con su frecuencia esperada siempre que todos los codones que codifican un aminoácido particular se utilicen igualmente (Paul M. Sharp & Li, 1986). Los valores de RSCU para todas las secuencias codificantes de los genomas se calculan para determinar los patrones de uso de codones sinónimos sin la influencia de la composición de aminoácidos o la longitud de secuencia. Los valores de RSCU se estiman de la siguiente manera (Paul M. Sharp & Li, 1986):

$$RSCU_i = \frac{X_{ij}}{\frac{1}{n}\sum_{i=1}^{n} X_{ij}} n_i$$

donde X_{ij} es el número observado del i-ésimo codón para el j-ésimo aminoácido que tiene n_i clases de codones sinónimos. Los codones sinónimos con los valores RSCU de <1.0, 1.0 y >1.0 representan el sesgo de uso de codones subrepresentados (codones menos abundantes), sin sesgo (uso igual de todos los codones sinónimos) y la presencia de uso de codones sobrerrepresentados (codones abundantes), respectivamente.

El número efectivo de codones (ENC) es el número total de codones diferentes usados en una secuencia y se calcula utilizando la siguiente fórmula (Wright, 1990):

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{2}{F_4} + \frac{3}{F_6}$$

donde F (i = 2, 3, 4, 6) es la media de los valores de F_i para el aminoácido degenerado i veces. Los valores de Fi se calculan mediante la fórmula general:

$$F_i = \frac{n \sum_{j=1}^{i} (\frac{n_j}{n})^2 - 1}{n - 1}$$

donde n es el número total de apariciones de los codones para ese aminoácido y n_j es el número total de apariciones del j-ésimo codón para ese aminoácido.

Los valores de ENC oscilan entre 20 y 61 (Wright, 1990). Un valor ENC de 20 indica un sesgo extremo de uso de codones (solo se usa uno de los posibles codones sinónimos para el aminoácido correspondiente); mientras que el 61 indica que no hay sesgo en absoluto (todos los posibles codones sinónimos se utilizan para el correspondiente aminoácido). En consecuencia, cuanto menor sea el valor de ENC, mayor será el grado de sesgo de uso de codones.

Finalmente, el gráfico ENC vs GC3 proporciona información cualitativa sobre la influencia de la presión de mutación y la selección natural en el sesgo de uso de codones. Aquí, los valores ENC corresponden a los valores del eje “Y” y GC3 (frecuencia de ya sea una guanina o citosina en la posición del tercer codón de los codones sinónimos, excluyendo Met, Trp y codones de paro) corresponden al eje X (Wright, 1990). El uso de codones está limitado solo por el sesgo mutacional de G + C cuando los valores de ENC predichos caen debajo o alrededor de la curva estándar (relación funcional entre ENC y GC3 esperados). De lo contrario, otros factores como la selección natural, el plegamiento del RNA y la deriva genética juegan un papel importante en la configuración del sesgo de uso de codones (Kumar et al., 2018).

1.4. **El virus de la hepatitis B: características generales**

El virus de la hepatitis B (HBV) fue identificado por primera vez por Baruch Blumberg en los 1960's (Blumberg et al. 1965). Este virus es el causante de enfermedades hepáticas alrededor del mundo; se estima que aproximadamente 250-260 millones de individuos están infectados crónicamente y un tercio de la población mundial tiene evidencia serológica de exposición. Esto convierte al HBV en un problema de salud pública global con niveles endémicos de infección en el sureste asiático y en África, con una tasa de prevalencia de al menos el 8% en muchas poblaciones. Además, diversos obstáculos no han permitido la erradicación de la infección por HBV: sesgos en la cobertura de vacunación, largos periodos entre la vacunación y sus efectos en la prevalencia de la población, así como la falta de una cura definitiva. Por lo anterior, es necesario tener un entendimiento detallado y robusto del genoma, estructura, diversidad y evolución del HBV para poder encontrar nuevos y/o mejores métodos que contribuyan a la eliminación de este problema de salud pública (McNaughton et al. 2019).

El HBV se transmite a través de la exposición a sangre y fluidos corporales infectados (en particular, semen y secreciones vaginales). Aunque se ha detectado el HBV en la saliva, las lágrimas, la leche materna, el sudor y la orina, existe una mínima evidencia de transmisión a través de la exposición a estos líquidos donde no hay sangre, y no se ha demostrado que la lactancia materna aumente el riesgo de infección (Zheng et al. 2011). La mayoría de las infecciones en todo el mundo se adquieren por transmisión perinatal al nacer, por transmisión horizontal a / entre niños pequeños, por contacto sexual y por consumo de drogas inyectables (Lok y McMahon 2009).

La epidemiología de la hepatitis B se puede describir en términos de la prevalencia del antígeno de superficie de la hepatitis B (HBsAg) en una población, clasificada en alta (> 8% de prevalencia de HBsAg), intermedia (2% -7%) y áreas de baja prevalencia (<2%). Entre los países del continente asiático existe una alta prevalencia, que también se presenta entre las poblaciones indígenas del Amazonas en Sudamérica y Alaska en Norteamérica. En contraste, existen zonas de endemicidad intermedia en Europa y baja en países como Estados Unidos y Canadá (MacLachlan & Cowie, 2015).

México se ha considerado como una zona de baja endemia, no obstante, se ha demostrado que en el país existen zonas de alta prevalencia, principalmente en poblaciones indígenas, al igual que en Centro y Sudamérica. Recientemente se ha descrito una alta frecuencia de hepatitis B oculta en poblaciones indígenas en México (Roman et al., 2010). Se han detectado zonas geográficas de alta prevalencia del antígeno de la proteína core (HBcAg) pero bajas prevalencias del HBsAg. Los estudios epidemiológicos más recientes y de revisión de la literatura muestran que en el país hay por lo menos tres millones de personas adultas que se han infectado (anti-HBcAg positivos) y de estos un mínimo de 300 000 portadores activos (HBsAg positivos) podrían requerir tratamiento. No obstante, si consideramos a la población indígena como zona de alta endemia, el número de pacientes que se han infectado podría aumentar hasta 7 u 8 millones de mexicanos y a cerca de un millón de portadores activos (Escobedo-Meléndez et al., 2011).

El HBV es el virus prototipo de la familia *Hepadnaviridae*, según el ICTV.Los virus de esta familia pertenecen al grupo VII (Baltimore, 1971) e infectan únicamente a animales. De manera clásica, los hepadnavirus se clasifican en dos géneros: *Orthohepadnavirus*, cuyos virus infectan mamíferos, y *Avihepadnavirus*, cuyos virus infectan aves. No obstante, como se puede observar en la Fig. 2A se han descubierto hepadnavirus que infectan tanto a peces como a anfibios, lo que sugiere que su diversidad puede ser aún mayor (Dill et al. 2016; Lauber et al. 2017).

Con base en la diversidad de la secuencia genómica, el HBV es clasificado en 8 genotipos principales: A, B, C, D, E, F, G, H (Fig. 2B). Cada genotipo tiene una distribución geográfica definida, sin embargo, dependiendo de la dinámica evolutiva entre los genotipos y dentro de ellos existen diversas variantes recombinantes y subgenotípicas (Kurbanov, et al 2010). Los genotipos A y D se distribuyen principalmente en Europa y Oceanía, mientras que los genotipos B y C tienen alta incidencia en diferentes partes de Asia. En África, el genotipo predominante es el E, principalmente en la parte occidental del continente. En E.U.A. y Canadá predominan los genotipos A, B, C y D, mientras que en el caribe se distribuye principalmente el A. En Latinoamérica, los genotipos con principal incidencia son F, G y H, predominando en general el genotipo F en toda la región. Sin embargo, la incidencia de los genotipos en Brasil y México difieren del resto de Latinoamérica. En Brasil, particularmente predominan los genotipos A y D, mientras que en México, los genotipos principales son el G (10.2%) y H (63.3%). Con lo anterior, es

claro que la distribución del HBV es cosmopolita, sin embargo, se muestran patrones similares entre países de la misma región del mundo pero varían en diferentes partes del mundo (Velkov et al., 2018).

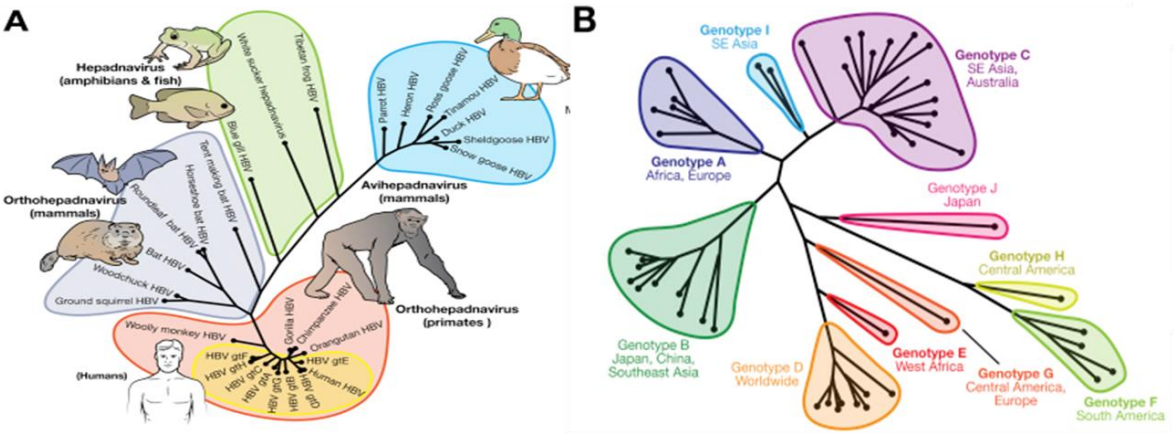


Figura 2 Relaciones evolutivas y distribución geográfica del HBV. A) Árbol filogenético en donde se muestra las relaciones evolutivas entre el HBV de humano y otros miembros de la familia *Hepadnaviridae*. B) Distribución geográfica de cada genotipo del HBV (Modificado de McNaughton et al. 2019).

El HBV es un virus esférico con un diámetro de aproximadamente 42 nm, presenta una cápside icosaédrica (Figura 3). La cápside está envuelta con una membrana lipoproteica hecha de tres proteínas de envoltura: *large* (L), *middle* (M) y *small* (S). Estas proteínas se encuentran en una proporción aproximada de 1:1:4. La cápside contiene un genoma circular de DNA parcialmente de doble cadena, el cual está covalentemente unido a la polimerasa en el extremo 5’ de la cadena negativa cadena (Seeger et al., 2013).

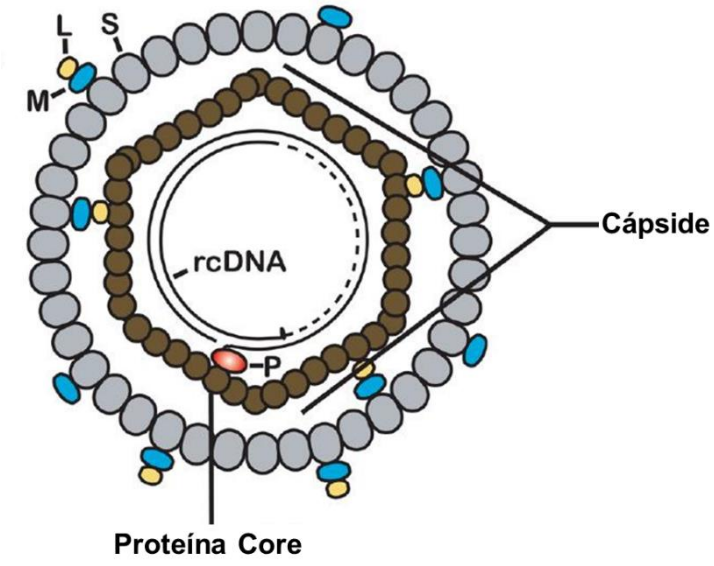


Figura 3. Representación esquemática del virión del HBV. L(Large), M (Medium) y S (Small) corresponden a las proteínas de superficie, la envoltura viral está representada con los círculos grises, seguido de cápside constituida por dímeros de la proteína Core, en este último compartimento se almacena el material genético en forma de DNA circular relajado (rcDNA), en el extremo 5' del genoma se une covalente mente la polimerasa viral (P). (Modificado de Lamontagne et al., 2016).

1.4.1. Organización y expresión genómica del HBV

El genoma del HBV presenta cuatro ORFs que codifican para siete proteínas: cuatro estructurales [proteínas de superficie (L, M y S) y la proteína Core] y tres no estructurales [polimerasa (P), la proteína X (X) y el antígeno e (HBeAg)]. Todos los ORFs se sobrelapan con algún otro ORF de alguna manera, el ORF S se encuentra en su totalidad embebido en el ORF P siendo éste el fragmento sobrelapado más grande de los virus animales. El ORF P abarca dos tercios del genoma y se sobrelapa con todos los demás genes. (Rancurel et al. 2009).

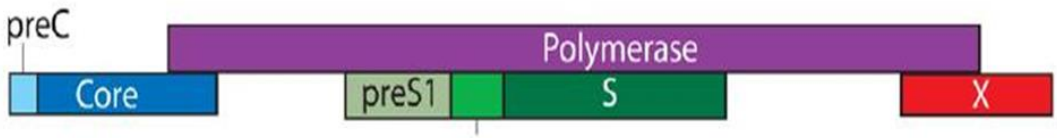


Figura 4 Representación lineal de los ORFs presentes en el genoma del HBV. El ORF P en morado, el ORF C en azul, el ORF S en verde y el ORFX en rojo. Algunos ORFs tienen diferentes regiones que codifican para dominios proteicos funcionalmente distintos, PreC y Core en el caso del ORF C y PreS1, PreS2y S en el ORF S (Modificado de Lamontagne et al., 2016).

La expresión de los ORFs da origen a cuatro transcritos de diferente tamaño (Figura 5A). E ORF C, que presenta las regiones Precore (PreC) y Core, codifica para dos transcritos de aproximadamente 3.5 kbp, el más grande de ellos se denomina Precore RNA y codifica para proteína Precore que mediante proteólisis en la parte N terminal, origina el HBeAg. El otro transcrito denominado RNA pregenómico (pgRNA) se utiliza ya sea como cadena molde para el proceso de retrotranscripción y origina el DNA genómico del HBV, o bien, se traducen las proteínas polimerasa y Core. El ORF S se divide estructural y funcionalmente en las regiones pre-S1, pre-S2 y S, a partir de este ORF se generan dos transcritos, uno de 2.4 kbp y otro de 2.1 kbp, del primero se traduce la proteína L, mientras que del segundo las proteínas M y S (Figura 5B). Esto es posible al uso de diferentes codones de inicio de la traducción mediante el mecanismo de leaky scanning ribosomal. La función del HBeAg permanece en gran medida indefinida. Sin embargo, con base en diversos estudios en animales, es evidente que la HBeAg no está involucrada en la infección viral, la replicación y el ensamblaje, pero es importante para la infección natural *in vivo* (Kramvis et al. 2018).

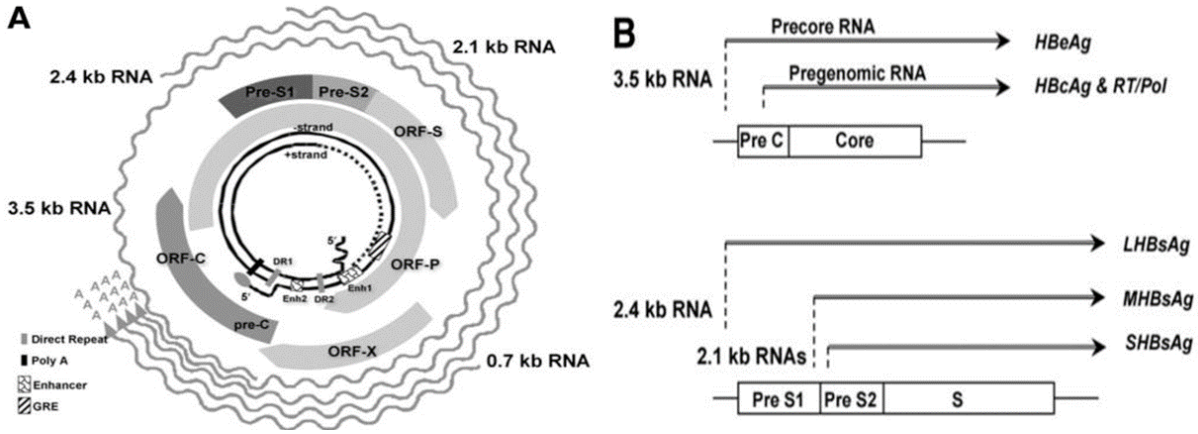


Figura 5. Características del genoma del virus de la hepatitis B. A) Organización genómica general, se muestran los diferentes ORFs así como sus cuatro transcritos principales B) Sitios de inicio de la transcripción y las proteínas que codifican, tanto para el ORF C (arriba) como para el ORF S (abajo) (Modificado de Liang, 2009).

La polimerasa (Pol) es una proteína de 84 kDa de aproximadamente 800 aminoácidos, es codificada por el ORF P y está dividida funcionalmente en cuatro dominios. El dominio de la proteína terminal (TP) participa en la encapsidación y el

inicio de la síntesis de la cadena negativa; el dominio Spacer conecta los dominios TP y RT, sin embargo, carece de alguna otra función definida; el dominio de la transcriptasa reversa (RT) cataliza la síntesis del genoma de DNA; y el dominio ribonucleasa H (RH)degrada el RNA pregenómico y facilita la replicación.

El ORF X codifica una proteína de 16.5 kDa de aproximadamente 154 aminoácidos. Tiene múltiples funciones, incluida la transducción de señales, la activación transcripcional, la reparación del DNA y la inhibición de la degradación de proteínas. El mecanismo de esta actividad y la función biológica de HBxAg en el ciclo de vida del virus sigue siendo en gran parte desconocido (Bouchard & Schneider, 2004). Sin embargo, está bien establecido que el HBxAg es necesario para la infección productiva del HBV *in vivo* y puede contribuir al potencial oncogénico del HBV (Liang, 2009).

1.4.2. Ciclo de replicación e historia natural del HBV.

La replicación del genoma del HBV se puede dividir ampliamente en tres fases: (I) Los viriones infecciosos contienen en su núcleo icosaédrico interno el genoma como un DNA parcialmente bicatenario, circular pero no cerrado covalentemente de aproximadamente 3.2 kb de longitud (rcDNA), (II) tras la infección, el rcDNA se convierte, dentro del núcleo de la célula hospedera, en un DNA circular covalentemente cerrado (cccDNA) similar a un plásmido, también conocido como unidad replicativa o episoma, (III) a partir del cccDNA, varios RNA genómicos y subgenómicos son transcritos por la RNA polimerasa II celular; de estos, el RNA pregenómico (pgRNA) se empaqueta selectivamente en cápsides de progenie y se transcribe de manera inversa por la proteína P presente en nuevos genomas de DNA-RC (Lamontagne et al. 2016). En el esquema de la Figura 6 se representan los 12 procesos principales. del ciclo de replicación del HBV.

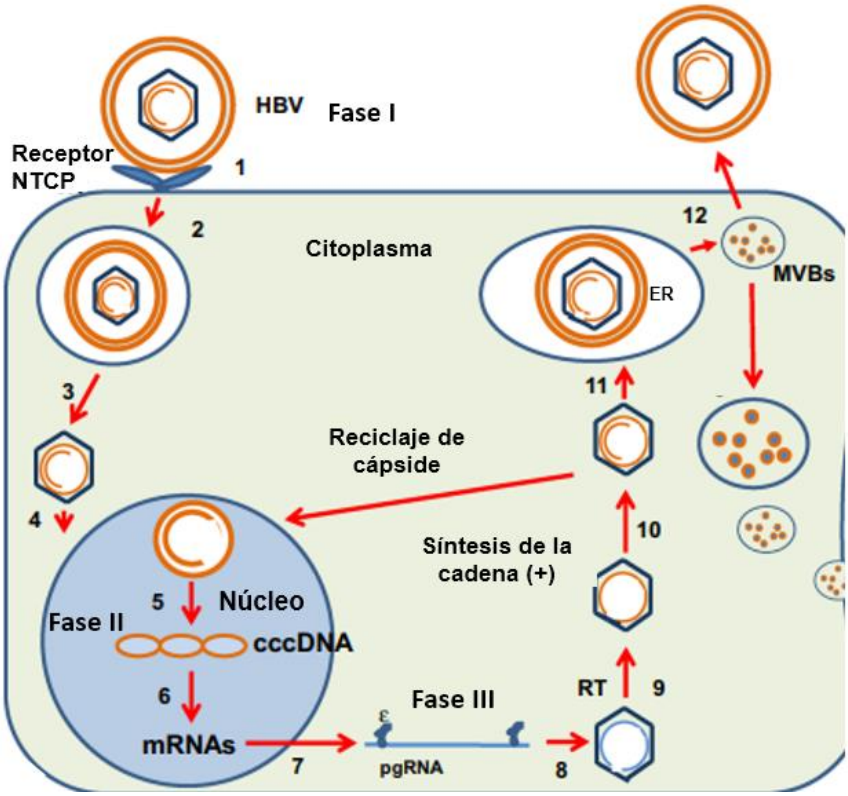


Figura 6. Ciclo replicativo del HBV. 1 Unión de la partícula viral al receptor NTCP, 2 Endocitosis, 3 Liberación de cápside, 4 Entrada de rcDNA al núcleo, 5 Síntesis de cccDNA, 6 Transcripción; 7 transferencia de RNAm al citoplasma; 8 Encapsidación; 9 Síntesis de cadena (-) de DNA por RT, 10 Síntesis de cadena (+) de DNA; 11 Entrada de viriones al lumen de RE, 12 Liberación de virus a través de un cuerpo multivesicular (MVB). (Modificado de Karayiannis,2017).

Tras la infección, el período de incubación del HBV varía de 28 a 180 días. En la mayoría de infecciones, el periodo de incubación es de 60 ± 110 días, El HBV suele causar hepatitis aguda, caracterizada por necrosis de las células hepáticas e inflamación periportal. La insuficiencia hepática es la complicación más grave de la hepatitis viral, que se presenta con coma. Menos del 1% de los pacientes desarrollan hepatitis fulminante. Desde un punto de vista serológico, el HBsAg se convierte en detectable en el suero 2 ± 8 semanas antes de cualquier elevación de aminotransferasas (ALT).

A medida que avanza la infección aumentan los niveles de ALT, se detectan fácilmente otros productos virales, como la DNA polimerasa viral y el HBeAg. La IgM anti-HBcAg son detectable al comienzo de la enfermedad clínica, seguida de la IgG anti-HBcAg. El HBcAg se encuentra solo en las células del hígado. A medida que avanza la recuperación y la convalecencia, aparecen en la sangre anticuerpos

contra otras proteínas virales. Cuando aparece anti-HBe, el HBeAg se vuelve indetectable. Los anti-HBsAg aparecen tardíamente en la convalecencia, después de 6 semanas a 6 meses el HBsAg desaparece. Si los antígenos permanecen detectables después de 6 meses, se considera que el paciente es portador de hepatitis B (Juszczyk, 2000). Específicamente, se vuelve un estado crónico de la infección y puede evolucionar a daño hepático progresivo hasta en 15 a 40% de casos, con desarrollo de cirrosis e insuficiencia hepática y/o hepatocarcinoma.

