Universitat Autònoma de Barcelona

Estudio de la quasiespecies del virus de la hepatitis B a través de secuenciación masiva para la identificación de dianas terapéuticas y factores pronósticos

Marçal Yll Picó

Universitat Autònoma de Barcelona

Programa de doctorado en Medicina

Departamento de Medicina

Tesis doctoral

Estudio de la quasiespecies del virus de la hepatitis B a través de secuenciación masiva para la identificación de dianas terapéuticas y factores pronósticos

Autor de la tesis

Marçal Yll Picó

Directores de la tesis

Dr. Francisco Rodríguez Frías Dra. María Asunción Buti Ferret Dra. Maria Francesca Cortese

Tutora de la tesis

Dra. María Asunción Buti Ferret

Universitat Autònoma de Barcelona. Barcelona, 2020

AGRADECIMIENTOS

Recuerdo perfectamente el día en el que me entrevisté con el Dr. Francisco Rodríguez Frías para hacer mis prácticas de máster con su grupo de investigación. Desde un primer momento ambos estábamos interesados en que ese proyecto inicial fuera evolucionando para acabar dando lugar a una tesis doctoral. Hoy, pasados 3 años y medio desde entonces, podemos decir que hemos cumplido el objetivo.

Esto no hubiera sido posible de no haber tenido la ayuda y asesoramiento que siempre he recibido por parte de mis directores de tesis: el ya mencionado Dr. Francisco Rodríguez Frías, la Dra. Maria Buti Ferret (también tutora de esta tesis) y la Dra. Maria Francesca Cortese. Con sus amplios conocimientos en la materia han ido guiando mis pasos, indicándome el camino a seguir cuando me encontraba un obstáculo, siendo los principales artífices del desarrollo y la progresión de mi carrera científica. Un ambiente laboral favorable en el que se te ofrezcan tanto buenos consejos como momentos es imprescindible para trabajar con una actitud positiva, lo que permite un aprendizaje fácil, cómodo y eficaz. Esto ha sido posible gracias a mis compañeros y compañeras de laboratorio: Sara, Irene, Cristina, Marta, Chari, David, Carol, Gerard y mucha otra gente que ha puesto su granito de arena en este proyecto, muchas gracias por todo. Por lo que hace a la bioinformática, sin la ayuda de Mercedes Guerrero y del Dr. Josep Gregori no sé cuánto hubiera tardado en analizar los datos obtenidos, por lo que su ayuda ha sido también imprescindible. No quisiera dejar de mencionar al Dr. Josep Quer, pues él como coordinador del máster del VHIR fue quien me asesoró en aquel entonces y me puso en contacto con este grupo de investigación con el que finalmente he llevado a cabo mi doctorado.

En un ámbito más personal, quisiera agradecer el apoyo y los ánimos que he recibido por parte de todos y cada uno de los miembros de mi familia. Especialmente quiero agradecerle a mi madre Mercè, a mi padre Joan y a mi hermana Gisela el apoyo incondicional que me han mostrado y los consejos que me han dado en esta etapa de mi vida, pues considero que gracias a ellos he crecido mucho como persona, aprendiendo algo nuevo cada vez que les pedía su opinión. La inteligencia y la fuerza mental de mi madre me han servido de referencia e inspiración para gestionar momentos difíciles, pues ella siempre ha sabido hacerme ver el lado bueno de las cosas y me ha cogido de la mano para sacarme del vaso de agua si me estaba ahogando en él. La actitud positiva y la determinación de mi padre me han enseñado que no hay camino difícil sino viajero cobarde y que la suerte se la busca uno mismo. El

carácter de mi hermana me ha ayudado a ser una persona fuerte y a saber diferenciar lo importante de lo banal, al mismo tiempo que me ha animado y me ha hecho reír siempre que lo he necesitado. Aunque esto no se reduce a este proyecto, sino que la actitud, los valores y los principios que he aprendido de estas tres personas a lo largo de mi vida me han hecho ser quien soy y me siento muy afortunado de haber podido crecer con ellas a mi lado.

No quisiera acabar estos agradecimientos sin mencionar a mi círculo de amistades, pues mis amigos y amigas son un punto de apoyo fundamental para mí. Siempre he considerado que en esta vida hay que trabajar duro para lograr tus objetivos, pero esto, al menos en mi caso, no sería posible sin los buenos momentos que he pasado a su lado y que me han permitido reírme y distraerme para volver al trabajo con todavía más fuerza y determinación. Quiero en primer lugar nombrar a mi amigo de la infancia, Pablo, pues, aunque ya no vivamos en el mismo sitio y nos veamos poco sigue estando muy presente en mi día a día y tenemos aquella confianza que solo los buenos amigos tienen. Como no, nombrar a mi amigo Dani que, aunque no nos conozcamos desde pequeños, el que nos parezcamos tanto en tantas cosas hizo que nos lleváramos increíblemente bien desde un principio y que hayamos compartido un sinfín de vivencias juntos. A Arnau no lo he querido nombrar hasta ahora ya que aparte de incluirlo entre mis mejores amigos (por esas largas charlas en las que argumentamos y debatimos sobre temas, a veces serios, a veces totalmente irrelevantes y sin sentido), también quisiera meterlo en el saco de mis compañeros de piso. Junto a él, Giovanni y J. Torras han sido personas clave a lo largo de este proyecto. Me han ayudado siempre en lo que hiciera falta y me han animado en momentos complicados de esta etapa. No quisiera acabar este párrafo sin nombrar a Pol, mi amigo de la universidad (en la que éramos uña y carne). Aunque hay que decir que han sido escasas, las veces que hemos podido vernos y hablar me han recordado muy buenos momentos de aquella época. Pol es una magnífica persona y un excelente científico, y aunque sé que no la necesita porque le espera un futuro brillante, le deseo toda la suerte del mundo en su carrera.

A parte de las personas que he mencionado, hay muchas otras a las que no he nombrado en estos agradecimientos y que han contribuido directa o indirectamente a que esta tesis doctoral fuera posible. Tanto a unos como a otros, fuera cual fuera vuestra aportación a mi vida tanto profesional como personal, MUCHAS GRACIAS.

ABREVIACIONES

A

A (nt)	Adenina
aa	Aminoácido
ADN	Ácido desoxirribonucleico
ADNccc	ADN circular covalentemente cerrado
ADNrc	ADN relajado circular
Ag	Antígeno
AGL	Loop antigénico presente en pre-S1
ALT	Alanina aminotransferasa
AN	Análogos de nuceós(t)idos
anti-HBc	Anticuerpos contra el antígeno core del VHB
anti-HBe	Anticuerpos contra el antígeno e del VHB
anti-HBs	Anticuerpos contra el antígeno de superficie del VHB
anti-PD1	Inhibidor del receptor de muerte celular programada 1
APOBEC	Apoliprotein B mRNA Editing enzyme, Catalytic polypeptide-like
ARN	Ácido ribonucleico
ARNm	ARN mensajero
ARNpc	ARN precore
ARNpg	ARN pregenómico
AST	Aspartato aminotransferasa
A.1	Amplicón 1
A.2	Amplicón 2

B

BCP	Basal core promoter

C

C (nt)	Citosina
C	Región core del genoma del VHB
CBP	CREB binding protein
CHB	Hepatitis crónica B / Pacientes con hepatitis crónica B sin daño hepático
CI	Contenido de información
Cp	Promotor preCore/pregenómico
CpAMs	Core protein Allosteric Modulators
CRE	Cyclic-AMP-regulated enhancer
CREB	Cyclic-AMP-response element binding protein
CRS	Señal de retención citoplasmática
CTD	Dominio carboxi-terminal de la proteína HBc
Cys	Cisteína

D

Del	Deleción
DR1/DR2	Repeticiones directas 1 y 2

E

EB	Elution buffer
Enh I/Enh II	Enhancer 1 y 2
ESCRT	Endesomal sorting complexes required for transport
ETV	Entecavir

F

F (aa)	Fenilalanina

G

G (nt)	Guanina
G (aa)	Glicina

H

HBc	Proteína core
HBC	Gen core
HBcAg	Antígeno core
HBcrAg	Antígeno core-related
HBeAg	Antígeno e
HBsAg	Antigeno de superficie
HBx	Proteína X
HBX	Gen X
HCC	Carcinoma hepatocelular / Pacientes con carcinoma hepatocelular
H_{GS}	Índice Gini-Simpson
Hpl	Haplotipo
H_{SN}	Entropía de Shanon
HSP	Heat shock protein
HSPG	Heparán sulfato

I

Ins	Inserción
InsDel	Inserción/Deleción
IQR	Rango intercuartil

K

Kb	Kilobase
KDa	Kilodalton

L

LC	Pacientes con cirrosis hepática
LHBsAg	Proteína de superficie grande

M

Mf	Frecuencia de mutación
Mfm	Average mutation frequency by molecule
MHBsAg	Proteína de superficie mediana
MHR	Major hydrophilic region
Min	Minutos
MIR	Major immunodominant region
mL	Mililitros
MVB	Cuerpos multivesiculares

N

NAP	Ácidos nucleicos fosfotioados
NF-Kb	Factor de transcripción de las cadenas ligeras kappa de las células B
NGS	Next generation sequencing
NPC	Poro nuclear
NLS	Señal de localización nuclear
nt	Nucleótido
NTCP	Péptido co-transportador de sodio-taurocolato
NTD	Dominio amino-terminal de la proteína HBc
NXF1	Factor de exportación nuclear 1

O

ORF	Open Reading frame

P

P (aa)	Prolina
P	Proteína polimerasa / Región P del genoma del VHB
P5/P7	Primers adaptadores de la secuenciación por MiSeq Illumina
P79Q	Sustitución aminoacídica de prolina por glutamina en la posición 79
pb	Pares de bases
PC	Proteína preCore
PCR	Polymerase chain reaction
PD1	Receptor de muerte celular programada 1
PKC	Proteína quinasa C
PreC/Core	Región preCore/core del genoma del VHB
Pt	Paciente

Q

Q (aa)	Glutamina
qPCR	Quantitative polymerase chain reaction
QS	Quasiespecie

R

R (aa)	Arginina
RB	Región bisagra de la proteína HBc
RH	Dominio ribonucleasa H de la polimerasa del VHB
RISC	Complejo de Silenciamiento Inducido por ARN
Rpm	Revoluciones por minutos
RT	Retrotranscriptasa

S

S	Región S del VHB
SBS	Secuenciación por síntesis
SHBsAg	Proteína de superficie pequeña
siRNA	small interference RNA

T

T (nt)	Timina
TAF	Tenofovir Alafenamide
TDF	Tenefovir
TLR	Toll-like receptor
TP	Dominio terminal protein de la polimerasa del VHB
TREX	Maquinaria de transcripción-exportación

U

UI	Unidades internacionales

V

VHB	Virus de la hepatitis B
VHC	Virus de la hepatitis C
VHD	Virus de la hepatitis D
VIH	Virus de la Inmunodeficiencia Humana

X

X	Región X del genoma del VHB

Y

Y (aa)	Tirosina

ÍNDICE

RESUMEN

A pesar de tener una vacuna preventiva eficaz, el virus de la hepatitis B (VHB) es un grave problema de salud mundial debido a su elevada prevalencia (afecta a más de 250 millones de personas en todo el mundo) y a la morbilidad que deriva de su cronicidad, pues es el primer factor virológico de riesgo en el desarrollo de carcinoma hepatocelular (HCC). La proteína HBc (codificada por el gen *HBC*) es imprescindible para el virus. Se autoensambla formando la cápside viral y gracias a su amplia red de interacciones interviene en múltiples procesos del ciclo viral. Aunque el tratamiento disponible actualmente permite controlar la replicación viral, hoy en día no es posible erradicar la infección, por lo que se necesitan nuevas estrategias terapéuticas. Además, considerando la gravedad de la evolución clínica de la enfermedad, la identificación de factores virológicos que pudieran pronosticar la progresión del daño hepático sería de gran ayuda en el seguimiento de los pacientes. En este proyecto de tesis doctoral se ha analizado, a través de secuenciación masiva, la conservación y la complejidad de la qusiespecies (QS) del VHB en la región del gen *HBC* en pacientes con hepatitis crónica B en diferentes estados de la enfermedad hepática. Con ello se han querido detectar, tanto regiones hiperconservadas (independientemente del cuadro clínico o el genotipo viral de los pacientes) que pudieran servir de posibles dianas para nuevas estrategias terapéuticas y/o diagnósticas como diferencias en términos de conservación, mutaciones y complejidad de la QS entre los distintos cuadros clínicos analizados que pudieran servir de factores pronósticos del avance de la enfermedad.