2. ANTECEDENTES PARTICULARES

2.1. **Características de los genes sobrelapados**

Los genes sobrelapados, se definen como regiones nucleotídicas ubicadas en un mismo locus que codifican para más de una proteína mediante el uso de diferentes marcos abiertos de lectura (Wei & Zhang, 2015). Este tipo de genes se encuentran presentes en el genoma de organismos pertenecientes a los tres dominios del árbol de la vida: Archea (Gomes-Filho, 2015), Bacteria (Rogozin et al., 2002) y Eukarya (Makałowska et al., 2007) así como en el genoma mitocondrial (Taanman, 1999). Sin embargo, y como se hablará con mayor detalle en el siguiente apartado, es en el genoma de los virus donde los genes sobrelapados tienen una mayor distribución, aproximadamente el 75% de los virus conocidos los presentan y se cree que fueron las restricciones físicas impuestas por la cápside de los viriones que generaron contracción del tamaño en el genoma y como mecanismo adaptativo se originaron las superposiciones (Chirico, et al., 2010).

Los genes sobrelapados presentan características muy particulares que los convierten en candidatos interesantes para su estudio evolutivo, en primer lugar codifican proteínas originadas de *novo* mediante un proceso denominado “overprinting”, en el cual mutaciones en un marco de lectura ancestral da lugar a la expresión de un segundo sin eliminar el primero (Keese & Gibbs, 1992; Sabath et al. 2012). Por otra parte, los genes sobrelapados representan un caso de conflicto adaptativo, es decir, las presiones selectivas serán diferentes en cada marco de lectura e incluso a nivel de sitio nucleotídico y al mismo tiempo existen restricciones evolutivas dependientes entre ambas proteínas codificadas (Mizokami et al.,1997; Pavesi, 2018).

De manera general, y como se puede observar en la Figura 7, los genes sobrelapados pueden ser clasificados en dos tipos: (1) genes sobrelapados en la misma cadena, los cuales se transcriben de la misma hebra de ácido nucleico (también conocidos como sobrelapes paralelos o unidireccionales); (2) genes sobrelapados en diferentes cadenas, los cuales son transcritos de las dos cadenas complementarias del DNA, también conocidos como sobrelapes antiparalelos (Sanna et al. 2008). El gen de referencia en un par de genes superpuestos se llama fase 0, los sobrelapes de la misma cadena pueden estar en dos fases (1 y 2). De

acuerdo al tipo de organismo, existen diferencias en la abundancia y distribución de sobrelapados paralelos o antiparalelos. En genomas bacterianos y virales los sobrelapes unidireccionales son los que predominan (Johnson and Chisholm, 2004), en los genomas eucariontes como el de los humanos y ratones alrededor del 90% de los genes sobrelapados son antiparalelos (Sanna et al. 2008; Saha et al. 2014).

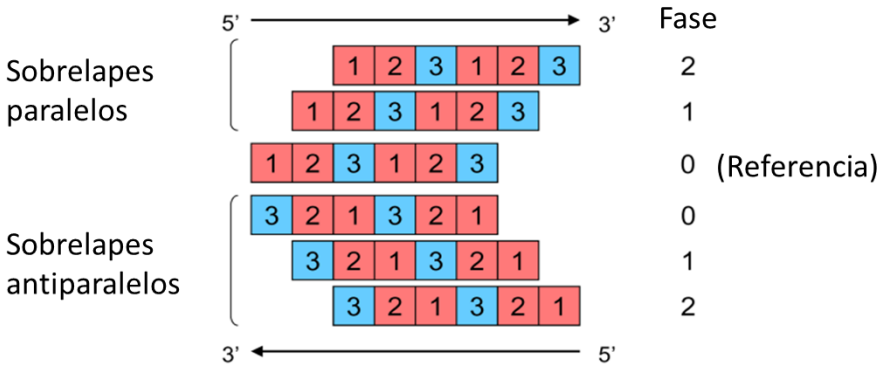


Figura 7. Fases y orientaciones de genes sobrelapados. Sobrelapes paralelos van en una dirección 5' a 3' y los antiparalelos de 3' a 5'. A partir de un marco de lectura de referencia (fase 0) existen otras cinco posibilidades en las que un ORF puede ser leído, ejemplificado con las diferentes fases 0, 1 y 2, tanto en la hebra sentido como antisentido. (Modificado de Sabath, 2008).

2.2. **Los genes sobrelapados en el mundo viral**

Una característica particular observada en los virus es la abundancia de marcos abiertos de lectura (ORF, por sus siglas en inglés) sobrelapados, los cuales se presentan a una mayor escala en comparación con los tres dominios de la vida. Los genes sobrelapados se originan por varios mecanismos, principalmente por el uso de codones de inicio alternativos, lectura ribosomal y cambio de marco ribosomal (Brandes & Linial, 2016). Se reporta que los eventos de sobrelapamiento son mayores en virus de RNA y en virus con genomas pequeños.

Por otra parte, se han sugerido diversas explicaciones de esta gran abundancia de genes sobrelapados, una teoría argumenta que los eventos de sobrelape surgieron como un mecanismo restricción selectiva ante la alta tasa de mutación, principalmente en los virus de RNA (Krakauer, 2000), otra teoría establece que el sobrelapamiento es un mecanismo efectivo para la generación de

nuevos genes al introducir nuevos marcos de lectura en uno ya existente, una tercera teoría y la más aceptada menciona que es la compresión genómica la principal fuerza evolutiva generadora de sobrelapes de genes, siendo esta compresión generada por la restricción física que ocasiona por la capside, especialmente en los virus icosaédricos (Chirico et al. 2010; Brandes & Linial, 2016).

Un modelo interesante para el estudio de los genes sobrelapantes desde el punto de vista evolutivo, es el el virus de la Hepatitis B (HBV), esto debido a que alrededor del 50% de su genoma (Zaaijer et al. 2007) presenta marcos abiertos de lectura (ORF) sobrelapados y el marco de lectura del gen que codifica para la proteína de superficie se encuentra totalmente embebido en el marco de lectura de gen que codifica para la polimerasa. Dado que el sobrelapamiento de genes cubre una gran parte del genoma y que una sustitución nucleotídica sinónima en un ORF puede dar lugar a una sustitución nucleotídica no sinónima en el ORF sobrelapado, se piensa que la evolución del genoma del HBV está limitada para mantener las funciones esenciales de sus proteínas (Cento et al. 2012).

.

2.3. **Composición nucleotídica en regiones sobrelapadas**

Los genes sobrelapados presentan una composición nucleotídica significativamente diferente de las regiones no sobrelapadas. En los genes sobrelapados existe una sobre representación del nucleótido C, los dinucleótidos CG, TC y CC; de los aminoácidos Arg, Ser y Gln, así como de los codones sinónimos CGA (Arg), TCG (Ser) y AGC (Ser). Estos patrones de composición están correlacionados, Arg es codificado por codones ricos en CG, por ejemplo CGA, el cual también es abundante; asimismo, Ser es codificado en gran parte por TCG, que es una combinación de los dinucleótidos en mayor abundancia TC y CG. Pavesi (2020), analizando 20 características de composición en genes sobrelapadas virales, encontró que existe una subrepresentación de T, los dinucleótidos TA, AT, TT y TG, los aminoácidos Tyr, Val, Phe e Ile, los codones sinónimos TAT (Tyr), TTT (Phe) , ATT (Ile) y GTT (Val) y de aminoácidos con baja degeneración de codones. De igual manera, estos patrones se correlacionan entre sí, los aminoácidos subrepresentados, Tyr, Phe e Ile están codificados por codones ricos en A y T (entre los cuales TAT, TTT y ATT están subrepresentados). La

disminución del contenido de dinucleótidos TA y TG tiene un claro significado biológico, ya que reduce la probabilidad de aparición de codones de terminación (TGA, TAG y TAA) en regioens sobrelapados (Pavesi et al., 2018).

Las proteínas sobrelapadas tienen una composición de secuencia generalmente sesgada hacia los aminoácidos promotores de desorden y se predice que contienen significativamente más desorden estructural que las proteínas no sobrelapadas, además, gran parte son creadas de *novo*. La mayoría de las proteínas creadas de *novo* se han descrito como proteínas huérfanas (es decir, restringidas a una especie o género), casi todas son proteínas accesorias que juegan un papel en la patogenicidad o diseminación viral, en lugar de proteínas centrales para la replicación o estructura viral. Se predice que la mayoría de las proteínas creadas de *novo* presentan estructuras intrínsecamente desordenadas y tienen una composición de secuencia altamente inusual (Rancurel et al 2009).

2.4. **Variabilidad genética en el HBV**

En el caso particular del HBV, una característica de la infección de este virus es la notable heterogeneidad genética a nivel inter e intrapaciente. El último caso de variabilidad como una población de genomas estrechamente relacionados, pero no idénticos se conoce como cuasiespecies virales. Es bien reconocido que tales mutaciones pueden tener implicaciones importantes con respecto a la patogénesis de la enfermedad viral. Por ejemplo, en la infección crónica, la mutación del punto G a A en el nucleótido (nt) 1896 en la región Precore (preC), así como las mutaciones A1762T y G1764A en la región promotora están altamente asociadas con la seroconversión de HBeAg, en este estado los pacientes presentan bajos niveles de viremia y la consiguiente curación clínica. En contraste, la infección aguda con el mutante pr-C G1896A representa un alto riesgo de insuficiencia hepática fulminante. Aunque estos hechos ilustran claramente las implicaciones clínicas de ciertas mutaciones virales, la creciente evidencia sugiere fuertemente que la heterogeneidad genética viral es más complicada de lo que se pensaba (Revill et al., 2020) .

La mayoría de los cambios genómicos observados en las variantes virales del HBV son sustituciones de bases únicas y se distribuyen ampliamente por todo el genoma. En un estudio cuatro de ocho (50%) pacientes con HBeAg negativo sin

terapia crónica mostraron una prevalencia relativamente baja del mutante Precore (preC) G1896A en los tejidos del hígado, lo que sugiere que otras mutaciones estaban involucradas en su seroconversión de HBeAg. Curiosamente, los tejidos hepáticos en 4 de 5 (80%) de los casos anti-HBe positivos tratados con análogos de nucleósidos (NA) presentaron frecuencias extremadamente bajas de la mutanción preC G1896A (0.0%, 0.0%, 0.1% y 1.1%), sugiriendo la alta sensibilidad de esta mutante a los NA. Además, gran abundancia de clones resistentes a NA fueron comunes tanto en el hígado como en el suero de pacientes sin tratamiento previo, y la proporción de mutantes M204VI resistentes a lamivudina y entecavir se expandió en respuesta al tratamiento con entecavir en el suero del 35,7% (5/14) de pacientes, lo que sugiere riesgo de desarrollar resistencia a los medicamentos de NAs. Por otra parte, se encontró menos diversidad utilizando Entropia de Shannon en genomas de HBV en pacientes crónicos sin tratamiento con NAs (Nishijima et al 2012)

2.5. **Evolución molecular de los genes sobrelapados**

Los genes sobrelapados están particularmente limitados evolutivamente porque una mutación en una región superpuesta afecta simultáneamente a los dos (u ocasionalmente más) genes involucrados en ese sobrelapamiento. Analizando las mutaciones en función de la posición del codón en la que caen, alrededor del 70% de las mutaciones que ocurren en la tercera posición del codón son sinónimas, frente a solo 5 y 0% de las mutaciones en la primera y segunda posición del codón, respectivamente (Sabath et al. 2008ª). Esto ocasiona que una mutación sinónima en un marco de lectura es muy probable que sea no sinónima en el otro marco; esto permite que haya una compensación genética, pero también cierta restricción evolutiva.. Se ha demostrado que existe una mayor restricción al cambio evolutivo en las regiones sobrelapadas, por ende, presentan una menor diversidad genética que las regiones no sobrelapadas (Mizokami et al. 1997; Sabath et al. 2008b, 2012; Simon-Loriere et al. 2013).

Además de la restricción, los sitios sobrelapados codifican residuos funcionalmente importantes para un gen u otro, pero nunca para ambos. Un método para determinar la restricción evolutiva en un gen sobrelapado es medir la tasa de sustitución de nucleótidos (número de variaciones por sitio), específicamente, la tasa de sustituciones sinónimas (dS, el número de sustituciones sinónimos por sitio

nucleotídico sinónimo) y la tasa de sustituciones de no sinónimas (dN, el número de sustituciones no sinónimas por sitio nucleotífico no sinónimo. dS es independiente de las presiones selectivas externas y suele ser mayor que dN. dN puede estar sujeto a presiones selectivas externas, es decir, la conservación estructural requerida de la proteína codificada y su restricción conformacional. En marcos de lectura sobrelapados, un cambio sinónimo en un gen puede no ser neutral en el otro. Por lo tanto, la dS de genes sobrelapadas podría verse limitado debido a los cambios no sinónimos que se deben mantener en la población en un marco de lectura que este bajo selección natural positiva. Para el producto proteico de un gen, el grado de restricción evolutiva y selección sobre él puede indicarse mediante la relación dN/dS < 1 para la selección negativa o purificadora dN/dS > 1 para la selección positiva. Bajo selección negativa, la diversidad genética disminuye a medida que un rasgo particular (frecuencia de fenotipo) se estabiliza en la población. En comparación, la selección positiva aumenta la frecuencia de ciertas variaciones y ocurre cuando no se alcanza un equilibrio en la población. (Fernandes et al.2016).

Por ejemplo, el virus de la leucemia bovina contiene una región pXBL que codifica tres partes de cuatro proteínas reguladoras (Tax, Rex, G4, R3) en marcos de lectura sobrelapados. Las tasas de sustituciones sinónimas fueron consistentemente más bajas en regiones sobrelapadas que en regiones no sobrelapadas o con poco sobrelape, el sesgo de codones fue también más alto. A un mayor nivel de superposición en las regiones codificadoras se correlaciona con una mayor restricción evolutiva. Tax, la más conservada entre las cuatro proteínas reguladoras, mostró una selección purificadora consistente con su importancia en el ciclo de vida viral (Zhao et al 2007).

El desarrollo de modelos evolutivos específicos para analizar genes sobrelapados ha sido un reto, sin embargo, diversos autores han realizado esfuerzos para el análisis de este tipo de genes (Hein and Stovlbaek, 1995; Krakauer, 2000). El enfoque estándar para detectar y analizar la selección natural actuando sobre regiones codificantes es mediante la estimación del cociente entre la tasa de sustituciones nucleotídicas no sinónimas (dN) y las tasas de sustituciones nucleotídicas sinónimas (dS). No obstante, como se ha mencionado, en los genes sobrelapados una mutación puede ser no sinónima en un marco de lectura y sinónima en el otro, por lo que los métodos clásicos no son útiles. Métodos

particulares para los genes sobrelapados se han desarrollado a lo largo del tiempo, por ejemplo Sabath et al. (2008) desarrollaron un método de máxima verosimilitud para estimar simultáneamente la intensidad de la selección en cada uno de los genes sobrelapados, sin embargo, el método no puede evaluar si el cociente dN/dS difiere significativamente de 1 para cualquier gen, limitante que Wei & Zhang (2014) resuelven separando los efectos de cada mutación sobre los dos genes y estimando la varianza asociada permitiendo evaluar la neutralidad de cada gen.

2.6. **Evolución molecular de los genes sobrelapados en el HBV**

Mizokami et al. (1997) demostraron que las tasas de sustitución sinónima en regiones sobrelapadas eran menores que en las regiones no sobrelapadas en el genoma del HBV, proporcionando evidencia de las restricciones evolutivas asociadas con las regiones sobrelapadas. Zaaijer et. al (2007) propusieron que el HBV puede, mediante la degeneración del código genético, sobrepasar estas restricciones. Hablando particularmente de la región sobrelapada entre el ORF P y el ORF S, debido al cambio de marco de lectura en esta region , la primera posición en el codón del ORF P corresponde a la tercera posición en el codón del ORF S (P1S3), la segunda posición en el codón P corresponde a la primera posición en el codón S (P2S1), y la tercera posición en el codón P corresponde a la segunda posición en el codón S (P3S2). Por lo tanto, una mutación sinónima en P puede producir una mutación no sinónima correspondiente en S (P3S2), que puede producir un cambio de aminoácido en S pero conserva el sitio correspondiente en P. Con lo anterior, se satisface las restricciones de cambio en ambos genes.

Zhang et al. (2010) proponen que las restricciones funcionales de los dominios de las proteínas del virus son las principales responsables de los diferentes patrones de selección exhibidos por la distribución de mutaciones y cambios de aminoácidos en las regiones donde se encuentran los genes sobrelapados. Por ejemplo, Chen et al. (2013) encontraron que la región que codifica para el dominio espaciador (Spacer) de la polimerasa, es una región en donde se pueden acumular mutaciones que proporcionan flexibilidad para que puedan ocurrir cambios conformacionales en el dominio proteico codificado y responder a la presión selectiva del sistema inmune.

Por otra parte, Pavesi (2015) encontró que la región sobrelapada P/S en HBV presenta dos patrones de uso de codones. Uno se localizó en el primer tercio del extremo 5’ del sobrelape y el otro en las dos terceras partes del extremo 3’, estos dos patrones se presentan en todos los hepadnavirus y sugiere que el dominio espaciador de la polimerasa y el dominio S de la proteína de superficie se originaron de *novo* por *overprinting*. Así pues, la medición y análisis de los patrones del uso de codones nos da información relevante para conocer procesos evolutivos en los genes sobrelapados así como de su origen (Kozlov, 1999; Paversi, 2006; Pavesi, 2013; Pavesi, 2015)

3. JUSTIFICACIÓN

Los trabajos previos se han enfocado en el análisis evolutivo de la región sobrelapada P/S del HBV. El presente trabajo pretende enfocarse en la caracterización de genes sobrelapados y no sobrelapados en el genoma completo del virus a través de la determinación de su composición nucleotídica, del uso y sesgo de codones, variabilidad genética y de los patrones de selección natural. De esta manera, se contribuirá al entendimiento de la capacidad adaptativa de algunas variantes del HBV ante las diferentes presiones inmunológicas y farmacológicas.

4. HIPÓTESIS

En las regiones sobrelapadas existe un mayor uso de codones con altos niveles de degeneración, así como un mayor número de mutaciones en sitios sinónimos que limitan el cambio evolutivo en comparación con las regiones no sobrelapadas del HBV.

5. OBJETIVOS

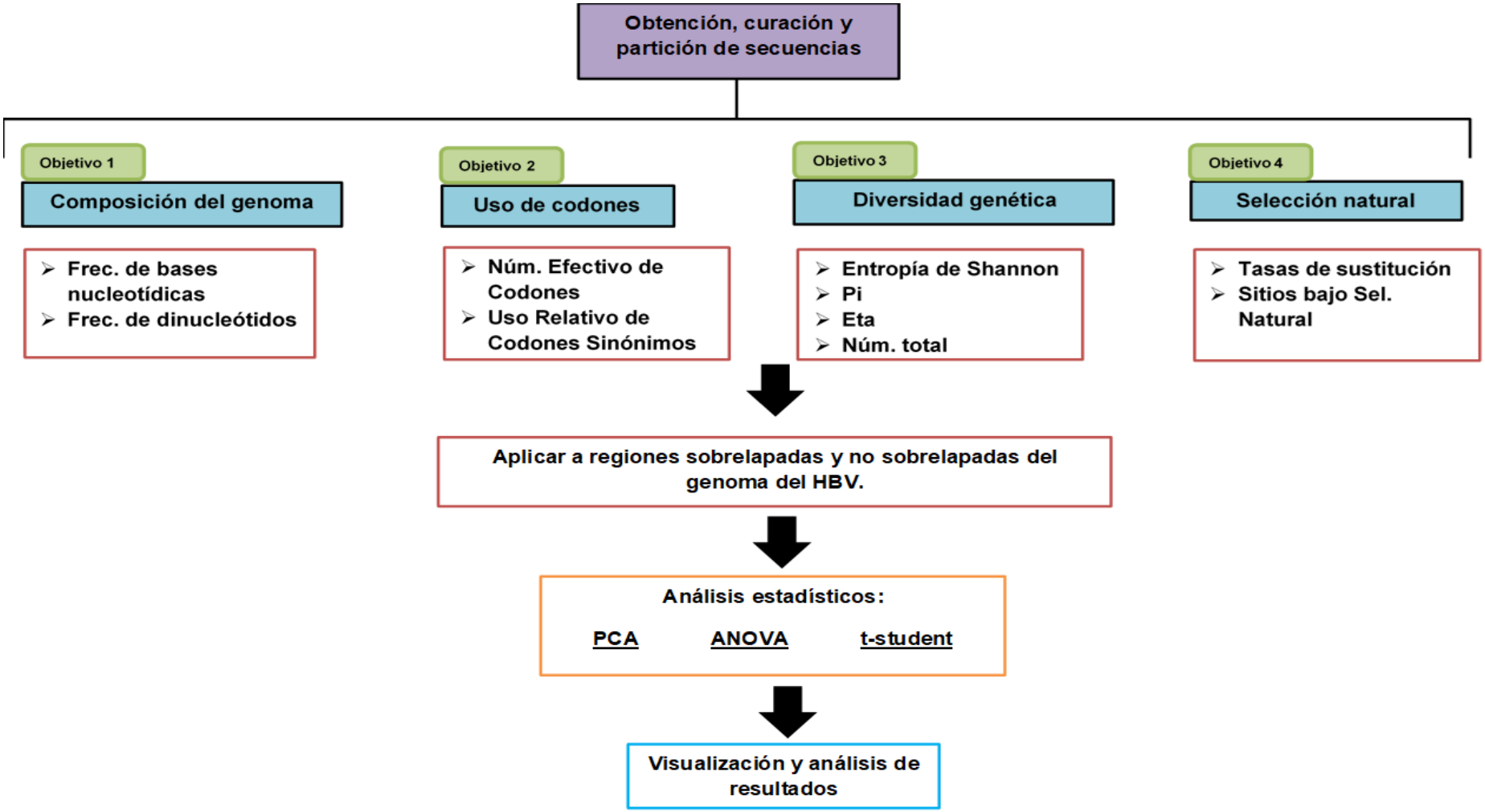
5.1. **Objetivo general**

Analizar la evolución molecular de las regiones sobrelapadas y no sobrelapadas presentes en el genoma del Virus de la Hepatitis B.

5.2. **Objetivos particulares**

- Determinar la composición nucleotídica de las regiones sobrelapadas y no sobrelapadas en el genoma del HBV.
- Determinar la diferencia en uso de codones a lo largo del genoma del HBV
- Determinar la variabilidad genética en las regiones sobrelapadas y no sobrelapadas del genoma del HBV.
- Identificar y analizar señales de selección natural en las diferentes regiones del genoma del HBV

6. ESTRATEGIA EXPERIMENTAL



7. MATERIALES Y MÉTODOS

7.1. **Obtención, curación y partición de secuencias**

En agosto del 2019 se obtuvieron las secuencias genómicas de los 8 principales genotipos del virus de la hepatitis B (A-H) de la base de datos especializada "The Hepatitis B Virus database, *HBVdb*" (https://hbvdb.lyon.inserm.fr/HBVdb/). La base de datos HBVdb contiene una colección de secuencias del virus de la hepatitis B anotadas, desarrollada por el equipo del Dr Fabien Zoulim en Francia. HBVdb tiene como objetivo (1) anotar las secuencias del HBV, (2) proporcionar información sobre el genotipo y sobre las proteínas codificadas por estas secuencias, (3) perfilar la resistencia al fármaco y (4) brindar acceso amigable a herramientas bioinformáticas para extraer y analizar estas secuencias del HBV (Hayer et al., 2013).

Los genomas del HBV de cada genotipo se alinearon utilizando el programa *SeaView* (Gouy et al, 2010). La curación consistió en la identificación y eliminación de secuencias genómicas duplicadas y con un porcentaje de identidad del 99% en cada genotipo a través de *JalView* (Waterhouse et al. 2009). Además, se eliminaron los genomas que presentaron variación en su tamaño con base en los tamaños reportados por los genomas de referencia de cada genotipo reportados por la herramienta de genotipado del NCBI (https://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi). Los genotipos de referencia se resumen en la Tabla 1:

Tabla 1 Genotipos de referencia utilizados en este proyecto

Genotipo	Acceso	Nombre	Tamaño (Nt)
A	X02763	Hepatitis b virus genome (serotype adw2)	3221
B	D00329	Hepatitis B virus subtype ADW DNA, complete genome, clone:pJDW233.	3215
C	X04615	Hepatitis B virus genome, subtype ayr	3215
D	X65259	Hepatitis B virus (ayw,patient E) genes PreS1, PreS2, PreC, C, X and polymerase	3182

E	X75657	Human hepatitis virus (genotype E, Bas) preS1, preS2, S, C, X, antigens, core antigen, X protein and polymerase	3212
F	X69798	Hepatitis B virus, subtype adw4 genes	3215
G	AF160501	Hepatitis B virus strain IG29227, complete genome	3248
H	AY090454	Hepatitis B virus strain 1853Nic, complete genome	3215

Hecha la curación se obtuvo al menos 100 secuencias genómicas de cada genotipo para los análisis posteriores (Tabla 2).

Tabla 2. Secuencias obtenidas posterior a la curación

Genotipo	Número de secuencias	Fecha de obtención de secuencias	Tamaño de genoma (nt)
A	140	Agosto 2019	3221
B	143	Agosto 2019	3215
C	135	Agosto 2019	3215
D	110	Agosto 2019	3182
E	108	Agosto 2019	3212
F	104	Agosto 2019	3215
G	32	Agosto 2019	3248
H	45	Agosto 2019	3215

Finalmente, utilizando la herramienta *Graphics* de cada genotipo de referencia en el NCBI, se identificaron y se delimitaron las diferentes regiones sobrelapadas y no sobrelapadas de los genomas. Con ayuda de los programas *Aliview* (Larsson, 2014) y *MEGAX* (S. Kumar et al., 2018) se delimitaron las diferentes regiones y se generaron 10 conjuntos de datos correspondientes la región sobrelapada y no sobrelapada de los ORFs de los ocho genotipos del HBV: C, P, S y X (Figura 8).

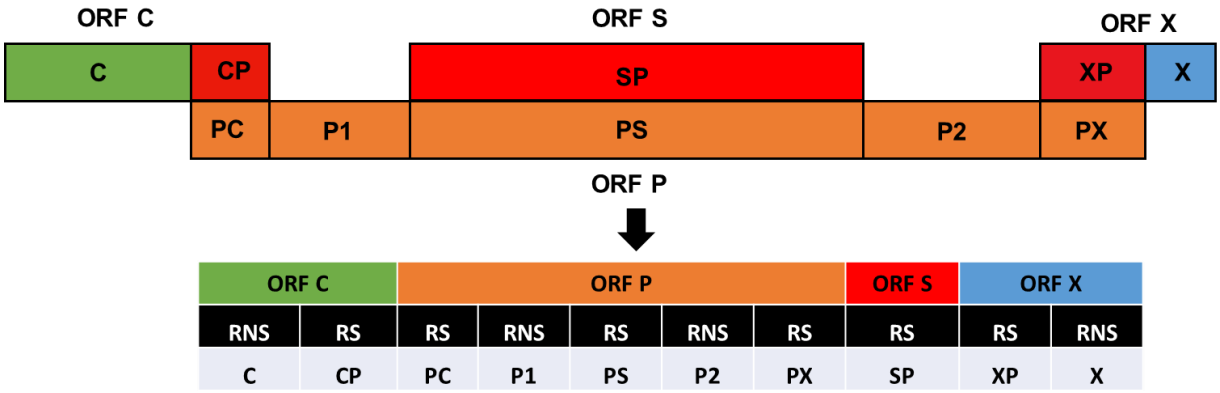


Figura 8. Esquema representado las diferentes regiones en las que fue dividido el genoma del HBV. Se dividió en 10 regiones correspondientes a todas las regiones sobrelapadas y no sobrelapadas del genoma del HBV; C y CP, región del ORF C no sobrelapada y sobrelapada con el ORF , respectivamente; PC, PS y PX, región de P sobrelapada con el ORF C, el ORF S y el ORF X, respectivamente; P1 y P2, regiones no sobrelapadas del ORF P; SP, corresponde al ORF S, se sobrelapada en su totalidad con el ORF S; XP y X, región del ORF X sobrelapada con el ORF P y no sobrelapada, respectivamente.

7.2. Caracterización de la composición nucleotídica de las regiones sobrelapadas y no sobrelapadas en el genoma del HBV

Para caracterizar la composición nucleotídica del genoma del HBV se utilizaron los 10 conjuntos de secuencias pertenecientes a todas las regiones del genoma y se calcularon diferentes elementos de composición: en primer lugar el contenido de Adenina (A), Timinia (T), Citocina (C) y Guanina (G) en general, seguido del contenido de las bases en la tercera posición de los codones (A3, T3, C3 y G3), el contenido GC en la primera, segunda y tercera posición de los codones (GC1, GC2 y GC3), el contenido de GC y AT en general y finalmente el contenido de los doce dinucleótidos (AA, AC, AG, AT, CA, CC,CG,CT,GA, GC, GG, GT, TA, TC,TG y TT). Para el cálculo se crearon scripts en *R* utilizando como base el paquete especializado para análisis de codones "*kodonz*" (Zhou & Kong, 2019) y "*R453Plus1Toolbox*" para los dinucleótidos. El contenido se calculó como frecuencias relativas de cada una de las variables de composición mencionadas anteriormente.

Con los datos obtenidos se generaron dos matrices de datos multivariados, una correspondiente a las frecuencias de las variables de composición de bases versus las diferentes regiones genómicas del HBV, mientras que la otra matriz corresponde a las frecuencias de los 16 dinucleótidos versus las diferentes

regiones. Posteriormente con el objetivo de identificar asociaciones entre las variables de composición y las regiones del genoma se realizaron Análisis de Componentes Principales (PCA) y de correlación de Pearson utilizando cada una de las matrices de datos.

En el caso particular de los datos de composición nucleotídica, posterior a identificar las variables que explicaban mejor la variación, se realizaron comparaciones de estas variables entre ORFs y sus regiones sobrelapadas y no sobrelapadas. La identificación de diferencias significativas se realizó mediante Análisis de Varianza (ANOVA) en *R.*

7.3. **Determinación del sesgo en el uso de codones**

Para determinar el sesgo en el uso de codones en las regiones del genoma del HBV se calcularon dos métricas. El primer parámetro se refiere al uso relativo de codones sinónimos (RSCU) que indica la frecuencia relativa de cada uno de los 59 codones sinónimos presentes en una secuencia determinada. Cuando el valor de RSCU es >1 indica sesgo de uso de codón y cuando es <= 1 no existe sesgo. Los cálculos del RSCU de los 59 codones sinónimos en las diferentes regiones genómicas del HBV se realizaron en el programa *MEGAX* (S. Kumar et al., 2018), utilizando los 10 conjuntos de secuencias obtenidas previamente.

El segundo parámetro utilizado es el número efectivo de codones (ENC), el cual indica el uso absoluto de codones en una secuencia. Los valores de ENC van de 20, indicando sesgo extremo en donde las secuencias utilizan sólo un codón para codificar cada aminoácido, a 61 en donde no existe sesgo utilizando todos los 61 codones disponibles. Estos parámetros se calculan en todas las regiones utilizando el paquete "*vhcub*" (Mostafa & Soudy, 2019) en *R*. Además, con estos datos se realizaron gráficas de ENC vs GC3 de la región sobrelapada y no sobrelapada del ORF P, las cuales nos indican de manera indirecta posibles mecanismos evolutivos que pueden estar actuando para determinar el sesgo de codones, la presión mutacional o la selección natural, principalmente.

7.4. **Variación genética en las regiones sobrelapadas y no sobrelapadas**

Para determinar la variabilidad genética, se generaron dos conjuntos de secuencias por cada genotipo (A-H), uno correspondiente a la región sobrelapada

y el otro a la región no sobrelapada de cada genoma, generando un total de 8 pares de archivos en formato fasta. A cada uno se le aplicaron los análisis que se explican a continuación:

1) La entropía de Shannon se calculó mediante el programa *Entropy* de la base de datos de HIV de los Alamos (http://www.hiv.lanl.gov/). La entropía de Shannon cuantifica la variación en cada posición de un alineamiento de DNA: si la entropía es baja, existe poca variación nucleotídica; de lo contrario, es e alta. Esta medida de diversidad también se aplicó en los genomas completos de todos los genotipos. 2) El índice de Pi (π), el cual mide el número promedio de diferencias nucleotídicas por sitio entre secuencias; 3) el número total de mutaciones (Eta); y 4) el número de sitios polimórficos, las últimas tres métricas se calcularon en el programa *DnaSP* (Rozas et al. 2017).

Para identificar visualmente los sitios en donde existía una mayor variabilidad por sitio nucleotídico en cada codón y en determinada región genómica , los valores de entropía de Shannon de todos los sitios del genoma completo de los genotipos B, C, E y F del HBV se graficaron utilizando el programa *ggplot* en R, la particularidad realizada fue graficar con diferentes colores las posiciones de las bases en cada codón, así, todas las primeras posiciones corresponden a color verde, las segundas posiciones en azul y las terceras a las rojas. Se eligieron esos cuatro genotipos porque tienen distribución geográfica diferencial, así como el mismo tamaño genómico (3215 nt), condición importante para este análisis.

7.5. **Inferencia de sitios bajo selección natural**

La inferencia de sitios bajo selección natural se realizó por dos métodos, el primero fue utilizando el algoritmo *Fixed Effects Likelihood* (FEL) (Kosakovsky et al, 2005) en la plataforma *Datamonkey* (Weaver, 2018); éste método clásico al igual que muchos otros realizan la inferencia de la selección a partir de secuencias codificantes de proteínas se basan en la siguiente observación: una sustitución no sinónima (o de reemplazo) cambia la secuencia primaria de la proteína codificada y es más probable que influya en la capacidad replicativa y de sobrevivencia (adecuación biológica) de un organismo que una sustitución sinónima aleatoria que no genera cambios en la secuencia de aminoácidos (Poon et al., 2009).

Si las mutaciones no sinónimas en un sitio de codón particular en la secuencia tienen un efecto insignificante en la función o expresión de la proteína (y por lo tanto en su adecuación biológica), entonces la tasa de sustituciones no sinónimas (dN) debería ser similar a la tasa de sustituciones sinónimas (dS), y el sitio evoluciona de forma neutral (Figura 9). Un exceso de sustituciones no sinónimas (dN> dS) puede interpretarse que ese sitio está bajo selección positiva, lo que sugiere que los cambios de aminoácidos aumentan la adecuación biológica. Una escasez de cambios de aminoácidos (dN <dS) indica que la selección negativa está funcionando para eliminar tales sustituciones del acervo genético (Poon et al., 2009).

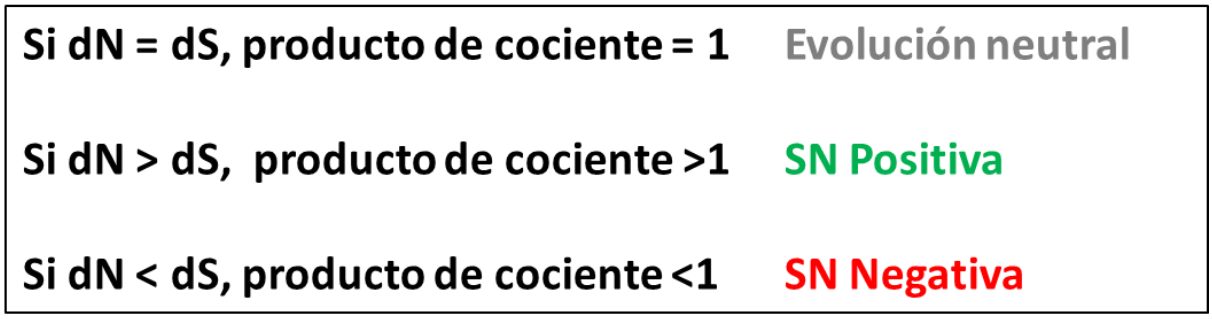


Figura 9. Posibles resultados del cociente dN/dS y su interpretación evolutiva. Cuando dN = dS, el producto es igual a 1, se infiere que el sitio nucleotídico está bajo evolución neutral; cuando dN > dS, el producto del cociente es > 1, se infiere que el sitio nucleotídico está bajo selección natural (SN) positiva; cuando dN < dS, el producto del cociente es < 1, se infiere que el sitio nucleotídico está bajo selección natural (SN) positiva.

No obstante, este tipo de métodos no están diseñados para inferir sitios bajo selección natural en regiones sobrelapadas, debido a que en este tipo de genes una posición puede ser simultáneamente sinónima para el ORF 1 pero no sinónima para el ORF 2. Por ello, se utilizó un segundo método implementado por el programa *OLGenie* (Nelson et al. 2019), considera los cambios sinónimos y no sinónimos observados en los dos ORFs sobrelapados.

OLGenie estima la tasa de sustitución no sinónima (dN) y sinónima (dS) observados en genes sobrelapados modificando el método de Wei y Zhang (2015). Se utilizan cuatro medidas ampliadas de dN y dS: dNN, dSN, dNS y dSS, donde el primer subíndice se refiere al gen de referencia y el segundo subíndice se refiere al gen alternativo (NN, no sinónimo / no sinónimo; SN, sinónimo / no sinónimo; NS, no sinónimo / sinónimo; SS, sinónimo / sinónimo). Por ejemplo, dNS se refiere al

número medio de sustituciones de nucleótidos por sitio que no son anónimos en el gen de referencia, pero son sinónimos en el gen alternativo (NS). Dados estos valores, dN / dS puede estimarse para el gen de referencia como dNN / dSN o dNS / dSS, o para el gen alternativo como dNN / dNS o dSN / dSS.

En cada caso, el efecto de las mutaciones en uno de los dos genes sobrelapados se mantiene constante (N o S), lo que garantiza una “comparación equitativa” en el otro gen. Por ejemplo, si los cambios no sinónimos observados en el gen de referencia son sinónimos desproporcionados en el gen alternativo (dNS> dNN), el resultado será dNN / dNS <1.0, y se puede inferir la selección purificadora en el gen alternativo (Hughes y Hughes 2005). En la práctica, dNN / dNS en lugar de dSN / dSS se usa típicamente para evaluar la selección en el gen alternativo, ya que los sitios SS generalmente son demasiado raros para permitir una estimación confiable de dSS.

La ventaja de FEL es su poder estadístico para la inferencia de sitios bajo selección, su desventaja es que fue hecho para ORFs no sobrelapados, respecto a *OLGenie,* su ventaja es que en un contexto de sobrelapamiento de ORFs, considera los cambios mutacionales de los dos marcos de lectura, su principal desventaja es que considera sólo los cambios observados en los dos ORF sobrelapados y no realiza filogenias para el mapeo de los cambios. Es por ello que se utilizaron ambos métodos para inferir, identificar y mapear sitios bajo selección natural tanto en regiones sobrelapadas como no sobrelapadas. Se utilizó el conjunto de secuencias genómicas del genotipo B (n=132) debido a que es de los genotipos de los cuales existe mayor estudios clínicos y biológicos en la literatura que fueron utilizados para realizar el análisis de resultados.

7.6. **Análisis estadísticos**

La normalidad de los datos se corroboró con el test de Shapiro-Wilk, posteriormente para identificar diferencias estadísticamente significativas en la composición, sesgo en el uso de codones y diversidad entre regiones sobrelapadas y no sobrelapdas en el genoma del HBV se realizaron pruebas de ANOVA o t-student dependiendo de las características de los datos.

8. RESULTADOS

8.1. **Composición nucleotídica del genoma del HBV**

Las bases nucleotídicas predominantes en el genoma del HBV son la citosina y la timina, la frecuencia relativa de estas en porcentaje fue de 27% y 28% respectivamente (Figura 10A). Al contrastar la composición de bases entre las regiones sobrelapadas y no sobrelapadas del genoma, existe una predominancia de adenina y timina en las regiones no sobrelapadas y de citocina en las sobrelapadas (Figura 10B). El contenido de GC en el genoma del HBV corresponde a un 48.8%, de este total el 55.5% se encuentra distribuido en las regiones sobrelapadas del genoma, mientras que el 45.3% en las no sobrelapadas, además existe una fluctuación del contenido de GC entre genotipos del HBV, existiendo una desviación estándar de ± 0.37.

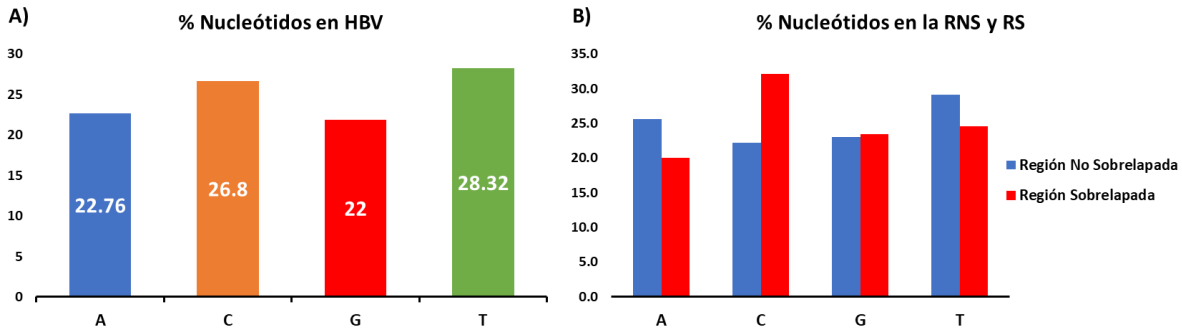


Figura 10 Composición de nucleótidos del genoma del HBV. A) Contenido general en porcentaje de las bases nucleotídicas, B) Contenido en porcentaje de las bases nucleotídicas contrastando las regiones sobrelapadas (en rojo) y no sobrelapadas (en azul) del genoma del HBV.

8.2. **Composición nucleotídica entre regiones sobrelapadas y no sobrelapadas del HBV**

Para analizar los patrones de variación de la composición del genoma sobre las regiones sobrelapadas y no sobrelapadas se realizó un Análisis de Componentes Principales (PCA) utilizando como datos el contenido (%) de A, T, G, C, GC, AT GC1, GC2, GC3, A3 y G3 de las diferentes regiones del genoma del HBV. Como se observa en la Figura 11, las variables relacionadas con el contenido de GC se encuentran asociadas a las regiones sobrelapadas (círculos negros), con excepción de la región sobrelapada del ORF C. Por otra parte, las regiones no

sobrelapadas se encuentran principalmente asociadas con las variables relacionadas con el contenido de AT (en cuadro azul).

Los resultados de correlación indican que el contenido de GC (r= 0.9623201, $p < 0.01$) GC3 (r=0.9081760 $p < 0.01$), GC3 (r= 0.8693677 $p < 0.01$), C (0.8325277 $p < 0.01$) y G (r= 0.8312320 $p < 0.01$) son los principales elementos de composición que están asociados con la variación de nucleótidos entre las diferentes regiones del genoma del HBV.

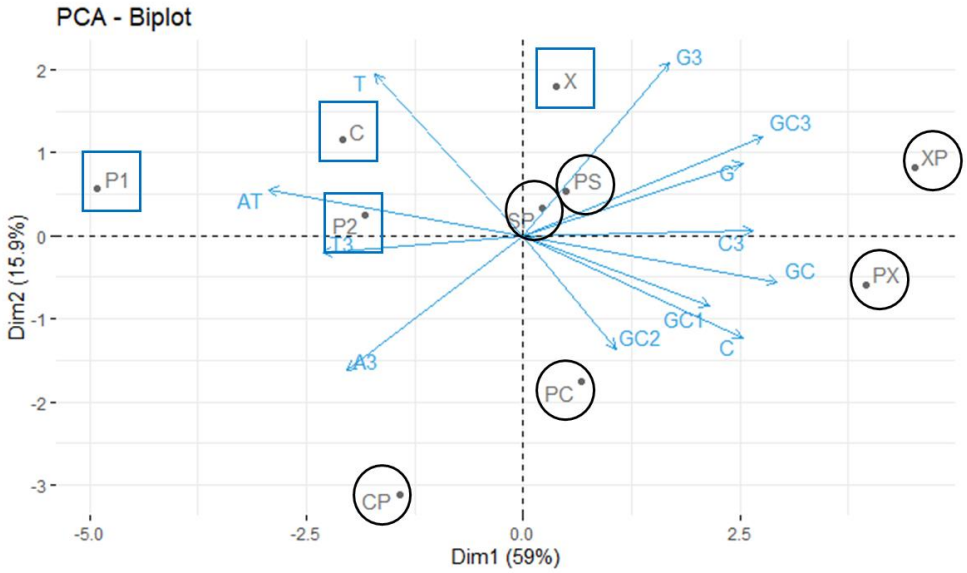


Figura 11. Gráfica bidimensional del Análisis de Componentes Principales utilizando los datos de composición de nucleótidos. Las regiones sobrelapadas (círculos negros) se asocian principalmente a las variables relacionadas al contenido de GC, mientras que las no sobrelapadas (cuadros azules) a las variables de contenido de AT.

8.3. Composición dinucleotídica entre regiones sobrelapadas y no sobrelapadas del HBV

De igual manera que con la composición nucleotídica, se realizó un PCA con los datos de frecuencia relativa de cada dinucleótido (Figura 12), los dinucleótidos que presentan contenido de GC se encuentran asociadas principalmente a las regiones sobrelapadas (cuadros rojos). Por otra parte, las regiones no sobrelapadas se encuentran principalmente asociadas con dinucleótidos que

presentan contenido de AT (cuadros azules), demostrando así que existe una diferencia en composición dinucleotídica entre regiones del genoma del HBV.