El genoma del VHB se extrajo a partir de muestras de suero de pacientes con hepatitis crónica por VHB en diferentes etapas clínicas de la enfermedad y la región *HBC* del genoma viral se amplificó, a través de un sistema de amplificación de tres pasos secuenciales, dividido en dos amplicones. Seguidamente estos se secuenciaron por Next Generation Sequencing (NGS) a través de la plataforma MiSeq Illumina. Una vez hecho el genotipado de la población viral, la presencia de mutaciones se detectó alineando las secuencias obtenidas para cada paciente con una secuencia consenso del genotipo correspondiente. La conservación de la QS, tanto global como por grupo, se analizó calculando el contenido de información. También se analizaron diversos indicadores de complejidad de la QS.

Se detectaron regiones (tanto nucleotídicas como aminoacídicas) hiperconservadas independientemente del cuadro clínico y del genotipo viral cuya elevada conservación podría evidenciar su importancia funcional, por lo que podrían ser valiosas dianas de terapia y

diagnosis. Además, se evidenciaron regiones conservadas específicamente en ciertos grupos clínicos, lo que sugiere un posible rol de estas regiones en la progresión de la enfermedad hepática. En los dos estudios realizados se detectó una sustitución aminoacídica (P79Q) en los pacientes con lesión tumoral (HCC). En estos mismos pacientes se detectó una elevada complejidad de la QS en la región de uno de los dos amplicones en los que se dividió el gen *HBC*. La elevada complejidad de la QS y la detección de la sustitución P79Q en los pacientes HCC podrían ser factores pronósticos de la transformación tumoral. Posteriores estudios serán necesarios para analizar con más profundidad la posible asociación entre estos factores y la lesión hepática.

En resumen, los resultados obtenidos podrían servir como base para el desarrollo de estrategias panclínicas y pangenotípicas de tratamiento y diagnosis de la hepatitis crónica por VHB, así como para la identificación de factores pronósticos que puedan ayudar en el seguimiento de la enfermedad hepática.

ABSTRACT

Despite having an effective preventive vaccine, the hepatitis B virus (HBV) is a serious global health problem due to its high prevalence (it affects more than 250 million people worldwide) and the morbidity derived from its chronicity, as it is the first virological risk factor in the development of hepatocellular carcinoma (HCC). The HBc protein (encoded by the *HBC* gene) is essential for the virus. It self-assembles forming the viral capsid and thanks to its wide network of interactions it intervenes in multiple steps of the viral cycle. Although the present therapeutic protocol allows to adequately control the viral replication, the eradication of the infection is not achievable, for which reason new therapeutic strategies are required. Moreover, considering the severity of the disease progression, the identification of viral prognostic factors could be extremely helpful in patients follow-up. In this doctoral thesis we have analysed, through massive sequencing, the conservation and complexity of the quasispecies (QS) of the *HBC* gene of HBV in patients with chronic hepatitis B at different clinical stages in order to detect both hyperconserved regions (regardless of the clinical stage or the viral genotype) that could serve as possible targets for new therapy and/or diagnostic strategies, and group-specific differences in terms of conservation, mutations and complexity of the QS that could serve as prognostic factors for the disease progression.

HBV genome was extracted from serum samples of patients with chronic hepatitis B at different clinical stages and the *HBC* region was amplified, using a three-steps amplification protocol, divided in two amplicons. The amplicons were later sequenced by Next Genetation Sequencing (NGS) using the MiSeq Illumina platform. Once the viral population was genotyped, the presence of mutations was detected by aligning the sequences obtained for each patient with a consensus sequence of the corresponding genotype. QS conservation, both general and group-related, was studied by calculating the information content. QS complexity was analysed by considering different indexes.

We detected some nucleotide and amino acid hyper-conserved regions (regardless of the clinical stage and the viral genotype) that could be used as therapeutic or diagnostic targets. Moreover, some group-specific conservation patterns, that could cover a role in disease progression, were observed. In both studies of the project an amino acid substitution (P79Q) was detected in patients with tumoral injury (HCC). The patients in this clinical stage showed a high QS complexity in one of the *HBC* amplicons. The P79Q mutation and the high

complexity in HCC patients could be used as prognostic factors of tumoral transformation. However, further studies are required to deeply clarify the possible association between these viral factors and the liver injury.

In summary, the results here obtained could serve as a basis for the development of panclinical and pangenotypic strategies for the treatment and diagnosis of chronic hepatitis B as well as for the identification of prognostic factors that could be helpful in liver disease progression follow-up.

INTRODUCCIÓN

1. INTRODUCCIÓN

1.1 Virus de la Hepatitis B

El Virus de la Hepatitis B (VHB) es un patógeno viral con tropismo específico para las células hepáticas y su infección crónica es la primera causa en el mundo de carcinoma hepatocelular (HCC) debido a infecciones virales.

Desde 1982 se dispone de una vacuna preventiva (eficacia del 95-99% (1)), lo que ha permitido una reducción de la incidencia y prevalencia del VHB en ciertas áreas endémicas (2). No obstante, la organización mundial de la salud reporta que la infección crónica por VHB afecta a 257 millones de personas en todo el mundo, llegando a causar hasta 887.000 muertes en 2015 (3). Esta infección es una importante causa de morbilidad. De hecho, el riesgo de desarrollar complicaciones hepáticas es elevado, pues causa el 30% de las cirrosis y el 53% de los HCC de todo el mundo (4).

1.1.1 Historia

En 1965 Blumberg y colaboradores identificaron un nuevo antígeno en la sangre de un aborigen australiano (5). Cuatro años después se relacionó con la hepatitis (6) y finalmente se supo que correspondía al antígeno de superficie del VHB (HBsAg) (7). En 1970 Dane y colaboradores visualizaron a través de microscopía electrónica las partículas virales del VHB ("partículas de Dane") (8) pero se tardaron alrededor de 10 años más en secuenciar totalmente el genoma viral (9) y en determinar la estructura y replicación del virus (10).

1.1.2 Taxonomía

El VHB pertenece al género *Orthohepadnavirus*. Este se incluye dentro de la familia *Hepadnaviridae*, una familia de virus de ácido desoxirribonucleico (ADN) que infectan el hígado. Se han detectado virus similares en otros organismos animales, tanto en mamíferos como en aves (11). Todos ellos tienen tropismo exclusivo para el hígado, poseen una organización genómica muy parecida y comparten una estrategia de replicación única en la que se da un paso de transcripción inversa de un ARN mensajero del virus (11).

1.1.3 Epidemiología y prevalencia

El VHB está presente en fluidos como la sangre, la saliva, el semen, las secreciones vaginales y la orina de las personas infectadas, por lo que se transmite en los adultos principalmente por vía parenteral (a través de agujas o de productos sanguíneos contaminados) y por vía sexual (12). En las zonas donde la infección es endémica, el virus se transmite mayoritariamente en la edad neonatal e infantil por transmisión vertical o perinatal. El VHB pueda permanecer estable hasta 7 días en la superficie de un material inerte, por lo que el contacto con una superficie contaminada debido a malas prácticas como compartir material de higiene personal o material sanitario puede comportar su transmisión horizontal (12).

El VHB se distribuye heterogéneamente a nivel global, con zonas geográficas donde su infección es endémica (Figura 1). Estas zonas incluyen África central y subsahariana, el este de Asia, islas del Pacífico, parte de la región de los Balcanes, la cuenca del Amazonas y en el norte de América, donde la infección afecta más del 8% de la población (en color azul oscuro en la Figura 1). En estas zonas la infección se adquiere mayormente por transmisión vertical o perinatal. Las zonas con prevalencia más baja (menos del 2% de la población) son las regiones tropicales i del centro de América latina, América del Norte y el oeste de Europa (2) (en azul más claro en la Figura 1). En estas regiones el virus se suele adquirir a edades adultas por transmisión parenteral debido a la inoculación de drogas por vía intravenosa o a prácticas sexuales de riesgo (13).

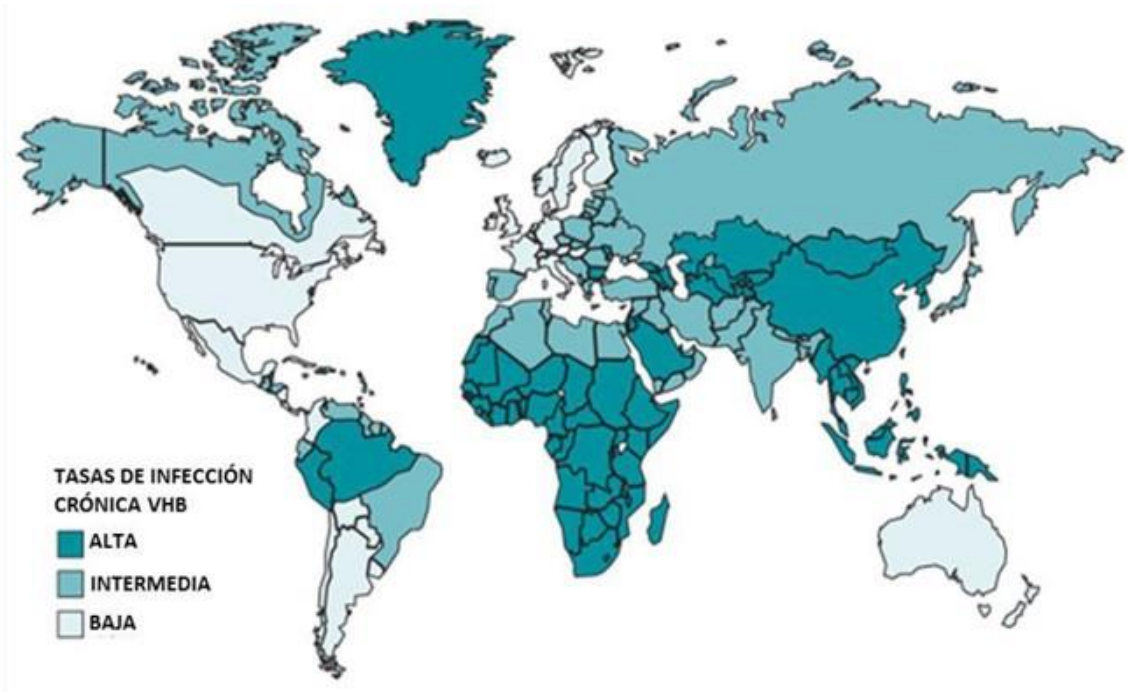


Figura 1: Prevalencia global de la infección crónica por VHB. Imagen modificada de Papastergiou V et al (4).

1.1.4 Genotipos del VHB

La elevada tasa de evolución del VHB ha causado la aparición y fijación diferencial de ciertas mutaciones en distintas poblaciones y zonas geográficas del mundo determinando así diferentes genotipos que se clasifican filogenéticamente en función de su divergencia genética (más del 7,5% en el genoma viral completo). Existen 9 genotipos (A-I) y un supuesto décimo genotipo (J) que ha sido aislado en solo una persona (14). Los genotipos se dividen y clasifican en subgenotipos cuando la diferencia intragrupal (entre los distintos genomas de un mismo genotipo) es del 4-8%. Se han detectado al menos 35 subgenotipos distintos (15).

Los distintos genotipos virales presentan una distribución geográfica concreta (como se muestra en la Figura 2). Los genotipos A y D, por ejemplo, son más típicos de poblaciones caucásicas, mientras que el genotipo C afecta mayormente a poblaciones asiáticas.

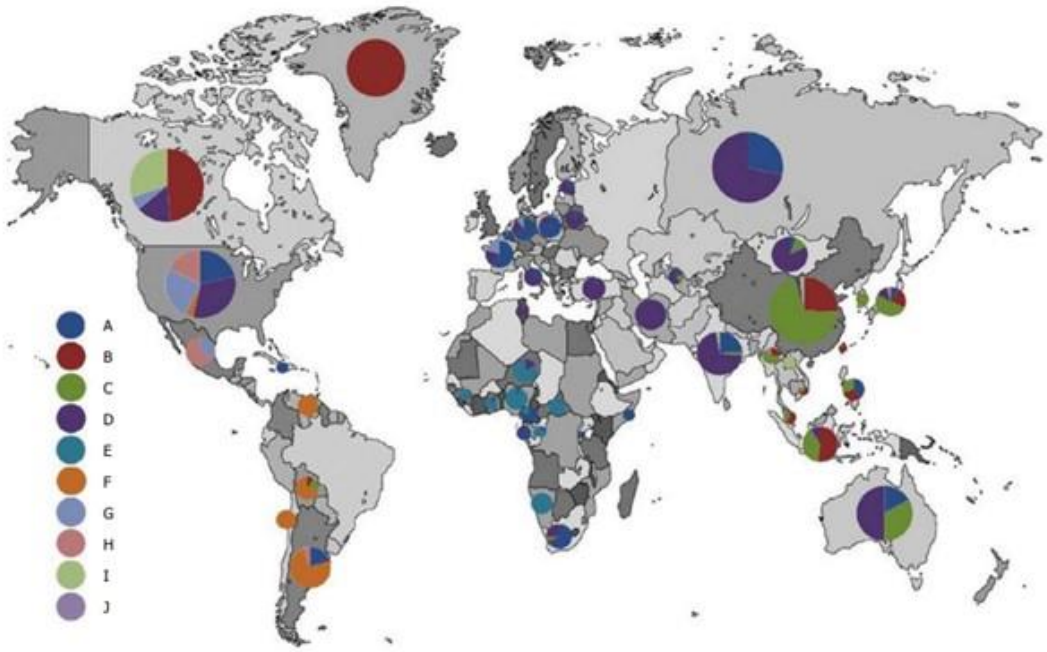


Figura 2: Distribución geográfica de los genotipos del VHB. Imagen extraída de Sunbul M (16).

Conocer los distintos genotipos es importante ya que se asocian a diferentes evoluciones clínicas como la progresión de la enfermedad o la respuesta terapéutica (16,17). Por ejemplo, los genotipos A y B muestran una mejor respuesta al tratamiento con interferón, seroconversión más temprana y mayor probabilidad de remisión sostenida, con menor actividad de la enfermedad a nivel histológico que los genotipos D y C que están más asociados a hepatitis severas con elevada probabilidad de desarrollar cirrosis y HCC (18,19).

1.2 Características clínicas de la infección por VHB

1.2.1 Marcadores virológicos en el estudio clínico de la infección por VHB

Existen diferentes marcadores virológicos para el estudio de la infección por VHB. El antígeno core (HBcAg, correspondiente a la proteína core o HBc) es uno de los primeros marcadores de la infección. A este le acompañan el ADN viral, el antígeno de superficie (HBsAg) y el antígeno e (HBeAg). Este último se considera un marcador de replicación viral. Los distintos marcadores son esenciales para definir correctamente el estado de la infección y para trazar una estrategia de tratamiento adecuada. A estos marcadores clásicos, se les han ido añadiendo otros como el antígeno *core-related* (HBcrAg) o el ARNpg circulante.

El ARNpg se usa como marcador de la actividad transcripcional del ADNccc en el hígado y desde que se determinó que el ARN del VHB en sangre periférica es ARNpg en 2016 ha habido un número creciente de estudios sobre el tema (20,21). Los niveles de ARNpg en suero pueden predecir la historia natural de los pacientes con hepatitis crónica por VHB (22). El antígeno *core-related* consiste en la detección por quimioluminiscencia de las proteínas HBc, HBeAg y p22 del VHB. Este se considera un marcador prometedor en pacientes con baja carga viral (como pacientes tratados) (23) y pacientes con infección crónica HBeAg-negativos (24).

1.2.2 Historia natural de la infección

1.2.2.1 Infección aguda

La infección aguda por el VHB se caracteriza, generalmente, por un cuadro o sintomático leve (astenia, cefaleas, náuseas, febrícula y malestar general entre otras) o del todo asintomático, siendo muy inusuales los cuadros severos que desarrollan una hepatitis fulminante (1% de los casos de hepatitis aguda ictérica) (25). La respuesta inmune que se da tras la infección genera una respuesta inflamatoria que se autolimita en más del 95% de adultos inmunocompetentes (26). El tiempo de incubación dura entre 30 y 180 días. El HBsAg y el ADN del VHB (pudiendo llegar a niveles muy elevados) son los primeros marcadores virológicos en aparecer, seguidos por HBeAg. A medida que los anticuerpos

empiezan a actuar, los marcadores virológicos empiezan a descender. Las ALT y AST (indicadores de daño hepático) aumentan al iniciarse la replicación viral como efecto del daño citotóxico ocasionado por la respuesta inflamatoria que se ha generado (27) (Figura 3).

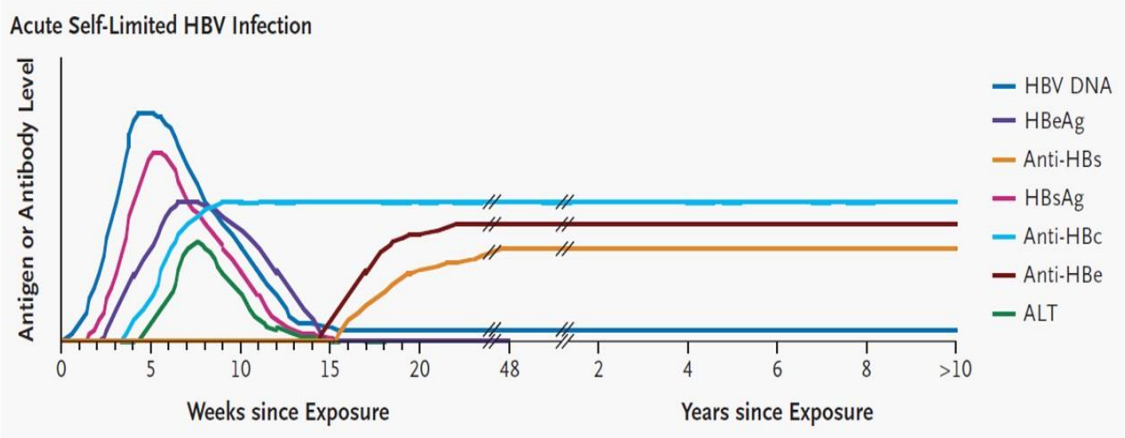


Figura 3: Marcadores serológicos de la infección aguda por VHB. La figura muestra la evolución de los diferentes marcadores, tanto virológicos como bioquímicos, que caracterizan la infección aguda por el VHB. El ADN (en azul oscuro) y el HBsAg (en rosa) son los primeros marcadores en aparecer, al poco tiempo de la infección, mientras que el HBeAg (lilla) es más tardío. Las ALT (en verde) empiezan a aumentar una vez que la replicación viral llega a su pico (alrededor de la semana 5). A medida que aumenta la respuesta de anticuerpos (en azul claro los anticuerpos anti-HBc, en granate los anti-HBe y en amarillo los anti-HBs), los niveles de los marcadores empiezan a descender. Imagen extraída de Ganem D et al (27).

La infección aguda por VHB se suele resolver a las 4-8 semanas sin necesidad de terapia. El fallo hepático agudo ocurre muy raramente y es favorecido por sobre o coinfecciones con el VHD o por un daño hepático previo (28).

En un porcentaje muy limitado de pacientes que se infectan en edad adulta y en la mayoría de los pacientes que se infectan por vía perinatal, la infección aguda deja paso a una infección crónica (29).

1.2.2.2 Infección crónica

La infección crónica por VHB se define por la persistencia de HBsAg en suero por más de seis meses (26). Los síntomas son diversos, pudiendo variar desde inespecíficos hasta una fibrosis hepática que progresivamente puede dar lugar a descompensación hepática o HCC como resultado independientemente de si se ha pasado por un estado de cirrosis (30).

El cuadro clínico de infección crónica perdura durante años y se caracteriza por la persistencia de los antígenos virales (HBsAg y en algunos casos del HBeAg), por un nivel de ADN viral detectable (desde niveles bajos a carga virales muy altas) y por la presencia de anti-HBc (indicativo de la pasada fase aguda) (Figura 4).

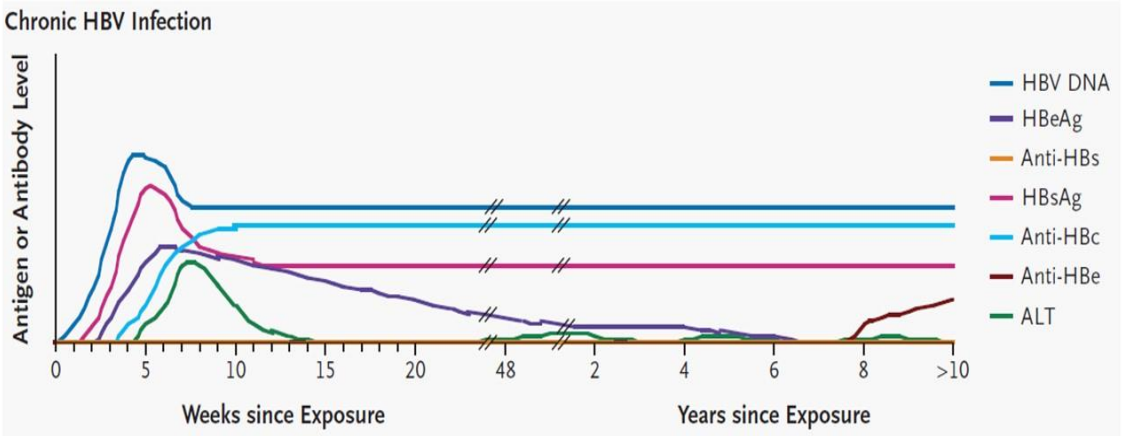
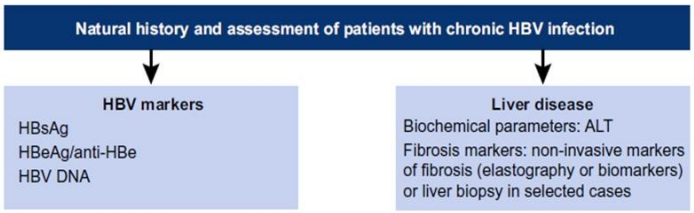


Figura 4: Marcadores serológicos de la infección crónica por VHB. La figura muestra los niveles de los distintos marcadores virológicos y bioquímicos en el caso de infección aguda que progresa a infección crónica. Los niveles de ADN viral y HBsAg se mantienen altos durante años después de la infección y resolución de la fase aguda. Los niveles de HBeAg disminuyen en cuanto aumentan los niveles de anticuerpos anti-HBe. Imagen extraída de Ganem D et al (27).