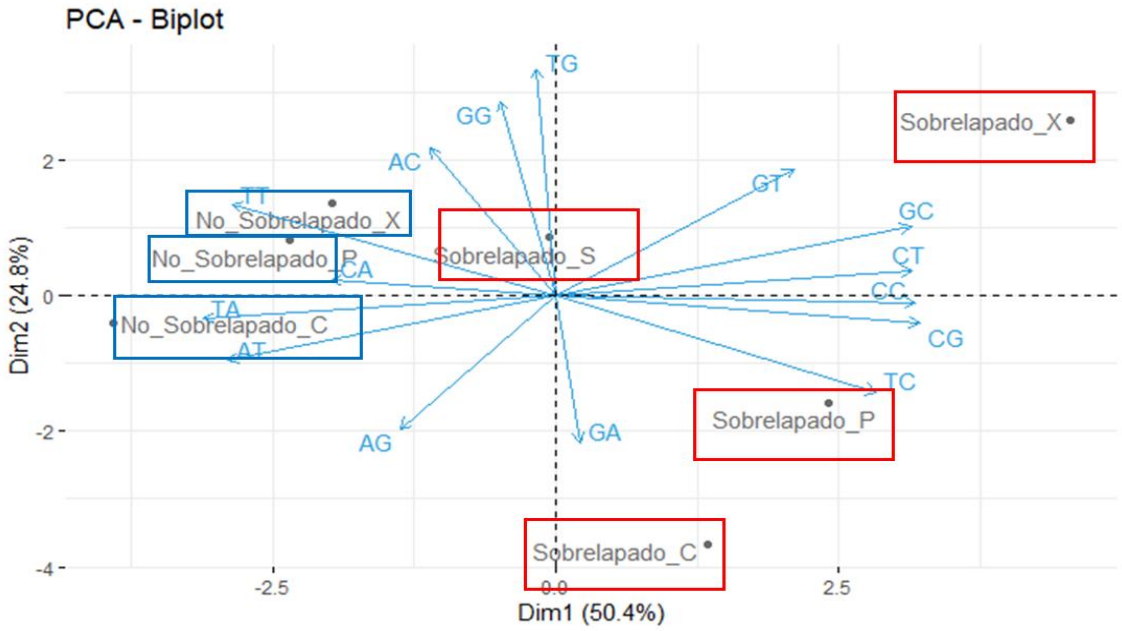


Figura 12. Gráfica bidimensional del Análisis de Componentes Principales utilizando los datos de composición de dinucleótidos. Las regiones sobrelapadas (cuadros rojos) se asocian principalmente a las variables relacionadas al contenido de GC, mientras que las no sobrelapadas (cuadros azules) a las variables de contenido de AT.

Para saber de manera más específica qué dinucleótidos se encontraban en mayor o en menor cantidad en las dos regiones principales del genoma que se analizan, se realizaron histogramas de las frecuencias relativas observables de cada dinucleótido (Figura 13). Se distinguió que en las regiones no sobrelapadas los dinucleótidos principalmente representados TpT, TpA, ApA. Por otro lado, en las regiones sobrelapadas los dinucleótidos GpC, CpC, TpG, fueron los que se encontraron en mayor cantidad. Estos resultados son congruentes tanto con los resultados de composición nucleotídica general como de dinucleótidos realizados en este trabajo.

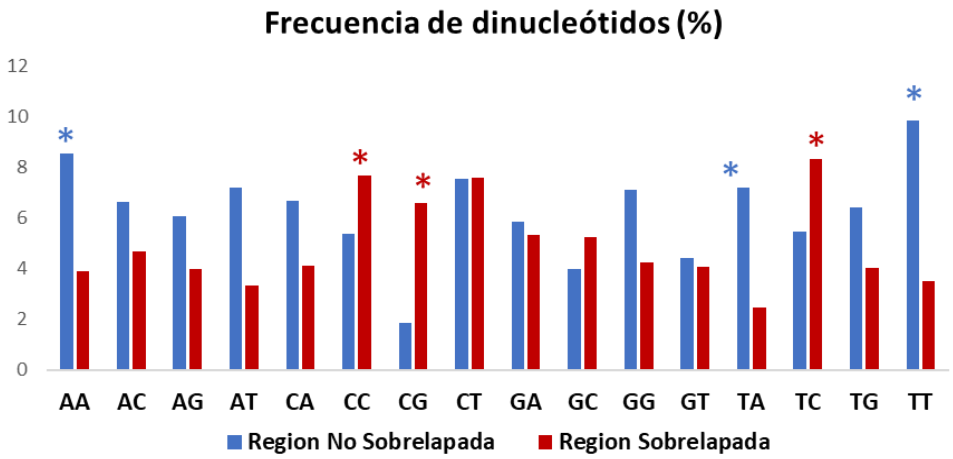


Figura 13. Frecuencia de dinucleótidos en regiones sobrelapadas y no sobrelapadas del genoma del HBV. En asterisco se indican los nucleótidos principalmente representados en las regiones no sobrelapadas (azul) y no sobrelapadas (rojo).

Ahora bien, una de las métricas más utilizadas para analizar la composición dinucleotídica es el *Odd Ratio,* el cual es un cociente de la frecuencia relativa observable entre la esperada de cada nucleótido, si el valor >1 ese dinucleótido se encuentra por arriba de lo esperado (sobrerrepresentado) y si es < 1 se encuentra por debajo de lo esperado (subrepresentado).

En las regiones no sobrelapadas, los dinucleótidos subrepresentados son CpC y GpG, mientras que en las sobrelapadas los sobrerrepresentados son TC y GA principalmente. LDentro de los dinucleótidos que se encuentran subrepresentados, se encontró que CpG presenta el principal contraste, en cuanto a las regiones sobrelapadas el contenido de TA se encuentra subrepresentado (Figura 14).

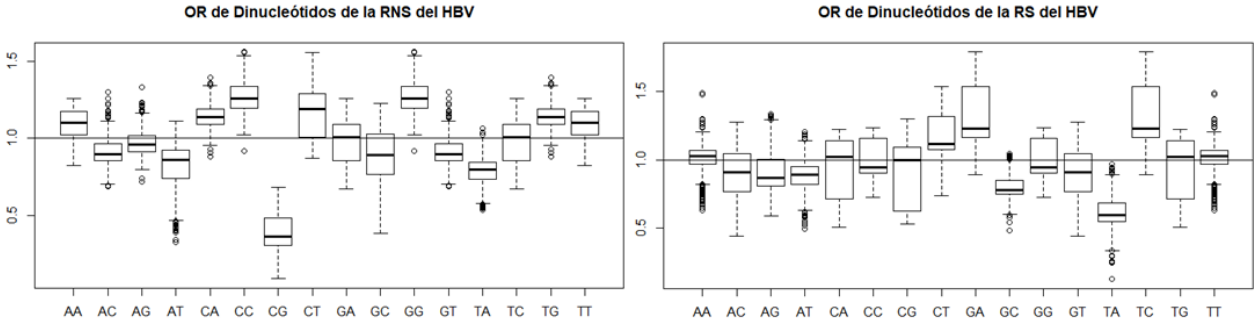


Figura 14. Odd Ratio de los dinucleótidos presentes en el genoma del HBV, contrastando las regiones no sobrelapadas (RNS) con las sobrelapadas (RS). Valores por arriba de 1 indica que existe una sobrerrepresentación de ese dinucleótido, mientras que por debajo de ese valor hay una subrepresentación

Con base en estos resultados, es claro que existen diferencias en la composición dinucleotídica; en primer lugar, de manera general entre regiones sobrelapadas y no sobrelapadas, pero también el PCA indica variación dentro de cada una de las regiones.

8.4. **Uso y sesgo de codones en el genoma del HBV**

8.4.1. Número Efectivo de Codones (ENC)

Los resultados indican que existe variación significativa en el uso absoluto de codones. En primer lugar, entre el ORF C (49.47, ±1.34), ORF P (55.46, ±0.82), ORF S (57.21, ±2.75) y ORF X (55.57, ±1.41). Es importante recordar que valores de ENC mayores a 60 implican un sesgo muy pequeño o nulo ya que se encuentran presenten en el ORF todos los codones sinónimos posibles, sin embargo, se considera que existe sesgo significativo cuando los valores de ENC son menores a 35 (He et al., 2019). Con base en lo anterior, en el genoma del HBV los ORFs que presentan un mayor y menor sesgo de codones son el ORF S y ORF X, respectivamente (Figura 15A).

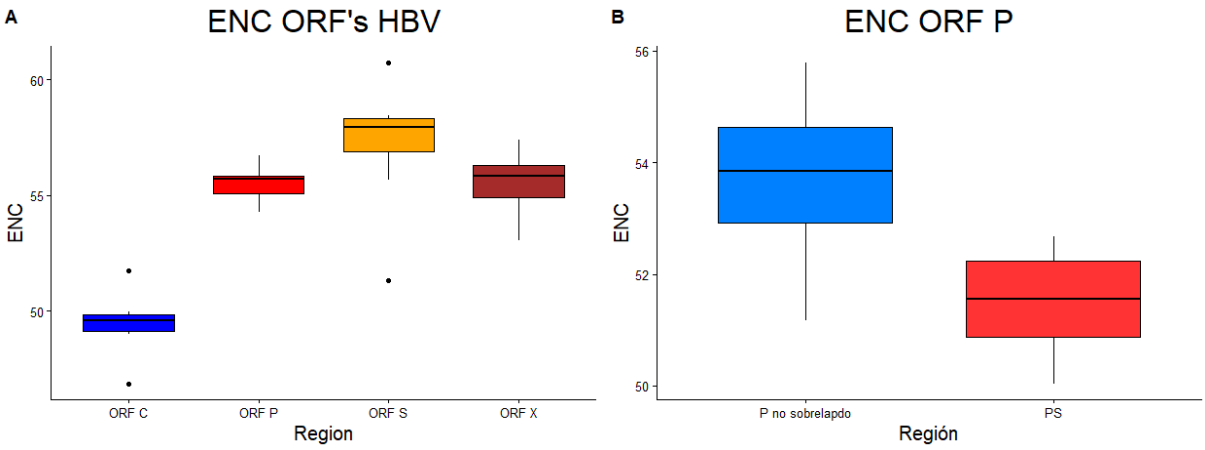


Figura 15. Número Efectivo de Codones (ENC) en el genoma del HBV. A) ENC de los cuatro ORFs del HBV, se puede observar un mayor sesgo en el ORF C, B) ENC de la región sobrelapada y no sobrelapada del ORF P, existe diferencia significativa con mayor sesgo en la región no sobrelapa del ORF P.

Para contrastar las diferencias del uso de codones entre regiones sobrelapadas y no sobrelapadas, se comparó el uso de codones entre estas dos regiones en el ORF más grande, el ORF P. En la Figura 15B se puede observar que la región de P, que se sobrelapa con el ORF S, presenta un sesgo significativamente mayor (51.47 ±1.00) en comparación con la región no sobrelapada de P (53.71 ±1.46).

8.4.2. Relación entre ENC y los sitios sinónimos de los codones (GC3)

Los cambios generados por presión mutacional en la tercera posición de cada codón son considerados neutros, cuando ocurren no generan cambios de aminoácidos debido a la propiedad degenerada del código genético. Una manera de analizar estos patrones es mediante la estimación de los cambios de contenido de G y C en la tercera posición (GC3) en un conjunto de secuencias. Por otra parte, se tiene establecido que en función de determinada frecuencia relativa de GC3 se espera la probabilidad de tener un determinado valor de ENC (A. Roth et al., 2012). Así, si cierto valor se sale de lo esperado significa que otro proceso además de la presión mutacional está sesgando el uso de codones en general; tal proceso es la selección natural.

Con lo anterior, una de las maneras para visualizar e identificar la influencia de la presión mutacional o selección natural en las secuencias es mediante gráficas de relación ENC vs GC3. Particularmente, en este trabajo se obtuvieron dos gráficas correspondientes a las regiones sobrelapadas y no sobrelapadas del ORF P (Figura 16). En cuanto a la región del ORF P sobrelapada, se observa que todos los valores esperados de ENC en función de cierto valor GC3 caen por debajo de la curva de valores esperados. Esto indica que la presión mutacional es el principal proceso que está determinando el uso de codones. En lo que respecta a la región de P no sobrelapada, además de la presión mutacional, la selección natural influye en el uso de codones ya que valores de ciertas secuencias caen por arriba de los valores esperados.

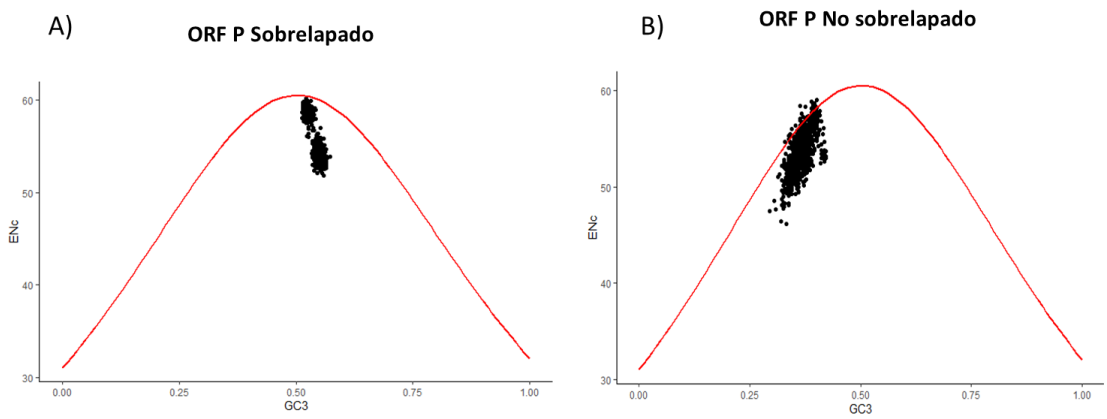


Figura 16. Gráficas GC3 vs ENC en las regiones del ORF P. A) El sesgo en el uso de codones es determinado por la presión mutacional en la región sobrelapada del ORF P, B) El sesgo es determinado por la presión mutacional y la selección natural en la región no sobrelapada del ORF P.

8.4.3. Frecuencia relativa del uso de codones sinónimos (RSCU)

El valor de RSCU de un codón es el producto del cociente entre la frecuencia observada entre y la frecuencia esperada de cierto codón. Valores igual a 1 indican que no hay sesgo, mientras que valores >1.5 y >0.6 son considerados como sobrerrepresentados y subrepresentados, respectivamente (He et al., 2019).

Se obtuvieron los valores de RSCU para cada región sobrelapada y no sobrelapada de cada ORF y en general presentes en el genoma del HBV. Los resultados de esto último se observan en la Figura 17, existen patrones de variación en la frecuencia de determinados codones en función de si la región genómica es sobrelapada o no. En el eje de las X se representan todos los codones sinónimos ordenados de mayor a menos nivel de degeneración y en el eje de las Y los valores de RSCU. Uno de los principales patrones encontrados es que el uso de codones con niveles de degeneración más altos se encuentra en mayor frecuencia en las regiones sobrelapadas comparado con las no sobrelapadas: 3/6 codones que codifican para Leu (CUU, CUC, CUG), 4/6 para Ser (UCU, UCC, UCG, AGU) y 4/6 para Arg (CGU, CGC, CGA, CGG) predominan en las regiones sobrelapadas.

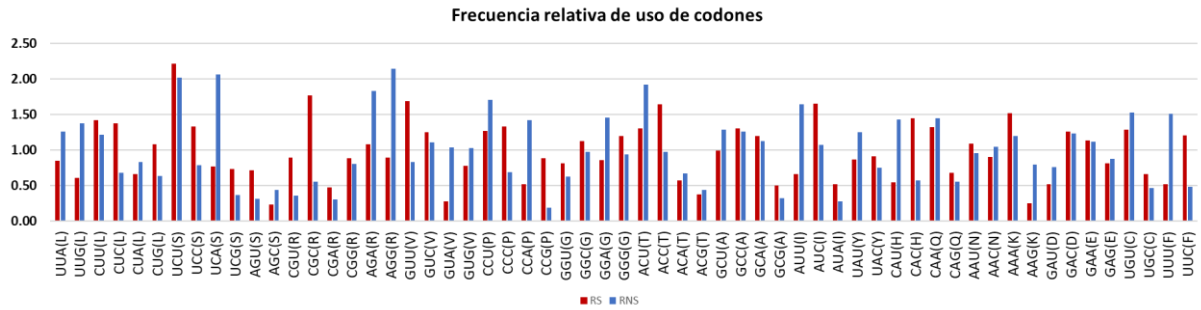


Figura 17. Valores del RSCU de cada codón correspondiente tanto a la región sobrelapada (RS, en rojo) como la no sobrelapada (RNS, en azul). Valores por arriba de 1 indican una sobrerrepresentación de ese codón en determinada región, valores por debajo de 1 una subrepresentación.

Se realizó un PCA utilizando los valores de RSCU de estas dos regiones, pero con énfasis en cada ORF. Se encontró que existen patrones de agrupamiento por similitud de uso de codones, específicamente, la región sobrelapada P con el ORF sobrelapado S por un lado (rectángulos rojos), y las regiones no sobrelapadas de P y C por el otro (rectángulos azules). Por otro lado, en general el ORF X tiene un uso de codones completamente diferente al del resto de ORFs, tanto su región sobrelapada como la no sobrelapada presentan valores atípicos, este patrón ocurre de igual manera con la región sobrelapada de C.

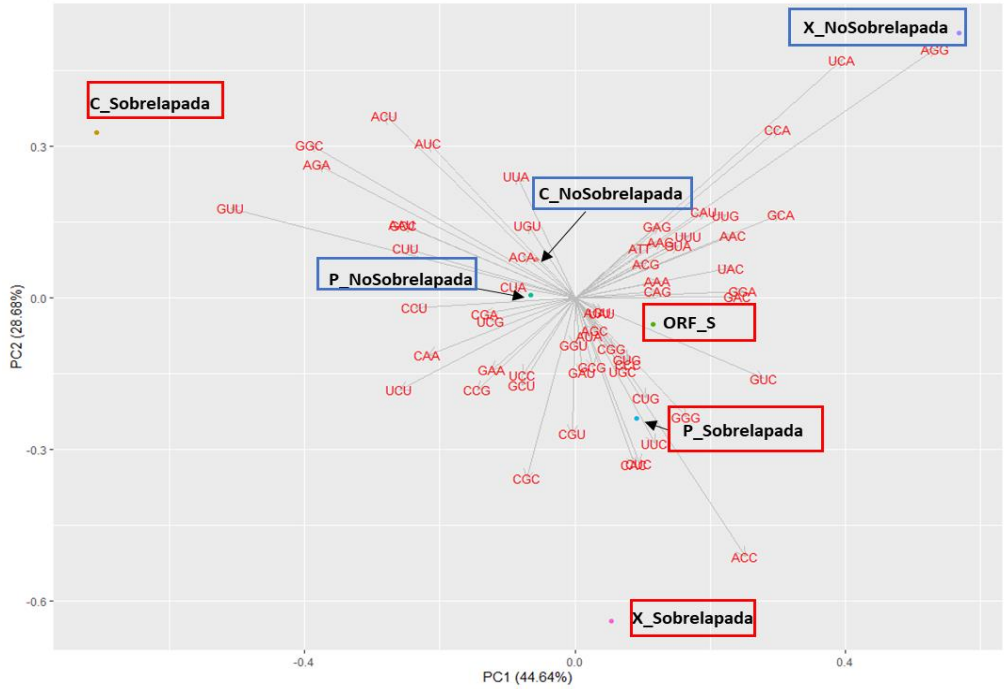


Figura 18. Gráfica bidimensional del Análisis de Componentes Principales (PCA) utilizando los valores de RSCU en las diferentes regiones genómicas del HBV. En rectángulo rojo están señaladas las regiones sobrelapadas y en azul las no sobrelapadas. Existe un patron de uso de codones

similares entre regiones sobrelapadas, particularmente del ORF S y la región sobrelapada del ORF P; la región no sobrelapada del ORF P y el ORF C presentan similitud en el uso de codones, no obstante, el ORF X presenta un uso de codones totalmente diferente al resto, tanto su región sobrelapada como su no sobrelapada.

Finalmente, para determinar la influencia de las composición nucleotídica en la variabilidad del uso de codones, se realizaron correlaciones de Pearson entre los datos originales de composición y el PC1 y PC2 de los datos de RSCU. Los resultados indicaron que las diferencias en la composición de GC (0.97, $p < 0.01$), GC3 (0.91, $p < 0.01$) y C3 (0.86, $p < 0.01$) son las variables mejor correlacionadas con PC1.

8.5. **Variabilidad genética en el genoma del HBV**

8.5.1. Patrones de variación genética en el genoma completo del HBV

Con la finalidad de analizar la variación genética a lo largo de las diferentes regiones genómicas del HBV, se utilizó la métrica de entropía de Shannon que mide la variabilidad de bases por cada sitio en genoma. Los resultados de entropía de Shannon se graficaron en *R* y se obtuvo la Figura 19. Se observa la variación genética en cada posición del genoma en las diferentes regiones, las regiones sobrelapadas se encuentran sombreadas de rojo y las no sobrelapadas de amarillo. Existen dos patrones que se pueden distinguir, en primer lugar hay una mayor variabilidad nucleotídica en las regiones sobrelapadas, especialmente en las regiones P y C. En segundo lugar, las regiones sobrelapadas presentan una menor variación de nucleótidos, particularmente, la primera parte de la región PS y la región PC.

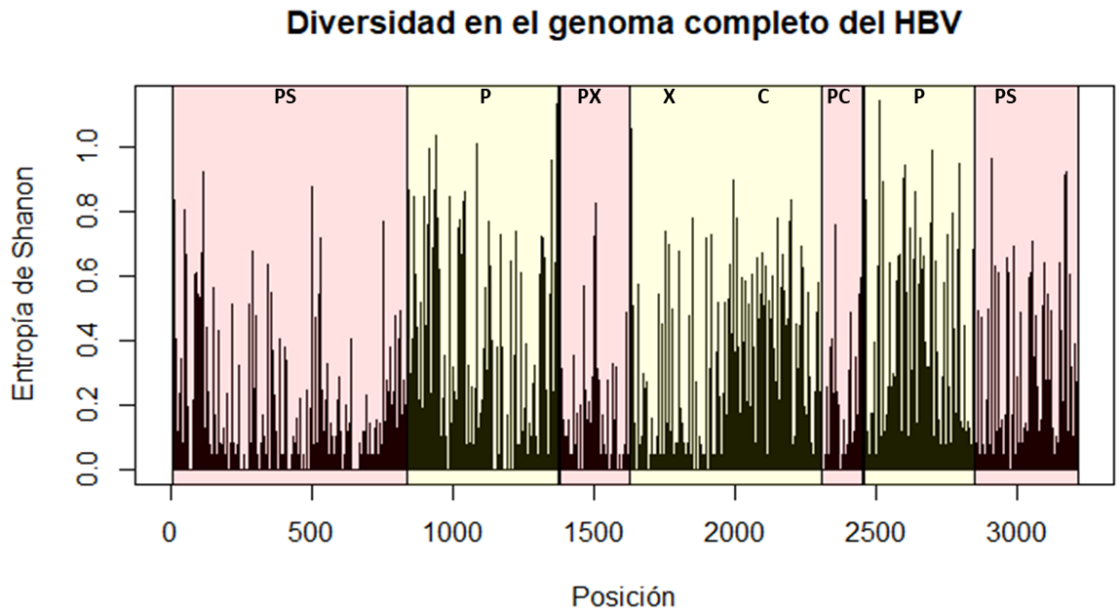


Figura 19. Patrones de variabilidad genética en el genoma del HBV utilizando la entropía de Shannon. Las regiones sobrelapadas están sombreadas de rojo mientras que las no sobrelapadas de amarillo. Se observa una mayor variación de bases por sitio en las regiones no sobrelapadas que en las sobrelapadas. PS= Región sobrelapada del ORF P con el ORF S, P= Región no sobrelapada del ORF P, PX= Región sobrelapada del ORF P con el ORF X, X= Región no sobrelapada del ORF X, C= Región no sobrelapada del ORF C, PC= Región sobrelapada del ORF P con el ORF C.