En función de los marcadores se diferencian distintos cuadros clínicos. Dependiendo de los niveles de ALT, se distinguen la infección crónica (no hay evidencias de necroinflamación, niveles de ALT normales) de la hepatitis crónica, la cual puede ser HBeAg positiva o negativa (Figura 5).



	HBeAg positive		HBeAg negative	
	Chronic infection	Chronic hepatitis	Chronic infection	Chronic hepatitis
HBsAg	High	High/intermediate	Low	Intermediate
HBeAg	Positive	Positive	Negative	Negative
HBV DNA	$>10^7$ IU/ml	10^4-10^7 IU/ml	<2,000 IU/ml°°	>2,000 IU/ml
ALT	Normal	Elevated	Normal	Elevated*
Liver disease	None/minimal	Moderate/severe	None	Moderate/severe
Old terminology	Immune tolerant	Immune reactive HBeAg positive	Inactive carrier	HBeAg negative chronic hepatitis

Figura 5: Fases de la historia natural y clasificación de pacientes infectados crónicamente por VHB basada en marcadores de enfermedad hepática y del VHB según las nuevas guías de tratamiento clínico de la EASL. Imagen extraída de Lampertico P et al (31).

A continuación, se detallan las características virológicas y bioquímicas de los distintos cuadros clínicos:

- Infección crónica por VHB HBeAg positiva (antigua fase de Inmunotolerancia): Se caracteriza por la presencia de HBeAg y viremia elevada, sin evidencias de necroinflamación ni de fibrosis (o fibrosis ínfima). El avance de la enfermedad hepática en esta etapa es mínimo a pesar de la alta viremia (26). No obstante, el genoma viral puede integrarse en el ADN del hepatocito aumentando el riesgo de padecer HCC en el futuro (29).

- Hepatitis crónica por VHB HBeAg positiva: En esta fase los niveles de ADN del VHB suelen ser altos pero variables, los niveles de ALT aumentan y la histología hepática revela necroinflamación con etapas variables de fibrosis (32). La activación de la respuesta inmune en muchos casos causa la seroconversión de HBeAg y un agravamiento de la necroinflamación y la fibrosis, por lo que una durada prolongada de esta fase incrementa el riesgo de desarrollar cirrosis y HCC (29).

- Infección crónica por VHB HBeAg negativo (antigua fase de portador inactivo): Los niveles de ADN viral son muy bajos (<2000 UI/mL) o indetectables y los niveles de ALT son normales. Esta fase tiene un buen pronóstico y el 1-3% de los caso llegan incluso a perder espontáneamente el HBsAg (26,29).

- Hepatitis crónica por VHB HBeAg negativo: se caracteriza por una viremia superior a 2000 UI/mL y niveles de ALT elevados o fluctuantes. En esta fase de infección activa la enfermedad hepática progresa a nivel histológico y hay una mayor probabilidad de descompensación hepática pudiendo ocasionar cirrosis y HCC (29,33).

1.2.3 Hepatocarcinoma y VHB

El HCC es una de las principales causas de muerte por cáncer en todo el mundo y el VHB juega un papel muy importante en esta enfermedad hepática severa, pues causa el 50-80% de todos los casos de HCC (34).

En un hígado dañado la deposición de tejido fibrótico en la cirrosis promueve la expansión clonal de los hepatocitos formando nódulos. Esta elevada tasa de proliferación celular podría promover los primeros eventos de transformación asociados a la carcinogénesis. No obstante, el HCC no siempre sucede a la cirrosis, pues se puede desarrollar en hígados no cirróticos y por lo tanto el nódulo tumoral, altamente replicativo, no siempre surge de un nódulo regenerativo preexistente (35).

El VHB puede promover la carcinogénesis de forma directa o indirecta (36). Este no es un virus citopático, y los daños observados en el tejido hepático de un paciente de VHB se asocian principalmente a la infiltración de células inmunes (37). Considerando la alta tasa de regeneración del tejido hepático, la proliferación de los hepatocitos derivada de este daño hepático podría favorecer la aparición de modificaciones preneoplásicas que aumenten el riesgo de carcinogénesis. El VHB también puede promover directamente la transformación celular mediante la integración del genoma viral en el genoma del huésped (38). Más del 80% de los casos de HCC asociados a VHB presentan esta integración genómica, siendo uno de los principales factores que desencadenan la carcinogénesis (39).

Así pues, la infección crónica por VHB promueve un estado de inflamación hepática que causa daños e irregularidades en los hepatocitos, ocasionando una regeneración descontrolada del tejido como consecuencia que favorecerá la aparición de nódulos fibróticos. Este estado de inflamación crónica sumado a la activa replicación celular puede favorecer una inestabilidad genómica que potencie procesos de carcinogénesis en estos nódulos, dando lugar al HCC.

1.3 VHB: características estructurales y genéticas.

Aunque el VHB sea uno de los virus más pequeños en términos de tamaño y de longitud del genoma, sus características estructurales y su complejidad genética hacen que sea un virus muy peculiar entre los virus a ADN.

Es un virus de 40-42 nm de diámetro (Figura 6), dotado de envuelta lipídica que recubre una nucleocápside proteica icosaédrica formada por 180-240 moléculas de la proteína HBc (HBcAg) (40,41). La envuelta consta de 3 tipos de proteínas de superficie que se diferencian entre ellas por el tamaño (*small* o SHBsAg, *medium* o MHBsAg y *large* o LHBsAg). Las 3 tienen la misma porción C-terminal, pero difieren en el tamaño de la región N-terminal (42). Dentro de la cápside proteica se encuentra el genoma viral, la polimerasa viral y otras proteínas del huésped que participan en el ciclo replicativo (como por ejemplo isoformas específicas de la proteína quinasa C (PKC)) (43).

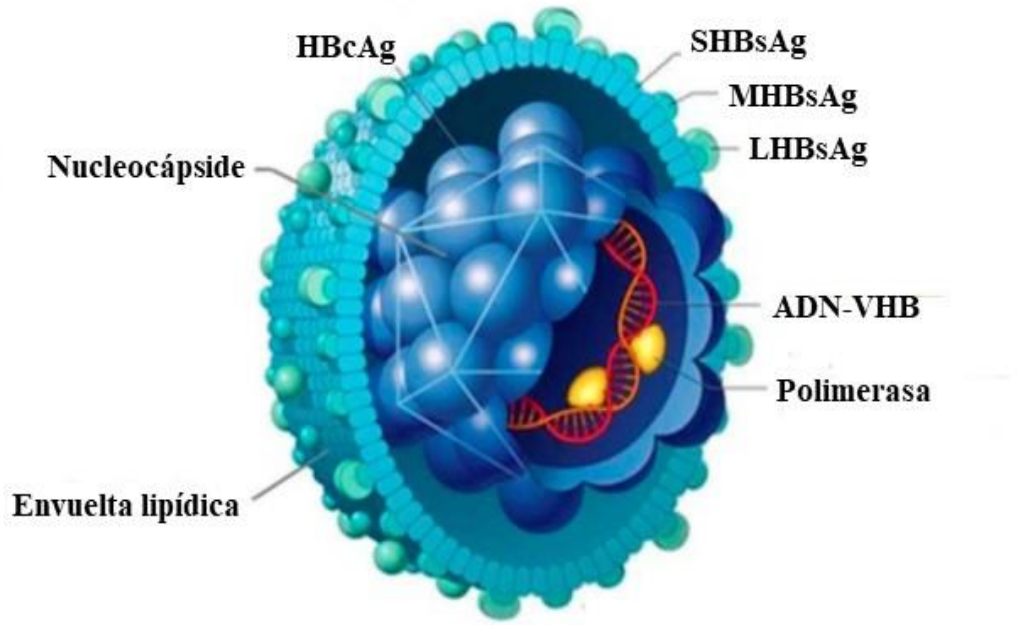


Figura 6: Composición de la partícula viral del VHB circulante en sangre. Se reportan sus componentes principales. Imagen extraída y modificada de © 2002 James A. Perkins, Medical and Scientific illustrations (http://people.rit.edu/japfaa/infectious.html).

1.3.1 Organización del genoma del VHB

Pese a su reducida longitud, el genoma del VHB codifica diferentes proteínas funcionales y estructurales. Este genoma se compone de una molécula de ADN de doble cadena parcialmente relajado, circular pero no cerrado covalentemente (ADNrc, ADN relajado circular) de 3,2 Kb (Figura 7). La cadena negativa (-ADN) contiene la secuencia completa del ADN viral mientras que la cadena positiva (+ADN) abarca 2/3 del tamaño del genoma. Al final de las dos cadenas hay dos secuencias cortas de 11 nucleótidos (nt) llamadas DR1 y DR2 (Repeticiones Directas), codificadas respectivamente por la cadena negativa y positiva, que mantienen la configuración circular del ADNrc por complementariedad y son esenciales para la replicación viral (44).

El ADN del VHB consta de 4 promotores (promotor preS1, promotor S, promotor preCore/pregenómico (Cp), que incluye el promotor basal del Core o BCP y el promotor X), 2 elementos de transcripción (*enhancing*, Enh I y Enh II) y 4 marcos de lectura abiertos (*Open reading frame,* ORF: P, S, PreC/Core y X) que se solapan entre sí (45) (Figura 7).

De la transcripción de estos 4 ORFs por la ARN polimerasa II celular se forman 4 ARN mensajeros (ARNm) virales subgenómicos (preS1, preS2/S, preCore (ARNpc) y X) que se traducen a las tres proteínas de superficie (SHBsAg, MHBsAg y LHBsAg), al antígeno "e" (HBeAg) y a la proteína X (HBx) y un ARNm pregenómico (ARNpg) que puede servir de molde para la síntesis de nuevas moléculas de ADNrc o bien traducirse a las proteínas HBc y polimerasa viral (P) (45,46).