Para determinar de una manera más cuantitativa la diferencia de variabilidad genética entre los dos tipos de regiones del genoma se calculó el índice Pi (**π),** el cociente de Shannon, el número total de mutaciones (**Eta**) y el número total de sitios polimórficos (**S**) en cada uno de los conjuntos de datos, es decir, en el conjunto de secuencias pertenecientes a las regiones sobrelapadas y no sobrelapadas del genoma de los principales genotipos (Figura 20).

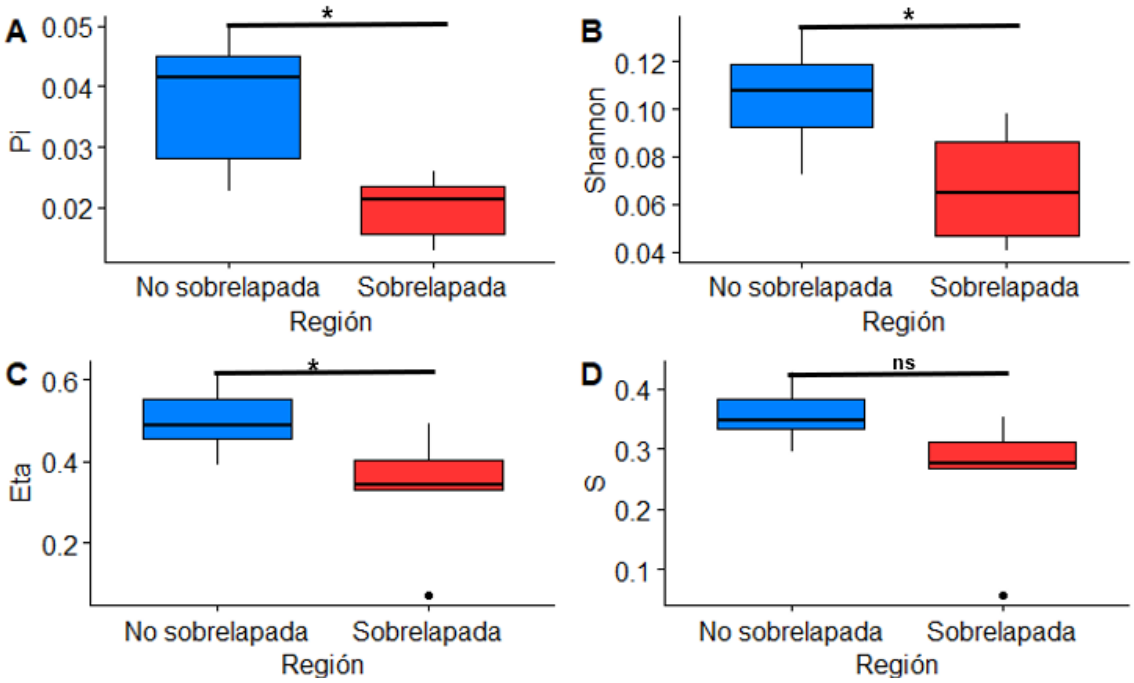


Figura 20. Comparación de variabilidad genética entre regiones sobrelapadas y no sobrelapadas del HBV. Se utilizaron cuatro métricas de variabilidad nucleotídica: A) Pi, B) Shannon, C) Eta (Número de mutaciones) y D) S (Sitios polimórficos), * $p < 0.05$ y ns= no significativo.

Se encontró diferencia significativa en tres de las cuatro medidas de variabilidad analizadas (Pi, Shannon y Eta) al comparar las dos regiones del genoma del HBV. Las regiones no sobrelapadas presentaron una variabilidad significativamente mayor en comparación de las sobrelapadas, estos resultados fueron consistentes con los visualizados en la Figura 19. Estos resultados obtenidos fueron comparando la diversidad general en cada región, no obstante, una mutación puntual puede generar o no alteraciones a nivel de aminoácido en función de en cuál posición del codón ocurra en un respectivo ORF. Mutaciones en la primera y segunda posición son consideradas como no sinónimas debido a que un

cambio en estas posiciones por lo general genera cambio de aminoácido, en contraparte de las mutaciones sinónimas las cuales ocurren en la tercera posición de los codones, las cuales casi en su totalidad no provocan cambios de aminoácido (Yang & Nielsen, 2000).

8.5.2. La variabilidad genética entre regiones del genoma del HBV

Hasta ahora, nuestros resultaron indican que las regiones sobrelapadas presentan una menor variabilidad genética en comparación con las no sobrelapadas (Figura 20). No obstante, aún no sabemos los efectos potenciales que tienen las mutaciones en cada codón de un marco de lectura correspondiente. Debido a la naturaleza sobrelapada del HBV, es importante analizar con mayor detalle los sitios en los que caen las mutaciones para entender el comportamiento de las restrucciones evolutivas en las regiones sobrelapadas. Para hacer esto, se idenficaron las posiciones por codon en las que ocurren las variaciones, tomando como referencia el marco de lectura del ORF P. Como punto de comparación se identificaron también las posiciones alternas en los ORFs con los que se sobrelapada.

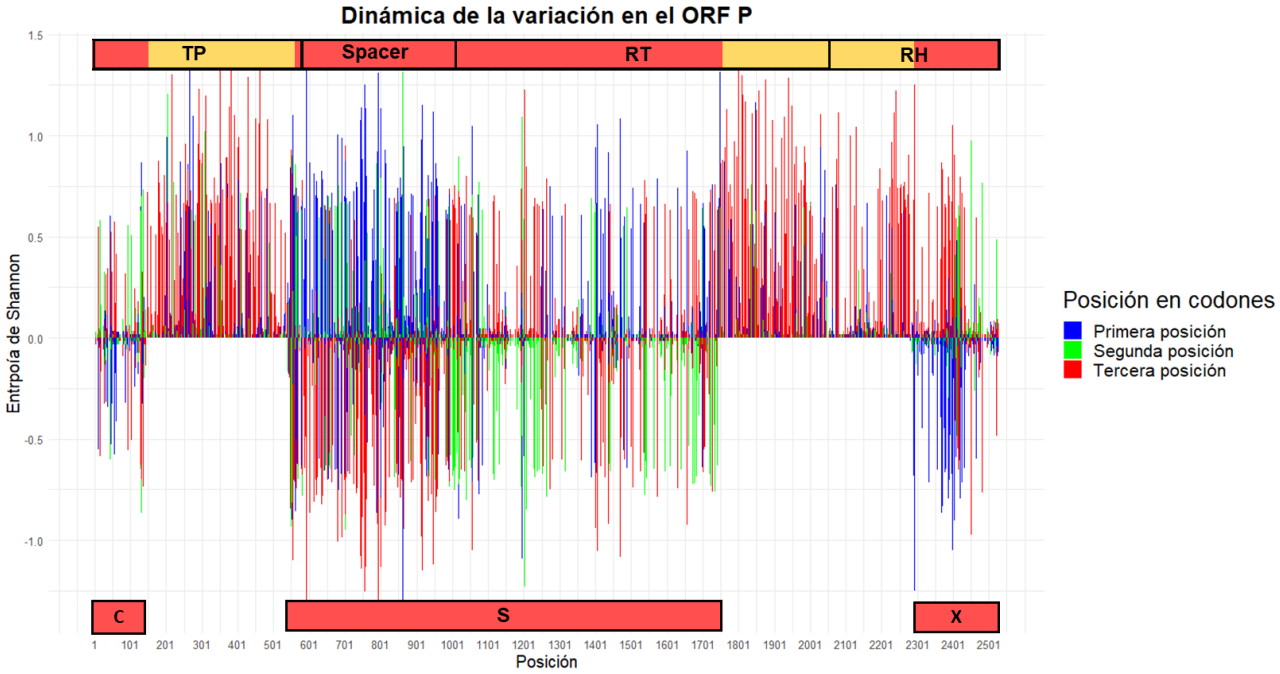


Figura 21. Patrones de variación a nivel codón en el ORF P y los ORFs con los que sobrelapada, el ORF C, ORF X y ORF S. Las variaciones en color azul pertenecen a la primera posición de los

codones, en verde a la segunda posición y en rojo a la tercera. Se indican las regiones del ORF P que codifican para los dominios de la polimerasa correspondientes: TP, Spacer, RT y RH.

En la Figura 21 se encuentra representada mediante una gráfica tipo “espejo” los valores de entropía por posición de base en los codones, los valores positivos corresponden a los valores por posición en función del marco de lectura de P, en la parte superior se encuentra un esquema representando las regiones que codifican para los diferentes dominios; los valores negativos corresponden a los valores por posición en función de C, S y X. La parte sombreada de rojo y amarillo corresponden a las regiones que se sobrelapan y que no se sobrelapan del genoma.

Se puede observar que en el ORF P, a pesar de que las regiones no sobrelapadas son las más diversas, gran parte de estas variaciones ocurren en la tercera posición de los codones. Interesantemente, en la región que sobrelapa con el ORF S (PS), gran parte de las mutaciones ocurren en la posición 1, esto particularmente en la región que codifica para el dominio Spacer de P. Este dominio se sobrelapa con la región Pre S1 y Pre S2 del ORF S, entre estos dos ORFs existe un corrimiento de marco de lectura de +1, por lo que la posición 1 de cierto codón en P corresponde a la posición 3 en S (1p/3s), la posición 2 corresponde a la posición 1 en S 2p/3s y finalmente la posición 3 en P corresponde a la primera posición en S 3p/1s. Sin embargo, la alta variación en las posiciones en las 1p/3s se debe al carácter funcional de Spacer, es un dominio que no presenta una función específica y es un sitio de flexibilidad mutacional (Chen et al. 2013).

8.6. **Análisis de sitios bajo selección natural en el genoma del HBV.**

Una manera de analizar la dinámica evolutiva en los genomas es mediante la estimación de las tasas de cambio en cada posición de codones mediante el uso de algoritmos que utilizan modelos evolutivos para inferir si una posición está bajo presión selectiva o no, las tasas de sustitución por sitio se obtuvieron en el servidor *DataMonkey* (Weaver et al., 2018). Para inferir los sitios específicos bajos selección natural tanto positiva como negativa se utilizó *DataMonkey* para las regiones no sobrelapadas y *OLGenie* (Nelson et al. 2019) para las sobrelapadas.

8.6.1. Patrones de sustitución nucleotídica.

En la Figura 22 se encuentra representado el ORF P completo, el cual está constituido por cuatro regiones que codifican para los dominios correspondientes de la proteína Polimerasa: TP, Spacer, RT y RH. Además, estos diferentes dominios se encuentran sobrelapados con regiones de los ORF C, S y X, respectivamente.

Existe una fluctuación en las tasas de sustitución a lo largo del ORF P, las tasas de sustitución sinónima (dS, en azul) son mayores en regiones no sobrelapadas en comparación con las sobrelapadas; en cuanto a las sustituciones no sinónimas (dN, en naranja) la intensidad parece resaltar en regiones sobrelapadas, especialmente en el dominio Spacer; además, en estas últimas regiones la intensidad de dS parece disminuir.

El ORF S se encuentra totalmente embebido dentro del ORF P, para analizar con un poco más a detalle la dinámica de variación de las tasas de sustitución entre dos marcos de lectura de sobrelapados se realizó un acercamiento específicamente a esta región.

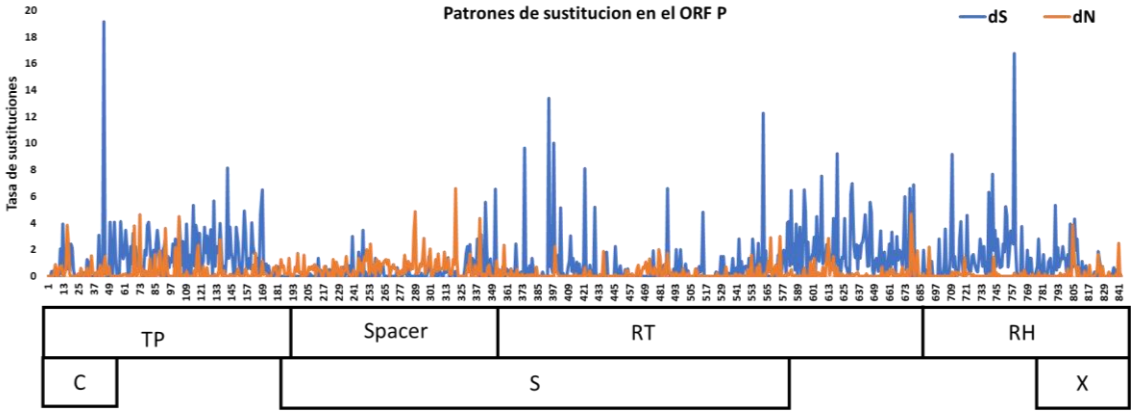


Figura 22. Patrones de sustitución en el ORF P. dS=Tasa de sustituciones sinónimas (en color azul), dN= Tasa de sustituciones no sinónimas (en color naranja). Se encuentran representados las diferentes regiones del ORF P que codifican para los dominios de la poli

8.6.2. Patrones de sustitución nucleotídica en regiones sobrelapadas

El ORF S está constituido por tres regiones que codifican para las regiones correspondientes de la proteína L: Pre-S1, Pre-S2 y S. Pre-S1 y Pre-S2 se encuentran principalmente sobrelapados con la región Spacer del ORF P, mientras que la región S se sobrepalapa totalmente con el dominio RT del ORF P. El primer patrón que se puede observar, es que en el ORF S en comparación con el ORF P en general existe una mayor intensidad en la señal de las tasas tanto sinónimas como no sinónimas (Figura 23). Esto podría tener dos explicaciones: en primer lugar, ocurre un menor número de mutaciones y, como consecuencia, genera la existencia de una menor variabilidad genética por sitio en el ORF P. Enn segundo lugar, con el paso del tiempo esto va a conllevar que ocurra una menor fijación de nucleótidos tanto en posiciones sinónimas como no sinónimas por lo que la señal es menor.

Por otra parte, en la región de Spacer existe una predominancia de sustituciones no sinónimas a diferencia de su contraparte en el ORF S en donde las sustituciones sinónimas predominan, así pues, existe una compensación de presiones selectivas entre los dos ORFs.

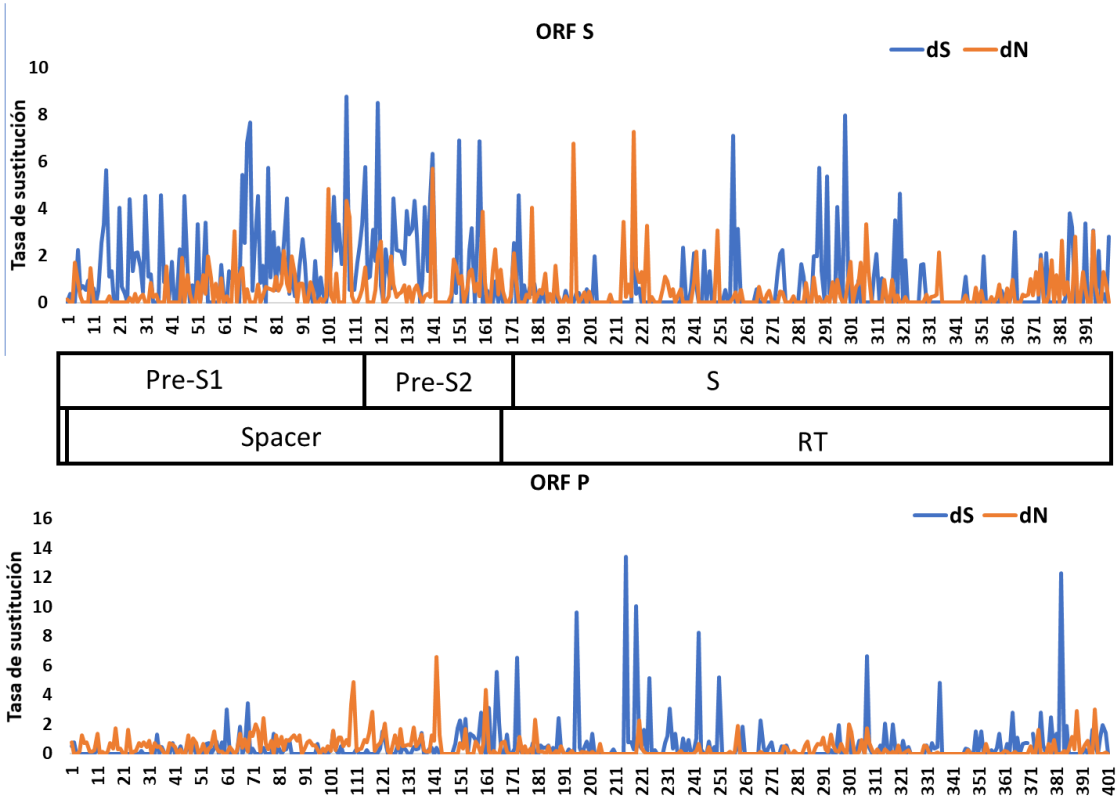


Figura 23. Patrones de sustitución en el sobrelape del ORF S y el ORF P. dS=Tasa de sustituciones sinónimas (en color azul), dN= Tasa de sustituciones no sinónimas (en color naranja). Se encuentran representados las diferentes regiones del ORF P que codifican para los dominios de la polimerasa (Spacer y RT), además de las regiones del ORF S con las que se sobrelapa (Pre-S1, Pre-S2 y S).

Dentro del ORF S, la región S es la que se encuentra bajo mayor presión selectiva debido a la identificación de regiones que codifican para epítopos, es decir, regiones de proteína que son blanco de elementos del sistema inmune Etas regiones se caracterizan por tener una dinámica de constante cambio como mecanismo de escape y sobrevivencia del virus. Para visualizar estos patrones evolutivos se realizó un acercamiento más a la región sobrelapada S/RT.

En la Figura 24 se encuentra el mapa con la variación de las tasas de sustitución en el sobrelape de la región que codifica para la proteína S y la RT de la polimerasa. En la parte superior se observa las tasas de sustitución en el marco de lectura del ORF S acotadas a las diferentes subregiones de S, en la parte inferior se visualizan las tasas de sustitución en el marco de lectura de P acotadas a la región RT, las letras representan las cajas conservadas con sitios catalíticos. Los rectángulos en verde representan epítopos reportados en S, los cuadrados A, B y C en RT representan los sitios catalíticos de la polimerasa.

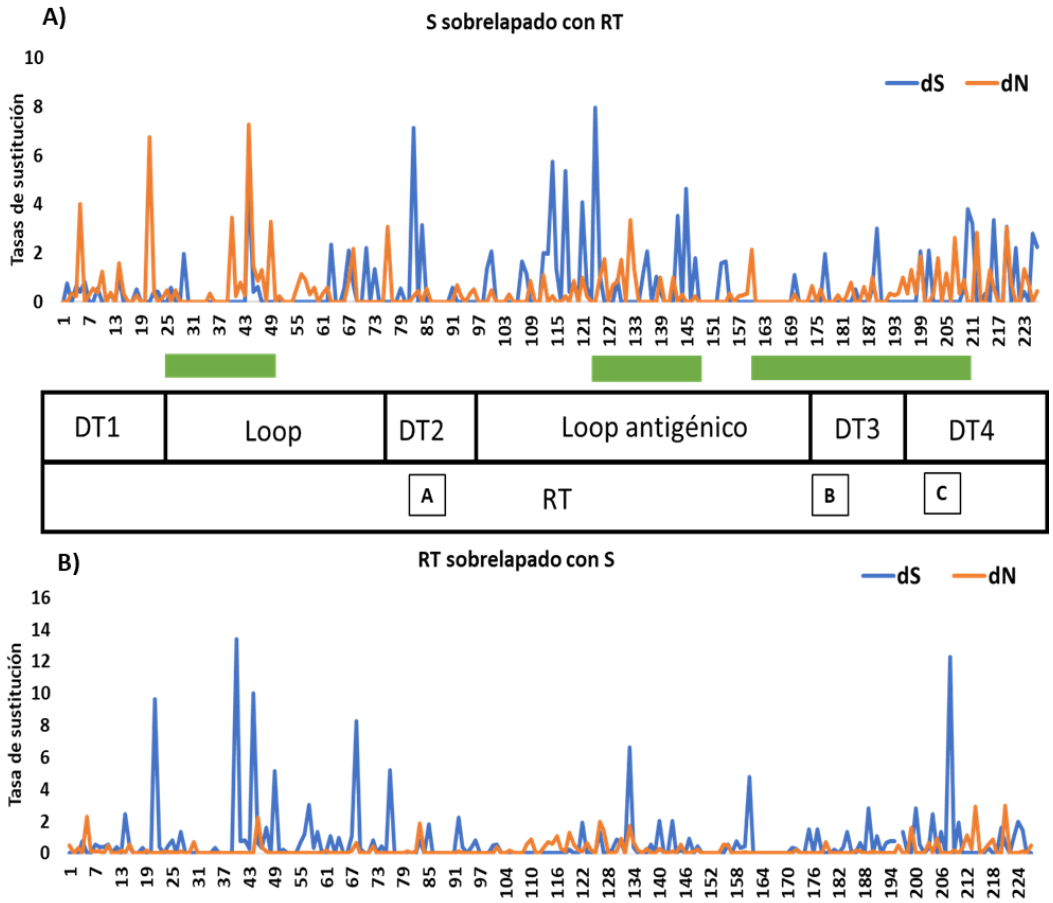


Figura 24. Tasa de sustituciones sinónimas y no sinónimas por posición a lo largo del sobrelape de la región S del ORF S con RT del ORF P. A) Tasas de sustitución de la región que codifica para S (proteína Small), B) tasas de sustitución de la región RT del ORF P, las letras representan las cajas conservadas con sitios catalíticos de la polimerasa. Las líneas verdes representan las regiones de S que presentan epítopos.

Analizando la región S del ORF S se puede observar que en los epítopos las sustituciones sinónimas tienden a ser elevadas en comparación con las demás regiones, esto es una señal de que la selección natural positiva juega un papel importante en favorecer cambios que benefician la replicación y dispersión del HBV. En general, la región S presenta un mayor número de sitios, más del doble, que están bajo cambios no sinónimos (12.3%) que la región sobrelapada RT (5.3%). Esto sugiere nuevamente que S está bajo mayor presión selectiva, probablemente debido a la interección con elementos del sistema inmune.

Además, se muestra que existen patrones de sustitución inversa entre los ORF sobrelapados. Mientras que en ORF los cambios sinónimos son elevados, en esos mismos sitios correspondientes al ORF alterno los predominantes son los no

sinónimos (véase Figura 24, del sitio 19 al 50) . Este proceso evolutivo “cooperativo” es probablemente uno de los aspectos clave en la evolución molecular de los genes sobrelapados en el HBV, ya sea para contrarrestar mutaciones que perjudiquen la replicación y propagación del virus o para generar cambios adaptativos ante las presiones selectivas del sistema inmune y de fármacos.

8.6.3. Inferencias de sitios bajo selección natural en regiones sobrelapadas.

Debido a la limitante de la identificación de sitios bajo selección natural en las regiones sobrelapadas mencionadas en la metodología, se utilizó *OLGenie*, un programa enfocado a cuantificar las diferencias sinónimas y no sinónimas considerando los dos ORFs sobrelapados. La limitante de este método es que solo toma en cuenta cambios observados, no realiza construcción de estados ancestrales para la inferencia de cambios a lo largo del tiempo. No obstante, es una herramienta complementaria a los métodos clásicos utilizados en DataMonkey.

En la Figura 25 se muestra dos gráficas obtenidas con los resultados de OLGenie, estas representan los sitios identificados bajo señales de selección natural positiva (asteriscos verdes) y selección natural negativa (asteriscos rojos) correspondientes a la región sobrelapada S/RT de la Figura 24.

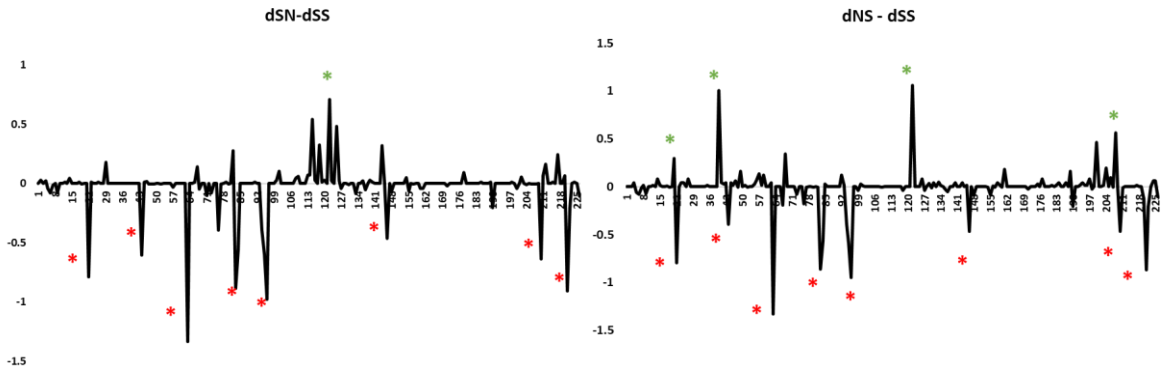


Figura 25. Sitios bajo señales de selección natural en región sobrelapada S/RT. Los asteriscos verdes representan sitios bajo selección positiva, los asteriscos rojos sitios bajo selección negativa. dSN-dSS= Diferencia de la tasa de sustituciones que son sinónimas en S pero no sinónimas en RT, entre la tasa de sustituciones que son no sinónimas en ambos ORFs; dNS-dSS= Diferencia de la tasa de sustituciones que son no sinónimas en S pero sinónimas en RT, entre la tasa de sustituciones que son no sinónimas en ambos ORFs.

En las siguientes tablas se representan los sitios identificados bajo selección positiva utilizando tanto OLGenie como FEL en DataMonkey, en la Tabla 3 se encuentran los sitios con referencia a la región sobrelapada S, seguido de la sustitución que ocurre tanto a nivel de aminoácidos como de codón, posteriormente se señala si existe una implicación biológica reportada y finalmente el ScoreB62 indica al grado de alteración fisicoquímica de determinada sustitución de aminoácido, valores negativos implican alteración significativa.

Tabla 3. Sitios bajo selección positiva en el ORF S utilizando ambos métodos.

Posición en S	Sustitución en S		Implicación	Score B62
21	Leu/Ser	T**T**G →T**C**G	Epítopo CTL	-4
40	Asn/Ser	A**A**T → A**G**T	Epítopo CTL	1
122	Lys/Arg	A**A**A → A**G**A	Epítopo Ab	3
208	Ile/Thr	A**T**C → A**C**C	Mutación de escape CTL	-3

De igual manera, en la Tabla 4 se muestran las características anteriormente mencionadas, sin embargo, correspondientes al marco de lectura de RT.

Tabla 4. Sitios bajo selección positiva en el ORF P utilizando ambos métodos.

Posición en RT	Sustitución en P		Implicación	Score B62
124	His/Asn	**C**AC → **A**AC	Mutación de resistencia	-3
127	Arg/Gly	CGG → **G**GG	-	-4
135	Ser/Tyr	T**C**C → T**A**C	Mutación de resistencia	-3
229	Leu/Val	**T**TG → **G**TG	Mutación de resistencia	1

En S los sitios bajo selección natural positiva corresponden a sitios de interacción con elementos del sistema inmune como las células citotóxicas o CTL, en especial el sitio 208. La sustitución encontrada en este trabajo ha sido reportada por trabajos experimentales como mutación de escape a células CTL. Respecto a

RT, los sitios bajo selección positiva identificados corresponden a sitios reportados como mutaciones de resistencia a fármaco

Finalmente, cuando se identifican sitios bajo selección negativa, significa que una presión selectiva, por ejemplo, fármacos, están generando mutaciones en sitios funcionalmente importantes y que pueden afectar negativamente la adecuación biológica del virus. Los sitios bajo selección negativa identificados fueron 22 , 63, 84 , 95, 146, 210, 221 para S y 30, 52, 71, 93, 95, 103, 154, 207 para RT. Específicamente, mutaciones en el sitio 207, el cual corresponde al sitio catalítico C de RT, se han reportado como consecuencia de la acción de la Lamivudina, un fármaco comúnmente usado para el HBV y HIV para inhibir la actividad de la retrotranscriptasa y, consecuentemente, impedir la replicación viral.

Así pues, el análisis de las dinámicas evolutivas y de sitios bajo selección natural en el HBV, así como en los demás genomas virales, se puede considerar como una herramienta importante para la identificación de sitios blanco del sistema inmune y posibles blancos farmacéuticos.

9. DISCUSIÓN

Al igual que otros miembros de la familia Hepadnaviridae, el Virus de la Hepatitis B (HBV) presenta una organización genómica compleja, en el genoma se encuentran cuatro marcos abiertos de lectura (ORFs) que codifican tanto para las proteínas estructurales (Core, L, M, S) como para las no estructurales (Polimerasa, X y HBeAg) necesarias para realizar el ciclo replicativo del HBV de manera exitosa. Esta compactación genómica es producto de la acción de diferentes presiones selectivas a lo largo de la historia evolutiva del HBV originando así sobrelapamiento de los ORFs para optimizar la producción de proteínas en un tamaño genómico limitado. En el HBV todos los ORFs se encuentran sobrelapados en algún grado, por lo que existen dos tipos de regiones en el genoma: las regiones sobrelapadas y las no sobrelapadas. Se ha reportado principalmente en virus de RNA que la composición nucleotídica y las tasas de sustitución entre estas regiones varían considerablemente (Hughes & Hughes, 2005; Pavesi, 2006, 2015; Rancurel et al., 2009). Sin embargo, no se conoce estudios integrales que consideren analizar las diferencias en composición, variabilidad y dinámicas evolutivas de las regiones sobrelapadas utilizando al HBV como modelo.

Por lo anterior, en el presente trabajo se llevaron a cabo análisis con dos principales enfoques con el objetivo de determinar las principales características genómicas y evolutivas de las regiones sobrelapadas del HBV que conlleven a mejorar el entendimiento de la capacidad adaptativa del virus. El primer enfoque tiene que ver con una caracterización de la composición nucleotídica y el sesgo en el uso de codones a lo largo de las diferentes regiones del genoma del HBV. El sesgo en la frecuencia observada de los codones está determinada principalmente por dos procesos evolutivos: (1) la selección natural puede aumentar la frecuencia de codones en genes que se presenten una expresión continua; la principal presión selectiva la ejerce la abundancia de determinados tRNA mensajeros en la célula hospedera, cuyo objetivo principal es optimizar la velocidad/eficiencia del proceso traduccional; y otros factores como la eficiencia transcripcional, estructuras secundarias de RNA y la estructura de la proteína pueden ejercer también presión selectiva (Shackelton et al., 2006), (2) la presión mutacional se refiere a las diferencias intrínsecas en la capacidad de los

genomas para generar mutaciones en los procesos replicativos, esto moldea los patrones generales de composición de bases y por ende en los genomas, gran parte de estas mutaciones suceden en sitios de los codones que no generan cambio de aminoácido por lo que son considerados procesos evolutivos neutrales (Bernardi & Bernardi, 1985; Sharp et al., 1993).

El segundo enfoque de análisis del genoma del HBV en este estudio tiene que ver con la caracterización de la variabilidad genética y de la dinámica evolutiva de las regiones sobrelapadas. Se ha reportado anteriormente, que las dinámicas de variación genética y de los tipos de sustitución (sinónima o no sinónima) que ocurren en los genomas de los organismos vivos y de virus varía ampliamente según la región genómica, por ejemplo, entre regiones codificantes y no codificantes, regiones ricas en GC o AT, regiones reguladoras, entre otras (Lynch, 2010; Smith et al., 2002). Particularmente, los estudios que contrastan el cambio genético entre regiones sobrelapadas y no sobrelapadas de un gen son más escasos; sin embargo, existen claras diferencias en la manera en que estos dos tipos de regiones evolucionan (Miyata & Yasunaga, 1978; Krakauer, 2000). En los virus, se requiere una mayor atención para su estudio debido a que la mayoría de las regiones o genes completos sobrelapados codifican para dominios o proteínas funcionales importantes que aumentan la probabilidad de replicación y subsistencia de las partículas virales (Rancurel et al., 2009). Por lo anterior, surgió el interés de analizar con mayor detalle la evolución de las regiones sobrelapadas en el HBV. Éstas corresponden al 50% de su genoma codificante por lo que los resultados generados podrían contribuir a la evolución del HBV y sus implicaciones en el ciclo replicativo.

Los virus sin importar si son de DNA o RNA, presentan diferencias en cuanto a la composición nucleotídica de sus genomas; estas variaciones ocurren entre virus, virus-hospedero, pero también dentro del mismo genoma viral. Nuestros resultados indican que existe variación nucleotídica significativa entre las regiones sobrelapadas y no sobrelapadas del genoma del HBV, en primer lugar, hay una predominancia de C y una disminución en el contenido de A en las regiones sobrelapadas. En cuanto a las regiones no sobrelapadas el nucleótido más abundante es T seguido de A, mientras que el menos abundante es C. Consecuentemente, el contenido de GC se encuentra principalmente asociado a las regiones de sobrelapadas y el contenido de AT a las no sobrelapadas (Figura 11). En otros genomas, el sesgo de composición de

determinado nucleótido es una firma genómica característica de determinada familia y grupo de virus, por ejemplo, en el HIV existe un enriquecimiento en el contenido de A y una disminución en C (Kuly & Berkhout, 2012).

Pavesi et al (2018) analizaron la composición de regiones de genes sobrelapados en virus encontrando de manera similar que el contenido de C es predominante y AT está subrepresentado. Además, encontraron la presencia de determinados dinucleótidos asociados a las regiones sobrelapadas tales como CpC y CpG, en nuestros resultados además de encontrar la predominancia de estos mismos dinucleótidos también se encontró un alto contenido de TpC en comparación con las no sobrelapadas en donde predominó el contenido de TpT, ApA y TpA (Figura 13). Se ha reportado que el sesgo en la presencia de determinados dinucleótidos en genomas virales puede tener importantes implicaciones para la replicación de los virus, y por lo tanto, son de interés biomédico.

La mayoría de los virus de RNA que infectan vertebrados y plantas, así como los virus de DNA pequeños presentan una disminución del contenido de los dinucleótidos CpG y TpA. Esto probablemente como respuesta adaptativa para simular la composición del RNAm del hospedero. Tomando en cuenta lo anterior, se ha realizado diversos estudios en donde la alteración en la frecuencia dinucleotídica influye en la capacidad replicativa y de sobrevivencia de los virus. En *Echovirus 7,* por ejemplo*,* el aumento del contenido de CpG y TpA mediante mutantes generó una reducción en la capacidad del virus para replicarse posterior a entrar a la célula hospedera, sugiriendo que el hospedero previene la replicación del material genético viral basándose en la composición sesgada del genoma viral (Atkinson et al., 2014; Fros et al., 2017; Karlin et al., 1994). Resultados similares se han observado en el virus de la influenza A (IAV) en donde se reportó un aumento en la atenuación de la replicación viral así como una inducción de la respuesta inflamatoria y adaptativa al maximizar la frecuencia de CpG, esto significa que la alteración de la composición de dinucleótidos podrían ser un método alternativo para la generación de vacunas inmunoreactivas (Gaunt et al., 2016). En el caso del HBV, las regiones sobrelapadas contienen un alto contenido de CpG y podrían ser sitios candidatos para realizar análisis experimentalmente con fines similares a los mencionados anteriormente.

Por otra parte, ha sido ampliamente estudiado que la composición de nucleótidos influye considerablemente en el uso y sesgo codones particularmente en virus de RNA, en donde la variación del contenido de AT es el principal elemento de composición que está asociado a la determinación del sesgo de codones (Belalov & Lukashev, 2013; Khandia et al., 2019; Sheikh et al., 2020). Este tipo de estudios en virus de DNA son menos comunes, particularmente en el HBV son escasos y no son claros, por ejemplo Qi et al. (2020) realizaron análisis del número efectivo de codones (ENC) y la frecuencia relativa de uso de codones sinónimos (RSCU) en los genomas de las genotipos B, C y D. No obstante, en su metodología y resultados no especifican en cuál ORF y/o región se aplicaron dichos análisis, debido a la complejidad de los marcos de lectura sobrelapados y la organización circular del genoma no es posible aplicar estos análisis en las secuencias de genomas completos sin considerar separar los marcos abiertos de lectura (ORFs).

El genoma del HBV consta de cuatro ORFs: C, P, S y X, los cuales codifican las proteínas necesarias para completar el ciclo replicativo del virus, se identificaron diferencias significativas en el uso total de codones entre los ORF mediante el ENC (Figura 15ª), se demuestra que los ORF C y ORF S son los que presentan un mayor y menor sesgo, respectivamente. Los ORFs o genes con poco sesgo están bajo menor presión selectiva por parte de los RNAs de transferencia (tRNAs) debido a que presentan la mayoría de los codones posibles, esta característica se ha asociado a genes que son altamente expresados debido a que las proteínas que codifican son cruciales para la sobrevivencia del virus (Paul M. Sharp & Li, 1986; Sheikh et al., 2020). En el caso del HBV, el ORF S codifica tres proteínas estructurales que forman parte de la membrana del virión (S,M,L), por lo que su expresión es constitutiva. Por otra parte, el ORF C codifica para dos proteínas, una estructural (Core) y una no estructural (HbeAg); ésta última en un estado crónico de la infección del HBV su expresión se detiene, por lo que en general el ORF C se expresa en menor proporción en comparación con el ORF S (Juszczyk, 2000).

Un resultado interesante fue el contraste en el uso de codones entre regiones sobrelapadas y no sobrelapadas del ORF P. Este ORF codifica para la polimerasa y abarca aproximadamente el 78% del genoma del HBV. Este mismo ORF P se encuentra sobrelapado aproximadamente en un 50% con los demás ORF por lo que es una condición óptima para comparar las limitaciones en el uso

de codones dentro de un mismo ORF. En la Figura 15B se muestra que en efecto, existen diferencias significativas en los patrones de uso de codones entre las regiones sobrelapadas y no sobrelapadas. Esto sugiere una vez más que las dinámicas de composición y evolutivas de las dos regiones son diferentes. Las gráficas de ENC vs GC3 del ORF P (Figura 16) corroboraron que las regiones están bajo procesos evolutivos diferenciales, por un lado, tenemos que tanto en la región sobrelapada como no sobrelapada, la presión mutacional es la principal fuerza que influye en determinar el uso de codones. Sin embargo, en la región sobrelapada se identificó la influencia de la selección natural, debido a que diversos valores de ENC de secuencias no sobrelapadas eran mayor a lo que se esperan bajo neutralidad.

Así pues, las regiones sobrelapadas en el HBV tienen un uso absoluto de codones diferente de las regiones no sobrelapadas, además de presentar significativamente un menor sesgo de codones y éste es determinado no sólo por acción de la presión mutacional, sino también por la selección natural. Además, al igual que en el genoma de bacteriófagos de la familia *Microviriadae* (Pavesi, 2006), en el HBV los ORFs tienen diferentes patrones de usos de codones. Existe un especial contraste entre regiones sobrelapadas y no sobrelapadas de los ORFs (Figura 18). De los cuatro ORFs del HBV, las regiones del ORF X presenta un uso de codones totalmente diferente al resto de las regiones genómicas; esto es un indicio de que probablemente este ORF fue el último en crearse dentro de la historia evolutiva del HBV. Lo anterior ha sido sugerido por Suh et al. (2014) en donde muestran que, mediante un enfoque de distribución filogenética y bajo la premisa que el ORF más reciente, se presenta en un menor número de taxa y encontraron que el gen X de la familia H*epadnaviridae* era el más restringido, presentándose sólo en el género *Orthohepadnavirus* (Pavesi, 2006, 2015; Pavesi et al., 2013).

A pesar de que el virus de la hepatitis B se clasifica como un virus de DNA, se replica con un intermediario de RNA, codifica su propia retrotranscriptasa y carece de capacidad de corrección de errores en el proceso de replicación (proofreading). Esto le otorga al HBV la capacidad de tener un alta tasa mutacional, que en conjunto con la acción de los procesos evolutivos le confiere una tasa evolutiva similar a la de los retrovirus (10^{-4} sustituciones/sitio/año). No obstante, discernir el origen y la estimación de las tasas evolutivas del HBV no ha sido fácil debido a la complejidad biológica del virus. El genoma del HBV

tiene diversas restricciones al cambio evolutivo, tiene un tamaño pequeño (3215 nt aprox.), marcos de lectura sobrelapados y estructuras secundarias de RNA, estas características permiten que exista una alta variabilidad en la diversidad genética y en las tasas de sustitución a lo largo del genoma, por ejemplo, entre regiones sobrelapadas y no sobrelapadas (Harrison et al., 2011; Kay & Zoulim, 2007)

Las regiones sobrelapadas presentan mayores restricciones evolutivas debido a que una mutación puede tener repercusiones simultáneamente en dos marcos de lectura. No obstante, esto no ha impedido al HBV el ser capaz de evolucionar adaptativamente ante las diferentes presiones selectivas de su estructura genómica, así como del hospedero a lo largo de la historia evolutiva de dicho virus. Si bien es cierto que las regiones sobrelapadas presentan una menor variabilidad genética (Figura 20), gran parte de estas variaciones ocurren en sitios no sinónimos, es decir, en la primera y segunda posición nucleotídica de los codones (Figura 21). En contraste, las variaciones genéticas en las regiones no sobrelapadas ocurren en sitios sinónimos de las terceras posiciones de los codones. Las dinámicas de variación genética en el genoma del HBV varían considerablemente en función de la región. En el ORF P, interesantemente en la región que sobrelapa con el ORF S (PS), gran parte de las mutaciones ocurren en la posición 1, específicamente en la región *Spacer. S*in embargo esta alta variación se debe al carácter funcional de Spacer, el cual codifica para un dominio de la polimerasa que no presenta una función específica y por lo tanto está bajo poca presión selectiva, lo que lo convierte en un sitio de flexibilidad mutacional (Chen et al. 2013).

Con base en estos resultados de variabilidad, es de esperar que las tasas de sustitución sinónima (dS) sean mayores en regiones no sobrelapadas y las tasas de sustitución no sinónima (dN) en las regiones sobrelapadas del ORF P (Figura 22). Estos patrones evolutivos también se han reportado en genomas con regiones sobrelapadas de papilomavirus, virus de inmunodeficiencia en virus, procariontes (Hughes et al., 2001; Hughes & Hughes, 2005; Rogozin et al., 2002). No obstante, en una región sobrelapada las dinámicas evolutivas pueden variar en función del dominio de la proteína codificada de los dos ORFs. Al analizar a más detalle las tasas de sustitución en la región sobrelapada del ORF P con el ORF (Figura 23) se encontró que efectivamente que, en cuanto al ORF S la dS fue mayor en las regiones Pre-S1 y Pre-S2 (Pre-S), mientras que en S

se presentan más sitios con alto dN. Biológicamente esto tiene implicaciones importantes, ya que la región Pre-S codifica para dominios primordiales para el ingreso del virus a las células hospederas, así como para la morfogénesis de las partículas virales. Es por ello que es crucial que haya elevadas sustituciones sinónimas para mantener la estructura y funcionalidad de la proteína (Bruss & Vieluf, 1995; Watashi et al., 2014). Por otra parte, en la región del ORF alternativo que se sobrelapada con Pre-S es la región *Spacer* del ORF Pexiste un patrón inverso en donde predominan sustituciones no sinónimas. Este proceso evolutivo de "cooperación" es probablemente uno de los aspectos clave en la evolución molecular de los genes sobrelapados en el HBV, ya sea para contrarrestar mutaciones que disminuya la capacidad de replicación y propagación del virus o para generar cambios adaptativos ante las presiones selectivas del sistema inmune y de fármacos.

Se han reportado diversas variantes genéticas en el genoma del HBV que son de interés clínico, en el ORF P las principales mutaciones se han identificado en la región que codifica para el dominio RT de la polimerasa, esto a consecuencia de las presiones selectivas generadas por los antriretrovirales cuyo objetivo es detener la replicación del HBV al inhibir el proceso de la retrotranscripción. La mutación rtM204V se ha reportado como una mutación de resistencia en pacientes bajo tratamiento con Entecavirm, rtl180M y rtM184V con Lamivudina y rtK103N con Nevirapina. Por otro lado, en el ORF S deleciones de la región Pre-S y la mutación F22L se han encontrado significativamente en mayor frecuencia en pacientes con un desarrollo avanzado de carcinoma hepatocelular, mientras que mutaciones en el ORF X, tales como A1383C, A1461G y T1485C se han asociado a una mayor sobrevivencia de pacientes con estee mismo padecimiento (Jayalakshmi *et al*., 2013; Liu *et al*., 2010; Zhang *et al*., 2016). Estas variantes tienen en común que se encuentran localizadas en regiones sobrelapadas, con base en los resultados de este trabajo de investigación. Es importante considerar los efectos que puedan tener estas mutaciones en los dos marcos de lectura y en las respectivas proteínas que codifican esto con el objetivo de tomar mejores decisiones para un adecuado diseño de estrategias terapéuticas.