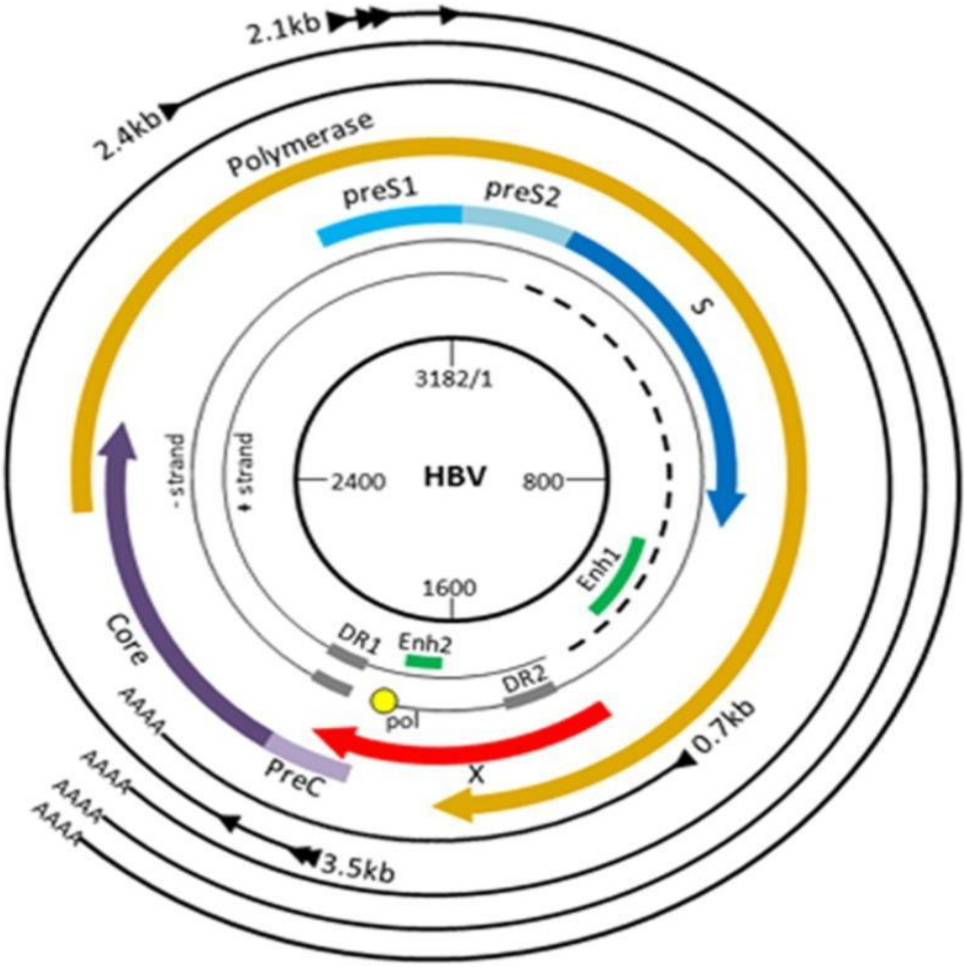


Figura 7: Organización del genoma del VHB. En la figura se muestran los 4 ORFs (representados en diferentes colores) y se aprecia el alto grado de solapamiento entre ellos. También se observa la posición de diferentes elementos como las repeticiones directas (DR1 y DR2) y los *enhancers* (Enh I y Enh II). Se reportan también los distintos ARN mensajeros que se forman durante la replicación del VHB: el ARN X de 0,7 kb, el ARN preS1 de 2,4 kb, el ARN preS2/S de 2,1 kb y los ARN preCore y pregenómico de 3,5 kb (ambos se transcriben bajo el control del mismo promotor, el preCore/pregenómico). Imagen extraída de Minor MM et al (47).

El ORF P abarca alrededor del 70% del genoma y codifica la polimerasa viral. Esta polimerasa incluye las funciones de transcriptasa inversa (retrotranscriptasa, RT), ADN polimerasa ADN dependiente (para la síntesis de la cadena positiva) y actividad RNAsa H (RH, para la eliminación del ARN usado como molde). Todas estas actividades enzimáticas se encuentran localizadas en lo que se denomina el dominio P. El otro dominio, dominio TP o *terminal protein* permite, a través de un residuo conservado de tirosina, la unión al ARN viral dando lugar al inicio de la síntesis de la cadena negativa de ADN por un proceso de retrotranscripción (48).

El ORF S codifica las tres proteínas de superficie de la envuelta. La proteína de superficie L (*Large,* LHBsAg) se traduce a partir de las regiones preS1, preS2 y S. La proteína de superficie M (*Medium,* MHBsAg) consta de las regiones preS2 y S, mientras la proteína de superficie S (*Small,* SHBsAg) se obtiene a partir de la región S solamente. Las tres proteínas de superficie contienen la región principal hidrofílica MHR (*Major hydrophilic region*, aa 99-169 de la secuencia de SHBsAg) donde se localiza el "determinante a" (aminoácidos (aa) 124-147) que es el principal dominio antigénico usado como diana para la respuesta neutralizante de anticuerpos de células B tras la infección o vacunación (49). El ORF S se solapa completamente con el ORF P y por lo tanto una mutación en uno de estos genes se puede ver reflejada en el otro (50).

El ORF más pequeño es el X, que codifica la proteína multifuncional y pleiotrópica X (HBx). Se solapa en gran parte con el dominio RH del ORF P y con el BCP del ORF PreC/Core. En su secuencia también encontramos la región DR2 y parte de los *enhancers* Enh I y Enh II. HBx es una proteína intracelular que tiene una gran capacidad transactivante e interfiere en una gran variedad de funciones celulares, por lo que tiene un papel importante en el desarrollo de HCC (51).

El ORF PreC/Core es objeto de estudio de este proyecto de tesis doctoral y sus características se detallan a continuación.

1.3.2 ORF PreC/Core

El ORF PreC/Core codifica las proteínas HBc y HBeAg. Está parcialmente solapado en sus extremos con los ORFs X y P, su transcripción se da bajo el control del promotor preCore/pregenómico (Cp, nt 1613-1849). Gracias a este promotor se producen dos ARN mensajeros de 3,5kb: el ARNpg y el ARNpc. Este ORF consta, de hecho, de dos codones de inicio de traducción en la misma pauta de lectura. El primero se encuentra en la posición nucleotídica 1814 del genoma del VHB y el segundo en la posición nucleotídica 1901.
Si la transcripción empieza en la posición 1901 se producirá HBc (también llamada p21) a partir de la traducción del ARNpg. Si la transcripción empieza en la posición 1841 se producirá la proteína preCore (PC o p25) a partir de la traducción del ARNpc. Esta es el

precursor inicial de HBeAg (p17), una proteína viral con función inmunomoduladora (52–54). El HBeAg se produce a partir de la proteína p25 como resultado de eventos de proteólisis llevados a cabo por furin proteasas en el lumen del retículo endoplasmático rugoso que formarán una molécula proteica intermedia llamada p22, que será el precursor final de HBeAg (55) (Figura 8).

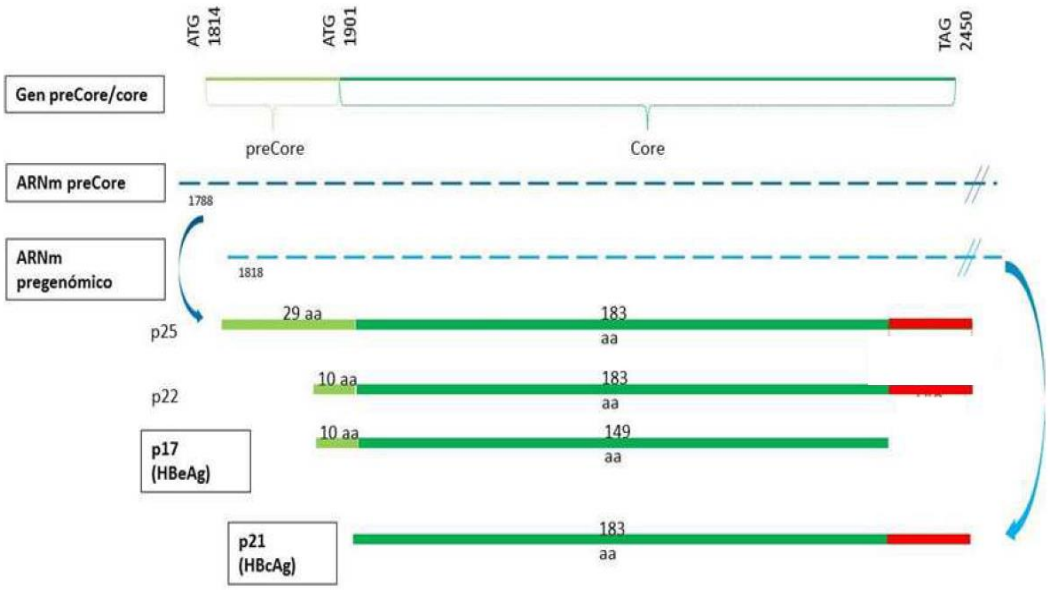


Figura 8: Traducción de las proteínas HBc (HBcAg) y HBeAg a partir del ORF PreC/Core. A partir del ORF PreC/Core se producen dos ARN mensajeros: el ARNm preCore y el ARNpg. Este último codifica por HBc (o p21), mientras que el ARNm preCore codifica la proteína p25, que gracias a la acción de las furin proteasa celulares, dará origen a la proteína p22 y finalmente al HBeAg (p17). Imagen extraída de: Caballero A (56).

1.4 Ciclo replicativo del VHB

El VHB entra en los hepatocitos gracias a un mecanismo de doble interacción (Figura 9, punto A). Primero, los residuos de lisina y arginina de un *loop* antigénico (AGL) presente en pre-S1 permiten la interacción electrostática de la partícula viral con HSPG, un proteoglicano muy abundante en la membrana plasmática de los hepatocitos (57). Este acoplamiento facilita la interacción entre los aa 2-48 del preS1 con el receptor específico NTCP (péptido co-transportador de sodio-taurocolato) en la membrana basolateral de los hepatocitos. Las partículas virales entonces son traspasadas al interior celular. Se han propuesto dos vías de entrada: por fusión de la envuelta viral con la membrana plasmática o por endocitosis (hipótesis más aceptada) (58).

Una vez liberadas en el citoplasma las partículas virales se dirigen a los complejos del poro nuclear (NPC) a través de microtúbulos gracias a un mecanismo de transporte activo (59) (Figura 9, punto B). Este complejo proteico atraviesa la doble membrana nuclear y permite el intercambio de moléculas entre el núcleo y el citoplasma a través del reconocimiento de un dominio NLS (señal de importación nuclear). La proteína HBc presenta una secuencia NLS que permite a la cápside atravesar el NPC hasta la canasta nuclear (estructura en forma de jaula situada en la cara nuclear del NPC).

Seguidamente (Figura 9, punto C), las cápsides se desacoplan y las proteínas HBc y el genoma viral se liberan en el carioplasma. En el carioplasma, las enzimas de reparación celular completan la cadena positiva del ADNrc, los extremos se ligan y, gracias a la interacción con proteínas virales (HBc) y celulares (histonas), se forma el ADNccc (molécula de ADN circular covalentemente cerrada de doble cadena) (60). Esta es una molécula de ADN estable que asume una estructura similar a un minicromosoma y cuya expresión está regulada por modificaciones epigenéticas (61). Se estima que se pueden acumular hasta 50 moléculas de ADNccc en una misma célula infectada (62). Esta acumulación de material genético viral en el núcleo de la célula huésped es el principal obstáculo para la erradicación de la infección.

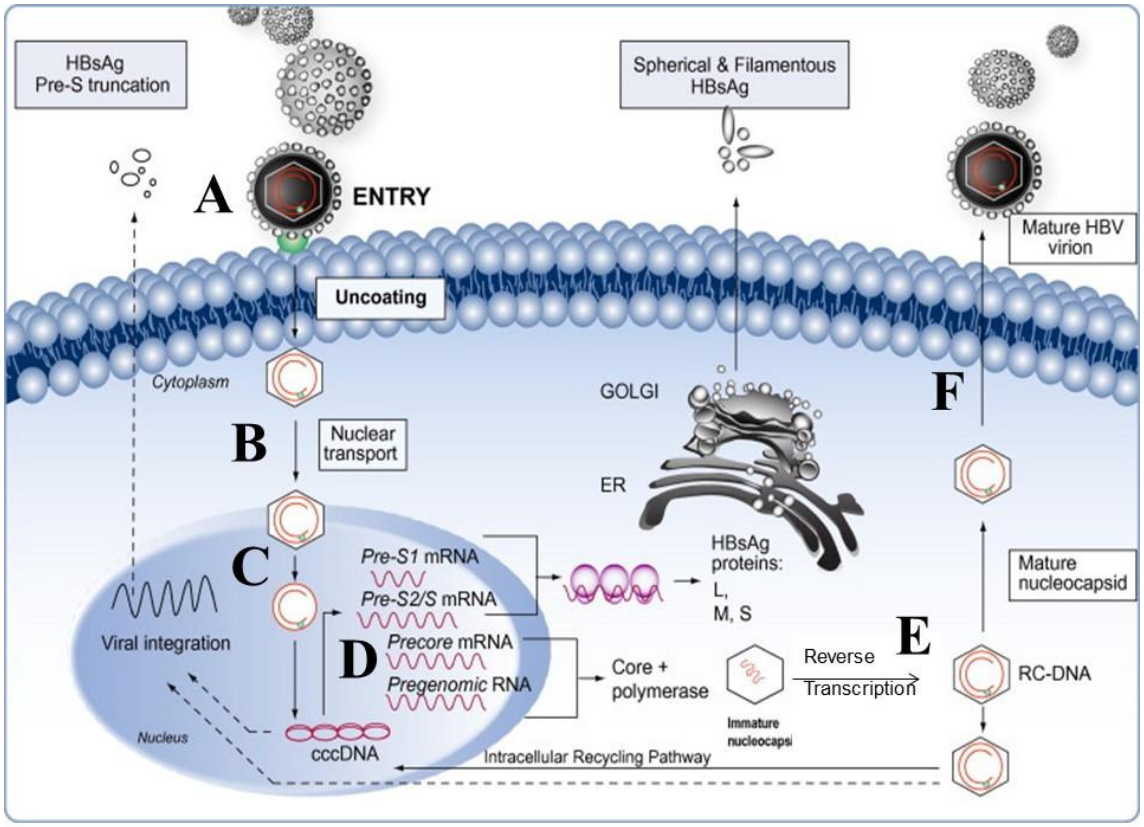


Figura 9: Esquema del ciclo replicativo del VHB. Los diferentes pasos explicados en el apartado están representados por letras (A: entrada, B: transporte nuclear, C: desacoplamiento de las cápsides, liberación en el carioplasma y reparación y formación del ADNccc, D: producción de los ARN mensajeros, E: retrotranscripción, F: adquisición de envuelta y secreción). Imagen modificada de Chan HLY et al (63).