Como se había mencionado previamente, las variaciones genéticas en una población viral, se mantienen o eliminan debido a la acción principal de la deriva génica y la selección natural. Esta última cuando actúa sobre mutaciones

que benefician al virus para que complete su ciclo de vida y se propague se dice que actúa la selección positiva, por otro lado, la selección negativa actúa sobre mutaciones deletéreas (Nielsen, 2005). Las mutaciones de escape ocurren en proteínas de membranas que tienen contacto directo con elementos del sistema inmune, y estas ocasionan una disminución de la afinidad entre receptor-ligando evitando así que células CTL reconozcan epítopos virales, estos sitios están bajo presión selectiva constante y por lo tanto en constante cambio por acción de la selección positiva (Frost et al., 2018).

En el HBV, la región S del ORF S codifica para la proteína "*small*", la cual es la proteína de membrana más abundante y presenta un *loop* antigénico altamente inmunorreactivo (Wu et al., 2012). En el presente estudio se identificaron los principales sitios bajo selección positiva en esa región (Tabla 3), mientras que en la región RT del ORF alternativo a diferencia de ser mutaciones de escape, se identificaron como mutaciones de resistencia (Tabla 4), identificadas y caracterizadas en genomas de HBV provenientes de pacientes bajo régimen de medicación antiviral (Liu et al., 2010). Respecto a los sitios bajo selección negativa, éstos se identificaron en sitios que debido a la alta importancia funcional de la proteína que codifican, se deben de mantener conservados, tal es el caso del sitio RT207 que pertenece al sitio catalítico C de la polimerasa, el mal funcionamiento de esta región implicaría la inhibición de la replicación viral.

10. CONCLUSIONES

La composición nucleotídica y el uso de codones entre las regiones sobrelapadas y no sobrelapadas en el genoma del HBV difiere significativamente. El contenido de GC y AT se encuentra principalmente asociado a las regiones sobrelapadas y no sobrelapadas, respectivamente.

Las regiones sobrelapadas del ORF P en el HBV presentan un mayor sesgo en el uso de codones; además, la presión mutacional es el principal proceso evolutivo que influye en determinar el sesgo.

Las regiones sobrelapadas presentan una menor variabilidad genética; sin embargo, las tasas de sustituciones no sinónimas son altas, por lo que varias de estas mutaciones se fijan por la selección positiva debido a diferentes presiones selectivas. Por lo contrario, las regiones no sobrelapadas, que suelen ser más variables,la mayoría de las mutaciones caen en sitios sinónimos.

La evolución molecular de las regiones sobrelapadas en el HBV está en función de tres aspectos: 1) tipo de sobrelape entre los dos ORFs; 2) rol funcional de los dominios de las proteínas codificados por las regiones sobrelapadas; y 3) la deriva génica que podría actuar sobre los ORFs con menor presión selectiva como sucede en la proteína P.

11. REFERENCIAS

1. Araujo, N. M., Waizbort, R., & Kay, A. (2011). Hepatitis B virus infection from an evolutionary point of view: How viral, host, and environmental factors shape genotypes and subgenotypes. Infection, Genetics and Evolution, 11(6), 1199–1207. doi:10.1016/j.meegid.2011.04.017
2. Atkinson, N. J., Witteveldt, J., Evans, D. J., & Simmonds, P. (2014). 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 Research*, *42*(7), 4527–4545. https://doi.org/10.1093/nar/gku075
3. Azeem Mehmood Butt, Izza Nasrullah, Raheel Qamar & Yigang Tong (2016) Evolution of codon usage in Zika virus genomes is host and vector specific, Emerging Microbes & Infections, 5:1, 1-14, DOI: 10.1038/emi.2016.106
4. Baltimore D. (1971). Expression of animal virus genomes. Bacteriological reviews, 35(3), 235–241.
5. Belalov, I. S., & Lukashev, A. N. (2013). Causes and Implications of Codon Usage Bias in RNA Viruses. *PLoS ONE*, *8*(2), e56642. https://doi.org/10.1371/journal.pone.0056642
6. Belshaw, R., Pybus, O. G., & Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research*, *17*(10), 1496–1504. https://doi.org/10.1101/gr.6305707
7. Bernardi, G., & Bernardi, G. (1985). Codon usage and genome composition. In *Journal of Molecular Evolution* (Vol. 22, Issue 4, pp. 363–365). Springer-Verlag. https://doi.org/10.1007/BF02115693
8. Blumberg, B. S. (1965). A “New” Antigen in Leukemia Sera. JAMA: The Journal of the American Medical Association, 191(7), 541.
9. Bohlin, J., & Pettersson, J. H. O. (2019). Evolution of Genomic Base Composition: From Single Cell Microbes to Multicellular Animals. In *Computational and Structural Biotechnology Journal* (Vol. 17, pp. 362–370). Elsevier B.V. https://doi.org/10.1016/j.csbj.2019.03.001
10. Bouchard, M. J., & Schneider, R. J. (2004). The Enigmatic X Gene of Hepatitis B Virus. Journal of Virology, 78(23), 12725–12734.

11. Brandes, N., & Linial, M. (2016). Gene overlapping and size constraints in the viral world. Biology direct, 11, 26.
12. Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends in Microbiology, 13(6), 278–284.
13. Bruss, V., & Vieluf, K. (1995). Functions of the internal pre-S domain of the large surface protein in hepatitis B virus particle morphogenesis. *Journal of Virology*, *69*(11), 6652–6657. https://doi.org/10.1128/jvi.69.11.6652-6657.1995
14. Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015). Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages. *Frontiers in Ecology and Evolution*, *3*(DEC), 143. https://doi.org/10.3389/fevo.2015.00143
15. Cento, V., Mirabelli, C., Dimonte et al. (2012). Overlapping structure of hepatitis B virus (HBV) genome and immune selection pressure are critical forces modulating HBV evolution. Journal of General Virology, 94(Pt_1), 143–149.
16. Chaitanya, K. V. (2019). Structure and Organization of Virus Genomes. In *Genome and Genomics* (pp. 1–30). Springer Singapore. https://doi.org/10.1007/978-981-15-0702-1_1
17. Chen P, Gan Y, Han N, Fang W, Li J, et al. (2013) Computational Evolutionary Analysis of the Overlapped Surface (S) and Polymerase (P) Region in Hepatitis B Virus Indicates the Spacer Domain in P Is Crucial for Survival. PLOS ONE 8(4): e60098
18. Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. Proceedings of the Royal Society B: Biological Sciences, 277(1701), 3809–3817.
19. Crooks, G. E. (2004). WebLogo: A Sequence Logo Generator. Genome Research, 14(6), 1188–1190.
20. Cui, J., Schlub, T. E., & Holmes, E. C. (2014). An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *Journal of Virology*, *88*(11), 6403–6410. https://doi.org/10.1128/jvi.00362-14
21. Dill, J. A., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., & Ng, T. F. (2016). Distinct Viral Lineages from Fish and Amphibians Reveal the Complex Evolutionary History of Hepadnaviruses. Journal of virology, 90(17), 7920–7933.

22. Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792–1797. doi:10.1093/nar/gkh340
23. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, Fourth Edition, 2016.
24. Escobedo-Meléndez, P., Fierro, N., Ruiz-Madrigal, B., Zepeda-Carrillo, E., & Román, S. (2011). Epidemiología de las hepatitis virales en México. *Salud Pública de México*, S37–S45.
25. Fleischmann WR Jr. (1996). Viral Genetics. In S. Baron (Ed.), *Medical Microbiology* (4th ed.). Galveston (TX): University of Texas Medical Branch at Galveston. https://www.ncbi.nlm.nih.gov/books/NBK8439/
26. Fros, J. J., Dietrich, I., Alshaikhahmed, K., Passchier, T. C., Evans, D. J., & Simmonds, P. (2017). CpG and upA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. *ELife*, *6*. https://doi.org/10.7554/eLife.29112
27. Frost, S. D. W., Magalis, B. R., & Kosakovsky Pond, S. L. (2018). Neutral Theory and Rapidly Evolving Viral Pathogens. Molecular Biology and Evolution, 35(6), 1348–1354. doi:10.1093/molbev/msy088
28. Gaunt, E., Wise, H. M., Zhang, H., Lee, L. N., Atkinson, N. J., Nicol, M. Q., Highton, A. J., Klenerman, P., Beard, P. M., Dutia, B. M., Digard, P., & Simmonds, P. (2016). Elevation of CpG frequencies in influenza a genome attenuates pathogenicity but enhances host response to infection. *ELife*, *5*(FEBRUARY2016). https://doi.org/10.7554/eLife.12735
29. Gomes-Filho, J. V., Zaramela, L. S., Italiani, V. C. da S., Baliga, N. S., Vêncio, R. Z. N., & Koide, T. (2015). Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea. RNA Biology, 12(5), 490–500.
30. Harrison, A., Lemey, P., Hurles, M., Moyes, C., Horn, S., Pryor, J., Malani, J., Supuri, M., Masta, A., Teriboriki, B., Toatu, T., Penny, D., Rambaut, A., & Shapiro, B. (2011). Genomic analysis of hepatitis B virus reveals antigen state and genotype as sources of evolutionary rate variation. *Viruses*, *3*(2), 83–101. https://doi.org/10.3390/v3020083

31. Hayer, J., Jadeau, F., Deléage, G., Kay, A., Zoulim, F., & Combet, C. (2013). HBVdb: A knowledge database for Hepatitis B Virus. *Nucleic Acids Research*, *41*(D1), D566–D570. https://doi.org/10.1093/nar/gks1022
32. He, Z., Gan, H., & Liang, X. (2019). Analysis of synonymous codon usage bias in potato virus M and its adaption to hosts. *Viruses*, *11*(8). https://doi.org/10.3390/v11080752
33. Hein, J. & Støvlbæk, J. (1995). A Maximum-Likelihood Approach to Analyzing Nonoverlapping and Overlapping Reading Frames J Mol Evol 40: 181
34. Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet.42:287–99.
35. Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. In *Annual Review of Genetics* (Vol. 42, pp. 287–299). Annu Rev Genet. https://doi.org/10.1146/annurev.genet.42.110807.091442
36. Holmes, E. (2009). *The Evolution and Emergence of RNA Viruses* (E. Holmes (ed.)). Oxford University Press. https://books.google.com.mx/books?id=fpoUDAAAQBAJ&dq=the+evolution+and+emergence+of+rna+viruses&lr=&source=gbs_navlinks_s
37. Hughes, A. L., & Hughes, M. A. K. (2005). Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Research*, *113*(2), 81–88. https://doi.org/10.1016/j.virusres.2005.03.030
38. Hughes, A. L., & Hughes, M. A. K. (2005). Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. Virus Research, 113(2), 81–88. doi:10.1016/j.virusres.2005.03.030
39. Hughes, A. L., Westover, K., da Silva, J., O'Connor, D. H., & Watkins, D. I. (2001). Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the tat andvpr Genes of Simian Immunodeficiency Virus. *Journal of Virology*, *75*(17), 7966–7972. https://doi.org/10.1128/jvi.75.17.7966-7972.2001
40. Hughes, A. L., Westover, K., da Silva, J., O'Connor, D. H., & Watkins, D. I. (2001). Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the tat and vpr Genes of Simian Immunodeficiency Virus. Journal of Virology, 75(17), 7966–7972. doi:10.1128/jvi.75.17.7966-7972.2001

41. Jayalakshmi, M., Kalyanaraman, N., & Pitchapp, R. (2013). Hepatitis B Virus Genetic Diversity: Disease Pathogenesis. In *Viral Replication*. InTech. https://doi.org/10.5772/53818
42. Juszczyk, J. (2000). Clinical course and consequences of hepatitis B infection. *Vaccine*, *18*(SUPPL. 1). https://doi.org/10.1016/S0264-410X(99)00457-0
43. Karlin, S., Doerfler, W., & Cardon, L. R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *Journal of Virology*, *68*(5), 2889–2897. https://doi.org/10.1128/jvi.68.5.2889-2897.1994
44. Kay, A., & Zoulim, F. (2007). Hepatitis B virus genetic variability and evolution. *Virus Research*, *127*(2), 164–176. https://doi.org/10.1016/j.virusres.2007.02.021
45. Keese, P.K. and Gibbs, A. (1992) Origins of genes: 'big bang'or continuous creation? Proc. Natl. Acad. Sci. U. S. A. 89, 9489–9493
46. Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., Das, J., Munjal, A. O., & Singh, R. K. (2019). Analysis of nipah virus codon usage and adaptation to hosts. *Frontiers in Microbiology*, *10*(MAY), 886. https://doi.org/10.3389/fmicb.2019.00886
47. Knipe et al. 2001. Fields Virology. 4th Ed. Lippincott, Williams & Wilkins. USA. Pp.
48. Kozlov, N.N. (1999) Demand of each of 64 codons in genetic overlapping areas. Dokl. Akad. Nauk 367, 544–547
49. Krakauer, D. C. (2000). Stability and evolution of overlapping genes. *Evolution*, *54*(3), 731–739. https://doi.org/10.1111/j.0014-3820.2000.tb00075.x
50. Kramvis, A., Kostaki, E. G., Hatzakis, A., & Paraskevis, D. (2018). Immunomodulatory Function of HBeAg Related to Short-Sighted Evolution, Transmissibility, and Clinical Manifestation of Hepatitis B Virus. Frontiers in microbiology, 9, 2521. doi:10.3389/fmicb.2018.02521
51. Kumar, N., Kulkarni, D. D., Lee, B., Kaushik, R., Bhatia, S., Sood, R., Pateriya, A. K., Bhat, S., & Singh, V. P. (2018). Evolution of codon usage bias in henipaviruses is governed by natural selection and is host-specific. *Viruses*, *10*(11). https://doi.org/10.3390/v10110604

52. Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096
53. Kurbanov, F., Y. Tanaka, M. Mizokami. (2010). Geographical and genetic diversity of the human hepatitis B virus. Hepatolog,y Research. 40(1):14-30
54. Lamontagne, R. J., Bagga, S., & Bouchard, M. J. (2016). Hepatitis B virus molecular biology and pathogenesis. Hepatoma Research, 2(7), 163. doi:10.20517/2394-5079.2016.05
55. Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531
56. Lauber, C., Seitz, S., Mattei, S., Suh, A., Beck, J., Herstein, J., … Bartenschlager, R. (2017). Deciphering the Origin and Evolution of Hepatitis B Viruses by Means of a Family of Non-enveloped Fish Viruses. Cell host & microbe, 22(3), 387–399.e6.
57. Li, M., Kao, E., Gao, X., Sandig, H., Limmer, K., Pavon-Eternod, M., Jones, T. E., Landry, S., Pan, T., Weitzman, M. D., & David, M. (2012). Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*, *491*(7422), 125–128. https://doi.org/10.1038/nature11433
58. Li, S., Wang, Z., Li, Y., & Ding, G. (2017). Adaptive evolution of proteins in hepatitis B virus during divergence of genotypes. *Scientific Reports*, *7*(1), 1990. https://doi.org/10.1038/s41598-017-02012-8
59. Liang T. J. (2009). Hepatitis B: the virus and disease. Hepatology (Baltimore, Md.), 49(5 Suppl), S13–S21. doi:10.1002/hep.22881
60. Liu, B.-M., Li, T., Xu, J., Li, X.-G., Dong, J.-P., Yan, P., Yang, J.-X., Yan, L., Gao, Z.-Y., Li, W.-P., Sun, X.-W., Wang, Y.-H., Jiao, X.-J., Hou, C.-S., & Zhuang, H. (2010). Characterization of potential antiviral resistance mutations in hepatitis B virus reverse transcriptase sequences in treatment-naïve Chinese patients. *Antiviral Research*, *85*(3), 512–519. https://doi.org/10.1016/j.antiviral.2009.12.006
61. Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

62. MacLachlan, J. H., & Cowie, B. C. (2015). Hepatitis B virus epidemiology. *Cold Spring Harbor Perspectives in Medicine*, *5*(5). https://doi.org/10.1101/cshperspect.a021410
63. Makałowska, I., Lin, C.-F., & Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. BMC Evolutionary Biology, 7(1), 193. doi:10.1186/1471-2148-7-193
64. McNaughton, A. L., D'Arienzo, V., Ansari, M. A., Lumley, S. F., Littlejohn, M., Revill, et al (2019). Insights From Deep Sequencing of the HBV Genome—Unique, Tiny, and Misunderstood. Gastroenterology. 156: 284-399.
65. Miyata, T., & Yasunaga, T. (1978). Evolution of overlapping genes. *Nature*, *272*(5653), 532–535. https://doi.org/10.1038/272532a0
66. Mizokami, M., Orito, E., Ohba, K. et al. (1997).Constrained evolution with respect to gene overlap of hepatitis B virus. Journal of Molecular Evolution. 44(Suppl 1): S83
67. Morley, V. J., & Turner, P. E. (2017). Dynamics of molecular evolution in RNA virus populations depend on sudden versus gradual environmental change. Evolution; international journal of organic evolution, 71(4), 872–883. doi:10.1111/evo.13193
68. Moya, A., Holmes, E. C., & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. In *Nature Reviews Microbiology* (Vol. 2, Issue 4, pp. 279–288). Nature Publishing Group. https://doi.org/10.1038/nrmicro863
69. Nasir, A., Kim, K. M., & Caetano-Anollés, G. (2012). Viral evolution: Primordial cellular origins and late adaptation to parasitism. Mobile genetic elements, 2(5), 247–252.
70. Nielsen, R. (2005). Molecular Signatures of Natural Selection. Annual Review of Genetics, 39(1), 197–218. doi:10.1146/annurev.genet.39.073003.112420
71. Nishijima N, Marusawa H, Ueda Y, Takahashi K, Nasu A, et al. (2012) Dynamics of Hepatitis B Virus Quasispecies in Association with Nucleos(t)ide Analogue Treatment Determined by Ultra-Deep Sequencing. PLOS ONE 7(4): e35052. https://doi.org/10.1371/journal.pone.0035052
72. Palanisamy, N., Osman, N., Ohnona, F., Xu, H.-T., Brenner, B., Mesplède, T., & Wainberg, M. A. (2017). Does antiretroviral treatment change HIV-1

codon usage patterns in its genes: a preliminary bioinformatics study. AIDS Research and Therapy, 14(1). doi:10.1186/s12981-016-0130-y

73. Pavesi A. (2006). Origin and evolution of overlapping genes in the family Microviridae. Journal of General Virology. 87(4): 1013-1017.
74. Pavesi A. (2015). Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution. Journal of General Virology. 96: 3577–3586.
75. Pavesi A, Magiorkinis G, Karlin DG (2013). Viral Proteins Originated De Novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses. PLOS Computational Biology 9(8): e1003162.
76. Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, et al. (2018) Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. PLOS ONE 13(10): e0202513
77. Pavesi, A. (2006). Origin and evolution of overlapping genes in the family Microviridae. *Journal of General Virology*, *87*(4), 1013–1017. https://doi.org/10.1099/vir.0.81375-0
78. Pavesi, A. (2015). Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution. *Journal of General Virology*, *96*(12), 3577–3586. https://doi.org/10.1099/jgv.0.000307
79. Pavesi, A. (2020). New insights into the evolutionary features of viral overlapping genes by discriminant analysis. Virology, 546, 51–66. doi:10.1016/j.virol.2020.03.007
80. Pavesi, A., Magiorkinis, G., & Karlin, D. G. (2013). Viral Proteins Originated De Novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses. *PLoS Computational Biology*, *9*(8). https://doi.org/10.1371/journal.pcbi.1003162
81. Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., & Karlin, D. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0202513
82. Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. In *Infection, Genetics and Evolution*

(Vol. 30, pp. 296–307). Elsevier. https://doi.org/10.1016/j.meegid.2014.12.022

83. Pickett, B. E., Liu, M., Sadat, E. L., Squires, R. B., Noronha, J. M., He, S. et al. (2013). Metadata-driven comparative analysis tool for sequences (meta-CATS): An automated process for identifying significant sequence variations that correlate with virus attributes. Virology, 447(1-2), 45–51.
84. Pickett, B., Greer, D., Zhang, Y., Stewart, L., Zhou, L., Sun, G., Scheuermann, R. (2012). Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community. Viruses, 4(11), 3209–3226.
85. Poon, A. F. Y., Frost, S. D. W., & Pond, S. L. K. (2009). Detecting signatures of selection from DNA sequences using datamonkey. *Methods in Molecular Biology*, *537*, 163–183. https://doi.org/10.1007/978-1-59745-251-9_8
86. Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biology direct*, *3*, 38. https://doi.org/10.1186/1745-6150-3-38
87. Pybus, O. G., & Shapiro, B. (2012). Natural selection and adaptation of molecular sequences. In P. Lemey (Ed.), *The Phylogenetic Handbook* (pp. 407–418). Cambridge University Press. https://doi.org/10.1017/cbo9780511819049.015
88. Qi, X., Wei, C., Li, Y., Wu, Y., Xu, H., Guo, R., Jia, Y., Li, Z., Wei, Z., Wang, W., Jia, J., Li, Y., Wang, A., & Gao, X. (2020). The characteristic of the synonymous codon usage and phylogenetic analysis of hepatitis B virus. *Genes & Genomics*, *42*(7), 805–815. https://doi.org/10.1007/s13258-020-00932-w
89. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R., & Karlin, D. (2009). Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *Journal of Virology*, *83*(20), 10719–10736. https://doi.org/10.1128/jvi.00595-09
90. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R., & Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. Journal of virology, 83(20), 10719–10736. doi:10.1128/JVI.00595-09
91. Revill, P. A., Tu, T., Netter, H. J., Yuen, L. K. W., Locarnini, S. A., & Littlejohn, M. (2020). The evolution and clinical impact of hepatitis B virus genome

diversity. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 17, Issue 10, pp. 618–634). Nature Research. https://doi.org/10.1038/s41575-020-0296-6

92. Revill, P. A., Tu, T., Netter, H. J., Yuen, L. K. W., Locarnini, S. A., & Littlejohn, M. (2020). The evolution and clinical impact of hepatitis B virus genome diversity. Nature Reviews Gastroenterology & Hepatology. doi:10.1038/s41575-020-0296-6
93. Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. In *Trends in Genetics* (Vol. 18, Issue 5, pp. 228–232). Trends Genet. https://doi.org/10.1016/S0168-9525(02)02649-5
94. Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. Trends in Genetics, 18(5), 228–232. doi:10.1016/s0168-9525(02)02649-5
95. Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. Trends in Genetics, 18(5), 228–232.
96. Roman, S., Tanaka, Y., Khan, A., Kurbanov, F., Kato, H., Mizokami, M., & Panduro, A. (2010). Occult hepatitis B in the genotype H-infected Nahuas and Huichol native Mexican population. *Journal of Medical Virology*, *82*(9), 1527–1536. https://doi.org/10.1002/jmv.21846
97. Roth, A., Anisimova, M., & M., C. G. (2012). Measuring codon usage bias. In S. A. Cannarozzi G. (Ed.), *Codon Evolution Mechanisms and Models* (pp. 189–215). Oxford University. https://www.researchgate.net/publication/230857012_Measuring_codon_usage_bias
98. Roth, Alexander, Anisimova, M., & Cannarozzi, G. M. (2012). Measuring codon usage bias. In *Codon Evolution: Mechanisms and Models*. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199601165.003.0013
99. Sabath, N., A. Wagner, D. Karlin. 2012. Evolution of Viral Proteins Originated De Novo by Overprinting, Molecular Biology and Evolution, Volume 29 (12): 3767–3780.