La ARN polimerasa II del huésped transcribe entonces el ADNccc produciendo los distintos ARN mensajeros (Figura 9, punto D). Como se ha explicado anteriormente, el ARNpg codifica las proteínas HBc y polimerasa viral y sirve de molde para la síntesis de nuevas moléculas de ADNrc mediante un proceso de retrotranscipción (Figura 9, punto E). Este proceso empieza con la unión de la polimerasa del VHB a la señal de encapsidación ε del extremo 5' del ARNpg, que desencadenará el co-empaquetamiento del complejo ribonucleoproteico (formado por la polimerasa viral y el ARNpg) en la cápside viral de nueva formación (64). Las nucleocápsides van madurando gracias a pasos secuenciales de fosforilación y desfosforilación de los residuos fosforilables del dominio C-terminal de las proteínas HBc, en los que su desfosforilación parcial final será necesaria para una óptima encapsidación y formación del ADNrc (65). Gracias a la función RT de la enzima polimerasa viral se sintetiza, primero, la hebra negativa del ADNrc a partir del ARNpg. Seguidamente la función RNAsa H de la enzima eliminará el ARN usado como molde, dejando solo un

pequeño extremo en 5’ que la actividad polimerasa de la enzima usará como *primer* para la síntesis de la hebra positiva del ADNrc (66).
Las nucleocápsides ya maduras pueden adoptar la envuelta para formar nuevas partículas virales pasando por los cuerpos multivesiculares (MVB), que generan vesículas intraluminales que desde el citosol irán hasta la membrana plasmática donde las partículas virales adquieren la envuelta lipoproteica (Figura 9, punto F). El proceso de formación de vesículas intraluminales está a su vez mediado por la maquinaria del ESCRT (complejo de clasificación endosómica necesario para el transporte) (67). Finalmente, las nuevas partículas virales infecciosas son liberadas de al torrente sanguíneo.

Hay otras características a tener en cuenta cuando se habla del ciclo replicativo del VHB. Las partículas neoformadas, en vez de adquirir la envuelta y ser liberadas al torrente sanguíneo, pueden volver al núcleo donde liberan el ADNrc aumentando así el pool de moléculas de ADNccc acumuladas en el núcleo (proceso llamado reciclaje nuclear) (68). Además, el genoma del VHB una vez en el núcleo puede integrarse en el genoma de la célula huésped. Este fenómeno se asocia particularmente con el desarrollo de HCC y se ha detectado en más del 80% de estos pacientes (39). Otra característica a tener en cuenta es el exceso en la producción de proteínas de superficie, que dan forma a unas partículas llamadas partículas subvirales (filamentosa o esféricas) que también se secretan al torrente sanguíneo como las partículas virales infecciosas pero lo hacen en una proporción mucho mayor (69).

1.5 Tratamiento contra el VHB

Las guías europeas recomiendan el tratamiento en todos aquellos pacientes con hepatitis tanto HBeAg positiva como negativa con viremia >2.000 UI/mL, ALT elevadas y signos de daño hepático (70). Hay dos estrategias de tratamiento principales.

El primer protocolo terapéutico se basa en la administración de análogos de nucleós(t)idos (AN) que inhiben la retrotrascripción viral permitiendo una rápida caída de la carga viral. Entre estos se encuentran como medicamentos de primera línea Entecavir (ETV), Tenofovir (TDF) y Tenofovir Alafenamide (TAF), que se caracterizan por una alta barrera genética, es

decir, menor probabilidad de que el virus adquiera mutaciones asociadas a resistencia (70). A pesar de su elevada eficacidad, se necesita un periodo de tratamiento muy largo, con un riesgo relativamente importante de reactivación de la infección si se interrumpe la terapia. Concretamente, casi todos los pacientes con niveles de ADN del VHB superiores a 3 log copias/mL sufren una reactivación de la infección aproximadamente al año de interrumpirla mientras que esta reactivación se da en el 70% de los pacientes con niveles de ADN viral inferiores al mencionado (71).

La segunda estrategia terapéutica consiste en la administración de interferón α pegilado para que se instaure un control inmunológico de la infección en un periodo de tiempo relativamente corto. No obstante, este protocolo de tratamiento presenta ciertos límites debido a sus reacciones adversas y a la imprevisibilidad de la respuesta virológica (72).

A pesar de que el tratamiento contra el VHB permite controlar adecuadamente la infección, su erradicación hoy en día no es alcanzable. Este hecho es debido a la persistencia del material genético viral (ADNccc) en el núcleo de los hepatocitos infectados. Por lo tanto, la comunidad científica ha establecido como objetivo terapéutico la cura funcional, es decir, el aclaramiento serológico de HBsAg. No obstante, con los tratamientos de los que se disponen hoy en día, solamente el 3-7% de pacientes alcanzan esta cura funcional (73). Por esta razón la comunidad científica está interesada en desarrollar nuevas estrategias terapéuticas, tanto tratamientos inmunomoduladores como moléculas antivirales de acción directa.

1.5.1 Nuevos tratamientos

La infección por VHB causa una alteración de la respuesta inmunitaria que se caracteriza por la inducción de la respuesta innata, una activación de citoquinas muy pobre y la presencia de células B y T disfuncionales (74). Por estos motivos, entre las nuevas estrategia de tratamiento en fase de desarrollo se encuentran agonistas de los receptores *toll-like receptors* (TLR), responsables de reconocer motivos de patógenos y así activar la respuesta de interferones (75) o el tratamiento con *checkpoint inhibitors*, como un inhibidor del PD-1 (un receptor coinhibitorio involucrado en la muerte programada de las células linfocitarias). Un estudio publicado recientemente muestra como el tratamiento en fase I con anti-PD-1 en pacientes con supresión virológica se asocia a un descenso del nivel de HBsAg (76).

Por otro lado, las nuevas estrategias terapéuticas incluyen también antivirales de acción directa (Figura 10). Entre estos cabe destacar el Myrcludex-B, un antagonista competitivo del receptor NTCP (75) o los moduladores del ensamblaje de proteínas HBc (*Core protein Allosteric Modulators*: CpAMs) que interaccionan con los dímeros de HBc alterando la formación de la cápside, inhibiendo la encapsidación del ARNpg y consecuentemente su retrotrascripción (77).

Otras tácticas consisten en promover la degradación del ADNccc (78), o su silenciamiento epigenético (por ejemplo modificando su estado de metilación o modificando las histonas que forman parte del su estructura) (79).

El tratamiento con polímeros de ácidos nucleicos fosfotioados (NAP) disminuyen la secreción de HBsAg mediante interacciones hidrófobas con la superficie del antígeno, evitando los cambios conformacionales o posibles interacciones entre hélices necesarias para su secreción al exterior celular (80). También se ha sugerido la posible efectividad de la inhibición de la actividad RNAsa H de la polimerasa viral para el tratamiento contra el VHB (81).

El silenciamiento génico a través de ARN de interferencia (*small interference RNA* o siRNA) se presenta entre las estrategias de acción directa como una opción muy prometedora (82). Estos son un tipo de ARN interferente de 20-25 nt de longitud que promueven el silenciamiento de genes mediante la unión a una región específica diana del ARNm que se quiere silenciar, formando así un ARN de doble cadena que activa el Complejo de Silenciamiento Inducido por ARN (RISC) para que lo degrade. La gran ventaja de esta estrategia es el alto nivel de superposición entre los diferentes ORFs del genoma del VHB por lo que un siRNA dirigido contra un transcrito podría interferir con otros transcritos virales. Se ha demostrado que siRNAs dirigidos a los transcritos de *HBC* o de *HBX* se asocian a una detención de la expresión génica del virus, evidenciándose en la disminución de los niveles de HBsAg, HBeAg y también del material genético viral. Por ejemplo, un estudio hecho con una combinación de siRNAs dirigidos al *HBX* (ARC-520) en combinación con entecavir mostró una disminución de HBsAg en pacientes HBeAg negativos (83–85).

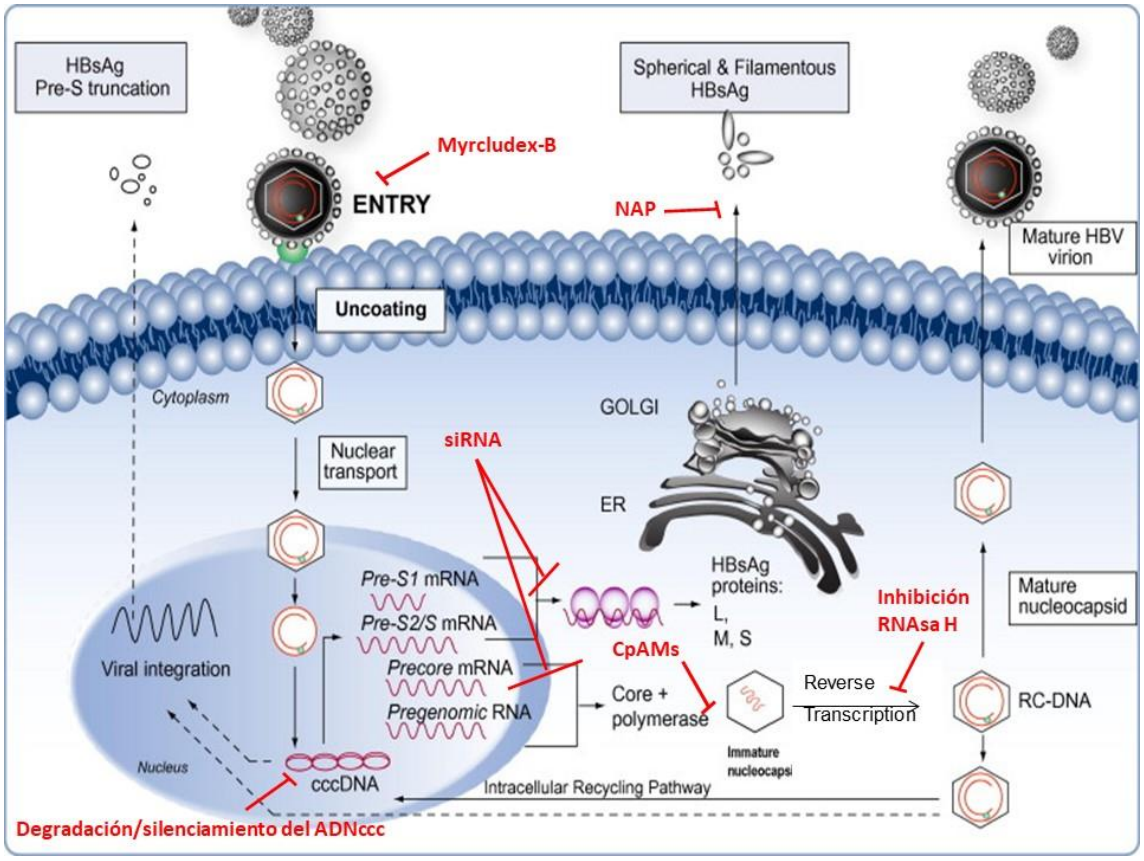


Figura 10: El ciclo replicativo del VHB y los puntos en los que podrían interferir los antivirales de acción directa. Las diferentes estrategias que se comentan en el texto están marcadas en rojo y se indica el punto del ciclo replicativo del VHB en el que interfieren. NAP: polímeros de ácidos nucleicos fosfotioados; CpAMs: moduladores alostéricos de HBc; siRNA: ARN pequeños de interferencia. Imagen modificada de Chan HLY et al (63).