100. Saha, D., Panda, A., Podder, S., & Ghosh, T. C. (2014). Overlapping genes: a new strategy of thermophilic stress tolerance in prokaryotes. Extremophiles, 19(2), 345–353

101. Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. In *Cellular and Molecular Life Sciences* (Vol. 73, Issue 23, pp. 4433–4448). Birkhauser Verlag AG. https://doi.org/10.1007/s00018-016-2299-6

102. Sanjuán, R., & Domingo-Calap, P. (2019). Genetic Diversity and Evolution of Viral Populations. *Reference Module in Life Sciences*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7157443/?tool=EBI

103. Sanjuán, R., & Domingo-Calap, P. (2019). Genetic Diversity and Evolution of Viral Populations. Reference Module in Life Sciences, B978-0-12-809633-8.20958-8. https://doi.org/10.1016/B978-0-12-809633-8.20958-8

104. Sanna, C. R., Li, W.-H., & Zhang, L. (2008). Overlapping genes in the human and mouse genomes. BMC Genomics, 9(1), 169.

105. Seeger, C., Zoulim, F., & Mason, W. S. (2013). Hepadnaviruses. In D. M. Knipe & P. M. Howley (Eds.), *Fields Virology* (6th ed.). LIPPINCOTT WILLIAMS & WILKINS.

106. Shackelton, L. A., Parrish, C. R., & Holmes, E. C. (2006). Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *Journal of Molecular Evolution*, *62*(5), 551–563. https://doi.org/10.1007/s00239-005-0221-1

107. Sharp, P. M., Stenico, M., Peden, J. F., & Lloyd, A. T. (1993). Codon usage: Mutational bias, translational selection, or both? *Biochemical Society Transactions*, *21*(4), 835–841. https://doi.org/10.1042/bst0210835

108. Sharp, Paul M., & Li, W. H. (1986). Codon usage in regulatory genes in Escherichia coli does not reflect selection for “rare” codons. *Nucleic Acids Research*, *14*(19), 7737–7749. https://doi.org/10.1093/nar/14.19.7737

109. Sheikh, A., Al-Taher, A., Al-Nazawi, M., Al-Mubarak, A. I., & Kandeel, M. (2020). Analysis of preferred codon usage in the coronavirus N genes and their implications for genome evolution and vaccine design. *Journal of Virological Methods*, *277*, 113806. https://doi.org/10.1016/j.jviromet.2019.113806

110. Shukla, A. (2015). Analysis of overlapping reading frames in viral genomes. Dissertation. University of Lübeck. Germany. Pp. 120.

111. Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? In *Nature Reviews Microbiology* (Vol. 9, Issue 8, pp. 617–626). Nature Publishing Group. https://doi.org/10.1038/nrmicro2614

112. Simon-Loriere, E., Holmes, E. C., & Pagán, I. (2013). The effect of gene overlapping on the rate of RNA virus evolution. Molecular biology and evolution, 30(8), 1916–1928. https://doi.org/10.1093/molbev/mst094

113. Smith, N. G. C., Webster, M. T., & Ellegren, H. (2002). Deterministic mutation rate variation in the human genome. *Genome Research*, *12*(9), 1350–1356. https://doi.org/10.1101/gr.220502

114. Suh, A., Weber, C. C., Kehlmaier, C., Braun, E. L., Green, R. E., Fritz, U., Ray, D. A., & Ellegren, H. (2014). Early Mesozoic Coexistence of Amniotes and Hepadnaviridae. *PLoS Genetics*, *10*(12). https://doi.org/10.1371/journal.pgen.1004559

115. Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 15;22(12):1540–2

116. Taanman, J.-W. (1999). The mitochondrial genome: structure, transcription, translation and replication. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1410(2), 103–123.

117. Torres, C., M. D., Blanco-Fernández, D., Martín-Flichman, R. Héctor-Campos, V. Andrea-Mbayed. (2013). Influence of overlapping genes on the evolution of human hepatitis B virus. Virology. 44 (1): 40-48.

118. Torresi, J. (2002). The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. 25(2): 97-106.

119. Tort, F. L., Castells, M., & Cristina, J. (2020). A COMPREHENSIVE ANALYSIS OF GENOME COMPOSITION AND CODON USAGE PATTERNS OF EMERGING CORONAVIRUSES. Virus Research, 197976. doi:10.1016/j.virusres.2020.197976

120. Van de Klundert, M. A. A., Cremer, J., Kootstra, N. A., Boot, H. J., & Zaaijer, H. L. (2011). Comparison of the hepatitis B virus core, surface and polymerase gene substitution rates in chronically infected patients. Journal of Viral Hepatitis, 19(2), e34–e40. doi:10.1111/j.1365-2893.2011.01506.x

121. van der Kuyl, A. C., & Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus.

In *Retrovirology* (Vol. 9, p. 92). BioMed Central. https://doi.org/10.1186/1742-4690-9-92

122. van der Kuyl, A. C., & Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology*, *9*, 92. https://doi.org/10.1186/1742-4690-9-92

123. Velkov, S., Ott, J. J., Protzer, U., & Michler, T. (2018). The global hepatitis B virus genotype distribution approximated from available genotyping data. *Genes*, *9*(10). https://doi.org/10.3390/genes9100495

124. Watashi, K., Urban, S., Li, W., & Wakita, T. (2014). NTCP and beyond: Opening the door to unveil hepatitis B virus entry. In *International Journal of Molecular Sciences* (Vol. 15, Issue 2, pp. 2892–2905). Molecular Diversity Preservation International. https://doi.org/10.3390/ijms15022892

125. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191

126. Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., & Kosakovsky Pond, S. L. (2018). Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular biology and evolution*, *35*(3), 773–777. https://doi.org/10.1093/molbev/msx335

127. Wei X.,J. Zhang, 2015. A Simple Method for Estimating the Strength of Natural Selection on Overlapping Genes, Genome Biology and Evolution.7 (1): 381–390

128. Willis, S., & Masel, J. (2018). Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes. Genetics, genetics.301249.2018. doi:10.1534/genetics.118.301249

129. Wright, F. (1990). The “effective number of codons” used in a gene. *Gene*, *87*(1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

130. Wu, C., Deng, W., Deng, L., Cao, L., Qin, B., Li, S., Wang, Y., Pei, R., Yang, D., Lu, M., & Chen, X. (2012). Amino Acid Substitutions at Positions 122 and 145 of Hepatitis B Virus Surface Antigen (HBsAg) Determine the Antigenicity and Immunogenicity of HBsAg and Influence In Vivo HBsAg Clearance . *Journal of Virology*, *86*(8), 4658–4669. https://doi.org/10.1128/jvi.06353-11

131. Yang, Z., & Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution*, *17*(1), 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236

132. Zaaijer, H. L., van Hemert, F. J., Koppelman, M. H., & Lukashov, V. V. (2007). Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. Journal of General Virology, 88(8), 2137–2143.

133. Zhang, D., Chen, J., Deng, L., Mao, Q., Zheng, J., Wu, J., … Li, Y. (2010). Evolutionary selection associated with the multi-function of overlapping genes in the hepatitis B virus. Infection, Genetics and Evolution, 10(1), 84–88. doi:10.1016/j.meegid.2009.10.006

134. Zhang, Z. H., Wu, C. C., Chen, X. W., Li, X., Li, J., & Lu, M. J. (2016). Genetic variation of hepatitis B virus and its significance for pathogenesis. In *World Journal of Gastroenterology* (Vol. 22, Issue 1, pp. 126–144). Baishideng Publishing Group Co., Limited. https://doi.org/10.3748/wjg.v22.i1.126

12. ANEXOS

12.1. Contenido de GC en los diferentes ORFs y genotipos

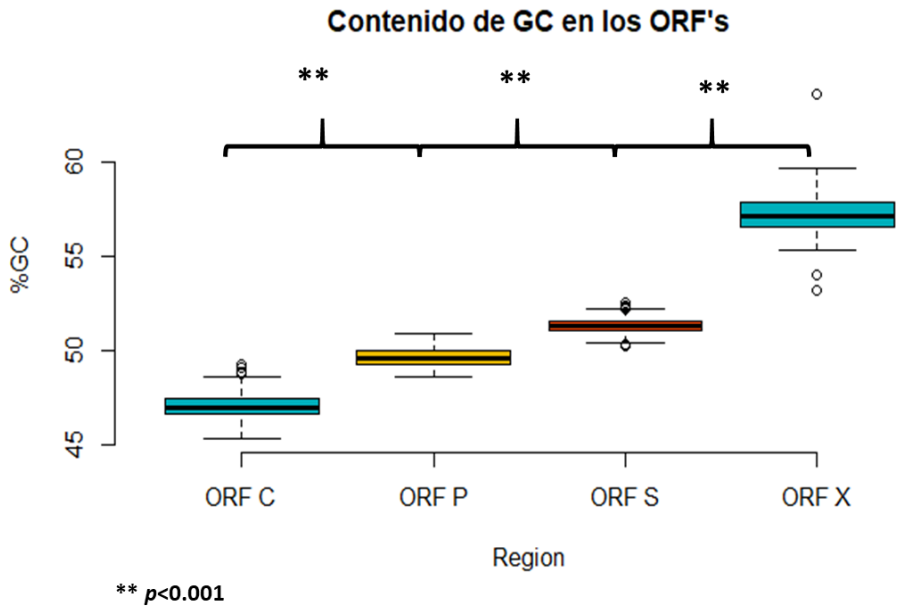


Figura 26. Contenido de GC (%) de los cuatro ORFs del genoma del HBV: C, P, S y X. ** p < 0.05.

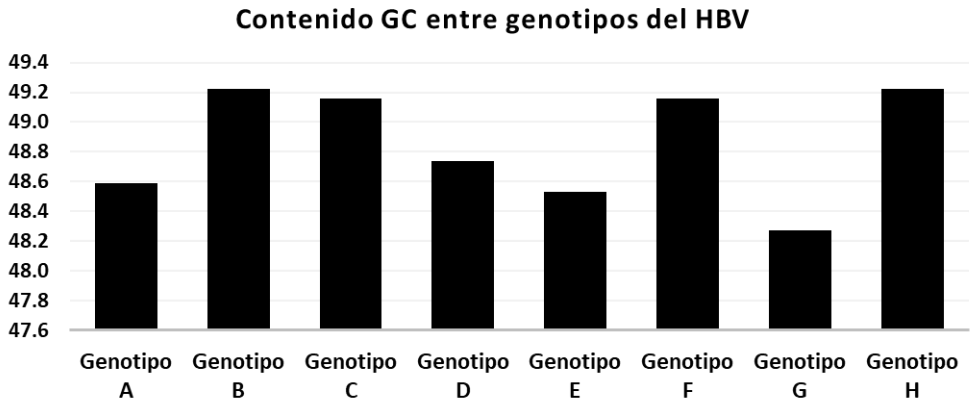


Figura 27. Contenido de GC (%) de los genomas de los principales genotipos del HBV.

12.2. Composición de aminoácidos en las diferentes regiones de los ORFs en función del nivel de degeneración de los codones que los codifican

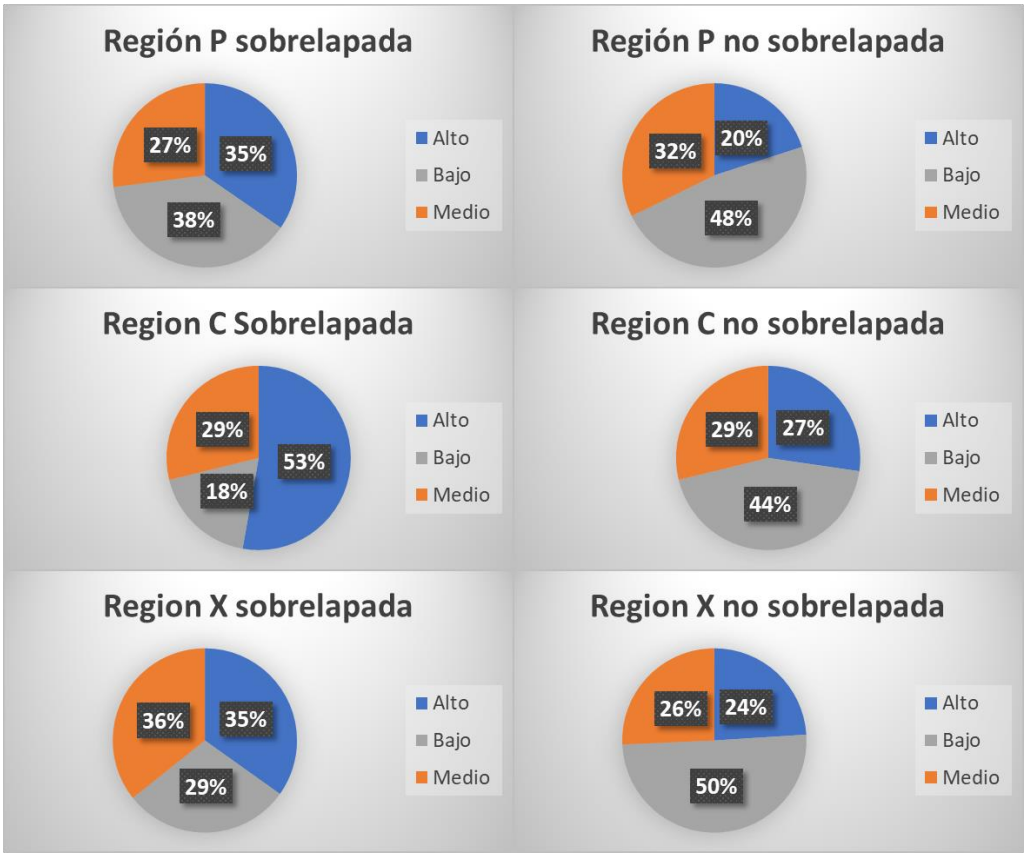


Figura 28. Composición de aminoácidos en las diferentes regiones de los ORFs en función del nivel de degeneración de los codones que los codifican.

12.3. Métricas de diversidad entre regiones y genotipos

Tabla 5. Valores de las métricas de diversas en la región sobrelapada y no sobrelapada de cada genotipo del HBV.

	Pi		**Shannon**		**No total de mutaciones (Eta)**		**No. sitios polimórficos (S)**	
Genotipo	Región no sobrelapada	Región soprelapada	Región no sobrelapada	Región sobrelapada	Región no sobrelapada	Región sobrelapada	Región no sobrelapada	Región sobrelapada
A	0.03917	0.02245	0.10267016	0.09765047	0.477	0.074	0.343	0.058
B	0.0436	0.02	0.113	0.063	0.504	0.357	0.353	0.282
C	0.04552	0.02597	0.13371509	0.09235767	0.569	0.493	0.393	0.351
D	0.04959	0.02383	0.1206221	0.06716156	0.619	0.421	0.427	0.320
E	0.02266	0.01281	0.07241808	0.04055702	0.448	0.333	0.331	0.267
F	0.02429	0.01403	0.08883114	0.04098872	0.392	0.331	0.294	0.266
G	0.01469	0.0106	0.04039227	0.02068233	0.128	0.091	0.123	0.087
H	0.0356	0.01928	0.06803327	0.03812445	0.235	0.134	0.198	0.126
Promedio	**0.03439**	**0.01862125**	**0.09246026**	**0.05756528**	**0.421**	**0.279**	**0.308**	**0.220**
SD	0.01248855	0.00557187	0.03103865	0.02736347	0.166	0.159	0.101	0.112

12.4. Tasa de sustitución entre en el dominio RT de la ORF P

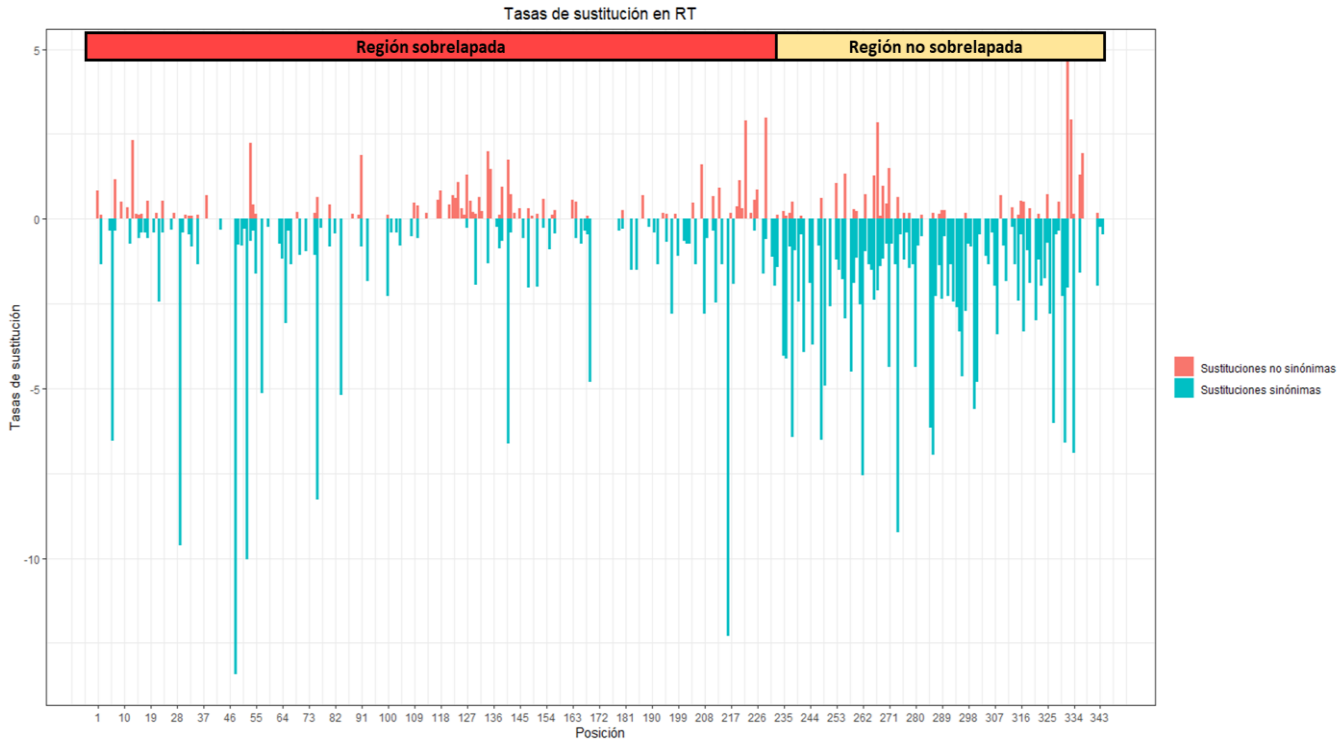


Figura 30. Tasa de sustitución entre en el dominio RT de la ORF P. Sustituciones no sinónimas en rojo y sustituciones sinónimas en azul. Arriba de la gráfica representada la región sobrelapada y no sobrelapada de la región que codifica para la RT de la polimerasa.

12.5. Patrones de sustitución nucleotídica en el ORF C y ORF X de HBV

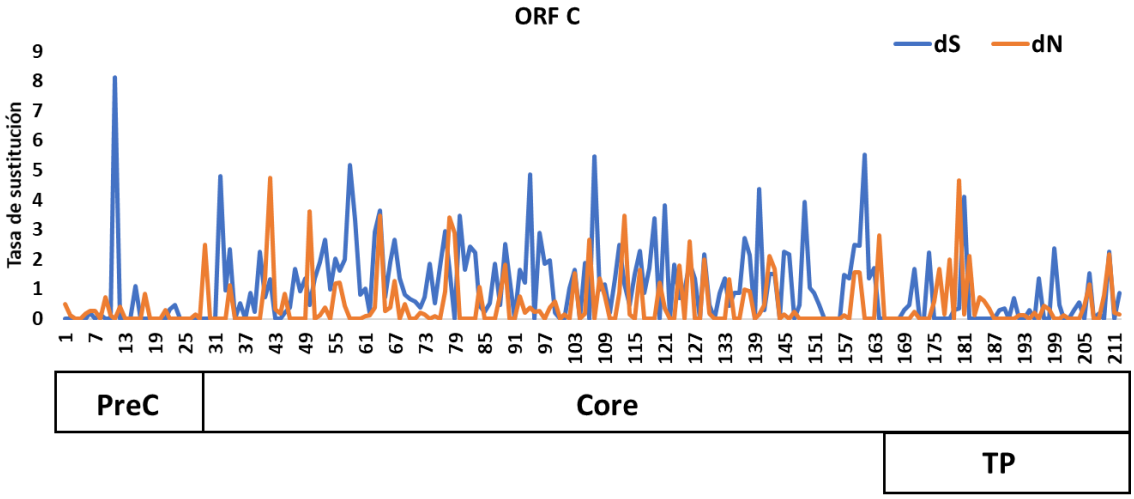


Figura 31. Patrones de sustitución en el sobrelape del ORF C y el ORF P. dS= Sustituciones sinónimas, dN = Sustituciones no sinónimas. Regiones del ORF C representadas, PreC y Core; TP =Terminal Protein, del ORF P.

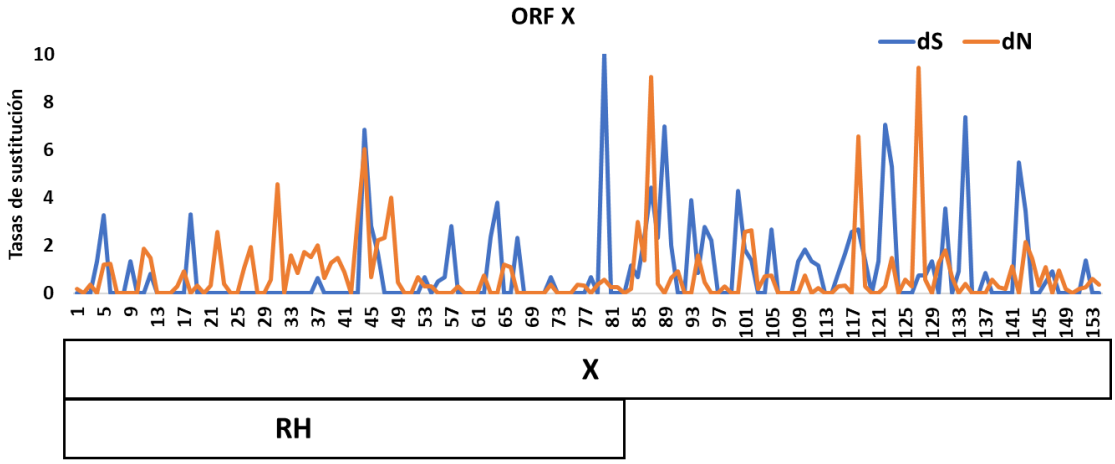


Figura 32. Patrones de sustitución en el sobrepe del ORF S y el ORF P. dS= Sustituciones sinónimas, dN = Sustituciones no sinónimas. ORF P representado; Región RH del ORF P.