1.6 HBc: una proteína estructural y funcional clave en la replicación del VHB

Como se ha detallado tanto en el apartado 1.3.2: ORF PreC/Core, como en el apartado 1.4: Ciclo replicativo del VHB, la proteína Core o HBc surge de la expresión del segundo codón de inicio de traducción del ORF PreC/Core (nt 1901, posición que indica el inicio del gen *HBC)*.

HBc es una proteína imprescindible a nivel estructural para el VHB y, además, desempeña importantes funciones en el ciclo viral. Puede encontrarse tanto en el citoplasma como en el núcleo del hepatocito infectado (41), por lo que desempeña un gran abanico de funciones que se detallan a continuación.

1.6.1 Dominios funcionales de la proteina HBc

HBc es una proteína de 183 aa (185 aa en el genotipo viral A) y de 21 KDa de peso molecular (por eso también se conoce como p21). Desde un punto de vista funcional, se diferencian dos dominios. Uno en el extremo amino-terminal (NTD, aa 1-149) y uno en el carboxi-terminal (CTD, aa 150-183) unidos por una región bisagra (RB entre los aa 140-149; normalmente incluida en el NTD) (Figura 11). El dominio NTD cumple con la función de autoensamblaje mientras que el CTD es el dominio funcional y versátil de la proteína, con un papel muy importante en el ciclo viral (86).

Figura 11: Representación esquemática de los dominios funcionales de HBc. Los distintos dominios están representados por distintos colores: en azul el NTD, en azul claro la región bisagra y en verde el dominio CTD. Las posiciones aminoacídicas de los dominios están reportadas encima de cada extremo.

En la estructura tridimensional de la cápside viral, el dominio NTD queda en la superficie de la cápside, formando una estructura de espículas hacia el exterior, posibilitando la dimerización y así, el ensamblaje. En este dominio se encuentran dos regiones importantes desde un punto de vista inmunológico: la principal región de reconocimiento por parte de células B (aa 74-83) (87), y dentro de esta, la región inmunodominante MIR (*Major Immunodominant Region*, aa 78-82) (88).

El dominio CTD, gracias a sus regiones ricas en arginina, está involucrado en la interacción con el genoma viral. Sus múltiples funciones serán descritas con más detalles en los apartados que siguen.

1.6.2 HBc y su rol estructural: ensamblaje y formación de la cápside viral

La estructura de HBc es helicoide (Figura 12 A) y consta de 5 hélices alfa de las cuales las hélices 2 y 3 crean una horquilla (“*hairpin”)* que marca el patrón de plegamiento de la proteína y permite la formación de dímeros que se estabilizan por un puente disulfuro entre los residuos cys-61 (Cisteína en la posición aminoacídica 61) de cada monómero (marcado en verde en la Figura 12 B) (89). Los dímeros tienen forma de T invertida, proyectando hacia

el exterior las 4 hélices encargadas de la dimerización (2 hélices por monómero), que conformarán las espículas de la cápside (90) en cuya punta se localiza la región inmunodominante MIR. Los dímeros se acoplan de 3 en 3 formando hexámeros (6 HBc en total, Figura 12 C) que finalmente se ensamblan formando la típica cápside icosaédrica viral (Figura 12 D). Esta podrá ser más grande o pequeña, con un diámetro de 34 o 30 nm y una simetría con número de triangulación del empaquetamiento T=4 o T=3 (donde T es el número de triángulos en que queda subdividida cada cara del icosaedro) en función de si contiene 240 o 180 moléculas individuales de HBc respectivamente (41). Los dos tipos de cápside comentados se pueden encontrar en los hepatocitos infectados, no obstante, las T=4 parecen seleccionarse preferentemente para ser envueltas y formar partículas infecciosas (91).

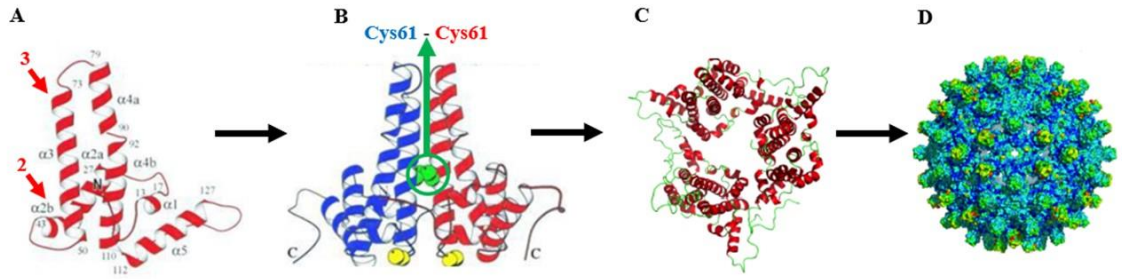


Figura 12: Formación de la cápside del VHB. A: Representación del monómero de HBc. Las hélices involucradas en la formación de la horquilla se indican e identifican con flechas y números en rojo. B: Estructura de los dímeros de HBc. Los dos monómeros se acoplan y su estructura se estabiliza mediante un puente disulfuro (evidenciado en verde) formado entre los residuos de Cisteína de la posición aminoacídica 61 de cada uno de los dos monómeros. C: Representación de un hexámero de HBc. D: Imagen de la cápside icosaédrica resultante del ensamblaje de HBc. Imagen modificada de Zlotnick A et al (89) y Wynne SA et al (92).

La cápside también está provista de unas fenestraciones que podrían ser cruciales para el acceso de los nt hacia su interior durante la síntesis de ADN (92).

1.6.3 HBc: una proteína funcional en la replicación viral y en la regulación celular.

La proteína HBc tiene un papel clave en la replicación viral, no solo por ser un elemento estructural, sino también por su rol funcional en el ciclo replicativo del VHB.

Una vez formada la cápside, los dominios CTD estarán orientados diferentemente según su estado de fosforilación. Este dominio presenta siete residuos de serina fosforilables en las posiciones aa 155, 162, 168, 170, 176, 178 y 181. Cuando su nivel de fosforilación es elevado, el dominio CTD quedará orientado hacia el interior de la cápside, y su posterior

desfosforilación parcial permitirá que ancle al ARNpg específicamente (interacción dada gracias a las cargas negativas de este), permitiendo así la encapsidación y retrotranscripción de este (93,94) (Figura 13). Las posteriores desfosforilaciones harán que el dominio CTD quede orientado al exterior donde interaccionará con proteínas del huésped. El estado de fosforilación del CTD se asocia también con la formación de partículas virales vacías (que no contienen el material genético viral en su interior). Se ha demostrado que una excesiva fosforilación determina un exceso de cargas negativas en el CTD que no permitiría la entrada del ARNpg en las cápsides neoformadas (93).

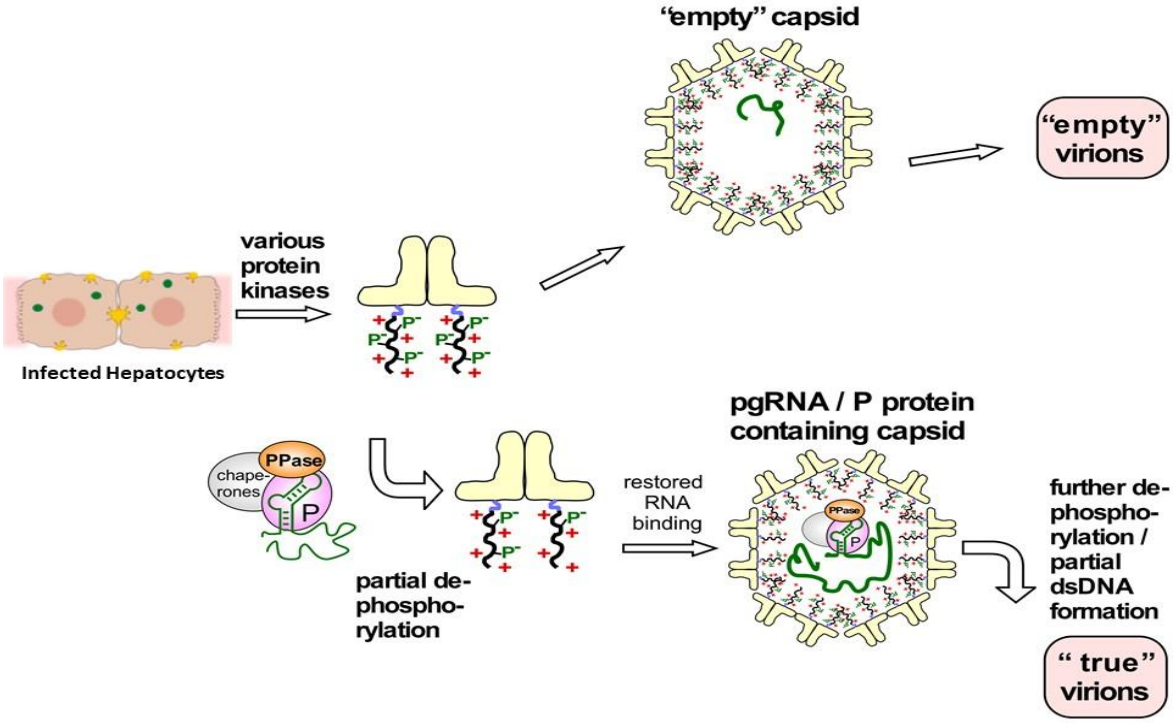


Figura 13: Estado de fosforilación del CTD de las proteínas HBc en las cápsides neoformadas y su rol en la replicación viral. La imagen esquematiza cómo el estado de fosforilación del CTD influye en la replicación del genoma viral. Partiendo de un estado altamente fosforilado del CTD, su desfosforilación parcial permite la encapsidación y retrotranscripción del ARNpg. De lo contrario, un exceso de fosforilación impide este proceso, lo que acaba resultando en cápsides vacías. Imagen modificada de Heger-Stevic J et al (93).

Los residuos fosforilables del CTD (en azul en la Figura 14) también son esenciales para regular la localización subcelular del virus a través de la activación de señales de localización nuclear (NLS), ya que al encontrarse fosforilados estas NLS quedan expuestas en la superficie (95). Estas NLS se encuentran en las repeticiones de arginina del CTD, entre los aa 150-152 y 164-167 (en rojo en la Figura 14). Hay otras dos repeticiones de arginina (entre los aa 157-159 y aa 172-175) en que se encuentran señales de retención citoplasmática (CRS, en verde en la Figura 14) que, de encontrarse más expuestas y por tanto dar una señal más

fuerte que las señales NLS harán que la cápside no se transporte al núcleo sino que se quede en el citoplasma, aunque se ha propuesto que en realidad podría desplazarse rápidamente entre el núcleo y el citoplasma de manera dinámica (96).

Figura 14: Dominio CTD de HBc y sus aa funcionales. La figura muestra la secuencia aminoacídica del dominio CTD y se reportan las posiciones de sus extremos. Sus aa funcionales más relevantes están reportados en distintos colores: en rojo se evidencian las repeticiones de arginina correspondientes a NLS, en verde las repeticiones de arginina correspondientes a CRS y en azul los residuos de serina fosforilables.

Además, gracias a las múltiples posibles interacciones del dominio CTD, HBc está involucrada en un amplio abanico de funciones diferentes (Figura 15):

- Interacciona con el ADNccc gracias a las cargas positivas de sus residuos de arginina del CTD, formando parte de la estructura del ADCccc actuando como una histona y aumentando hasta un 10% su nivel de compactación (60).
- Gracias a sus regiones ricas en arginina del CTD actúa como chaperona directa de ácidos nucleicos (41).
- A través de su interacción con chaperonas endógenas regula la estabilidad y el ensamblaje de la cápside (la interacción con HSP90 facilita la formación de la cápside mientras que la interacción con HSP40 promueve la degradación de HBc) (41).
- Interacciona con factores de localización como el factor de exportación nuclear 1 (NXF1/p15) y la maquinaria TREX que posibilitan la exportación nuclear de HBc como ribonucleoproteína junto con el ARNpg (41).
- Actúa como activador transcripcional del promotor Cp al cebar la unión entre NF-kB y el Enh II viral (41).
- Ceba la transcripción mediada por CRE via CRE/CREB/CBP (41).
- Funciona como represor transcripcional de ciertos genes del huésped. Por ejemplo, inhibe la transcripción de la proteína p53 mediante la interacción con el factor de transcripción e2f1, interfiriendo así con la apoptosis celular (41).

- A través de espectrofotometría de masas se ha postulado que HBc podría también interaccionar con varias proteínas celulares, incluyendo quinasas, interviniendo así con la expresión génica del huésped (41).

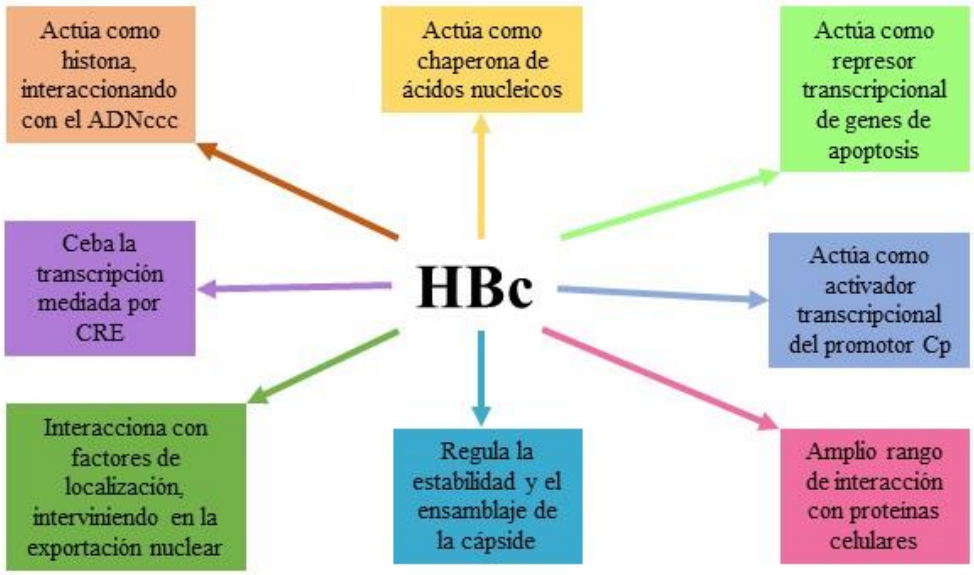


Figura 15: Funcionalidad de la proteína HBc. El diagrama esquematiza las diferentes propriedades funcionales que caracterizan a la proteína HBc y que se detallan en el texto del apartado.

1.7 Variabilidad del VHB y quasiespecies

Como demuestra la gran cantidad de genotipos y subgenotipos virales, el VHB se caracteriza por una gran variabilidad genética, por lo que circula como un conjunto de variantes genéticas estrechamente relacionadas entre sí, pero no idénticas.

1.7.1 Origen de la variabilidad genética del VHB

Como se ha detallado anteriormente en el apartado 1.4: Ciclo replicativo del VHB, el VHB es un virus de ADN que se replica mediante un intermediario de ARN (el ARNpg), por lo que su ciclo de replicación presenta una fase de retrotranscripción. Este paso hace que el VHB presente una alta variabilidad genética debido a que la retrotranscriptasa viral no presenta actividad exonucleasa 3'→5' de corrección de errores, por lo que el virus se replica

con una tasa de error elevada, que se traduce a un ritmo evolutivo (tasa de mutación) de entre 10^{-4} y 10^{-5} sustituciones/posición/año (97). Este ritmo evolutivo es similar al de algunos virus de ARN (98).

Este elevado nivel de variabilidad está además potenciado por la capacidad del virus de dar lugar a eventos de recombinación entre genotipos, fenómeno muy frecuente en el VHB (99). Esta recombinación es el resultado de la coinfección de un mismo huésped con distintos genotipos. Se han identificado nuevas variantes generadas por recombinación entre genotipos en todo el mundo. Por ejemplo, el genotipo predominante en la región del Tíbet es un recombinante de los genotipos C/D (100) y se han identificado recombinantes A/D, G/C y D/E en África (101).

Esta variabilidad genética también puede enriquecerse por el reciclaje nuclear de las cápsides neoformadas (68) o por la infección del mismo hepatocito por múltiples partículas virales (102), lo que permite eventos de recombinación genética viral sin necesidad de que sea entre genotipos distintos.

Los factores del huésped también juegan un papel importante en la variabilidad genética viral ya que hay proteínas que pueden favorecer la aparición de mutaciones como, por ejemplo, la enzima antiviral *Apoliprotein B mRNA Editing enzyme, Catalytic polypeptide-like 3G* (APOBEC3G) que induce la hipermutación de residuos de guanina-adenina en virus como el VHB, VIH y retrotransposones (103,104).

1.7.2 Quasiespecie viral

La suma de los factores que aportan variabilidad genética comentados en el apartado anterior genera un amplio espectro de genomas del VHB que se seleccionarán con mayor o menor intensidad dependiendo de su capacidad replicativa (*fitness*) y de la presión de selección determinada por la respuesta inmune y por la presencia de antirretrovirales. Por esta razón, el virus circula como un conjunto de secuencias o variantes genéticas estrechamente relacionadas entre sí que dan lugar a lo que se conoce como quasiespecie viral (QS) (105). En otros términos, la QS es una distribución ordenada y estable de mutantes dominados por un genoma principal, denominado secuencia consenso (*master secuence*) (106) (Figura 16).

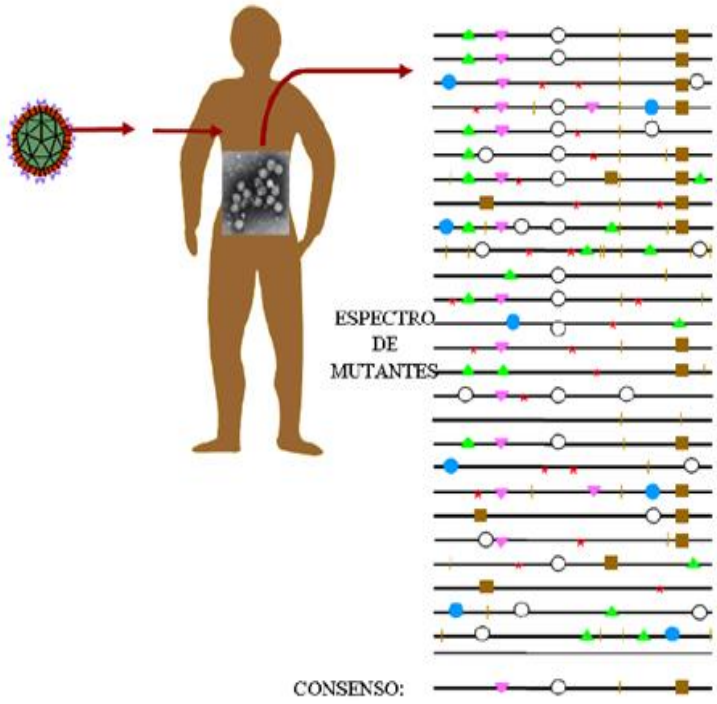


Figura 16: Representación gráfica de una quasiespecie viral. La figura esquematiza la composición de una quasiespecie viral, en la que el genoma que se presenta en la parte inferior representa la secuencia consenso (o *master secuence*), y los genomas superiores representan todas las posibles distintas variantes genéticas. Los símbolos de colores representan diferentes mutaciones en las secuencias respecto a la secuencia consenso.

El elevado grado de variabilidad permite al virus su continua adaptación a un ambiente cambiante, donde la QS que forma la población viral actúa como una unidad de selección, un espectro de mutantes con *fitness* diferentes en cada ambiente determinado. Por ese motivo muchas características o propiedades del virus no se pueden analizar correctamente a través de una sola secuencia consenso, sino que para ello es necesario conocer las variantes genéticas que conforman la QS (107).

Durante muchos años la variabilidad genética viral se estudió mediante técnicas de secuenciación. Primeramente, se utilizaba la secuenciación Sanger, pero esta no permitía abarcar todo el abanico de secuencias que componen una QS viral. De hecho, esta técnica de secuenciación solamente permite identificar mutaciones con una frecuencia superior al 15-20%, sin poder evaluar realmente la incidencia de cada cambio específico en la población completa (108). Tener en cuenta este hecho es muy importante en el estudio de la QS del VHB si, además, se considera que el elevado grado de solapamiento de su genoma hace que la mutación de un gen se pueda ver reflejada en otro (50), influenciando así la capacidad del virus de replicar y generar una enfermedad con distintos grados de gravedad.

Se ha reportado una conexión entre la variabilidad genética del virus y los distintos cuadros clínicos, evidenciando por ejemplo el impacto los distintos genotipos en la infección crónica por el VHB (17). Distintas mutaciones se han asociado también a la evolución de enfermedad. Por ejemplo, las mutaciones T1753C y A1762T / G1764A (K130M / V131I en HBx) del BPC se identificaron como posibles marcadores pronósticos para el HCC (109,110).

Asimismo, la complejidad de una QS se conoce como la propiedad que cuantifica la diversidad y frecuencia de los conjuntos de genomas con la misma secuencia (haplotipos) con independencia del tamaño de la población viral de estudio (111). Esta complejidad informa de factores como el potencial patogénico, la respuesta al tratamiento antiviral, la evolución clínica y la seroconversión (112–117). Más concretamente, la complejidad de la QS del VHB se ha relacionado con los niveles de ALT, el genotipo viral, la cuantificación del ADN viral y el estado del marcador HBeAg (118).

Por lo tanto, el estudio de la población viral que infecta un paciente a través de plataformas de secuenciación masiva puede ser determinante en la práctica clínica ya que esto permite analizar todas las variantes que componen la QS viral, lo que resulta muy relevante tanto a nivel terapéutico como de seguimiento clínico.

1.7.3 Secuenciación masiva

Las técnicas de *Next Generation Sequencing* (NGS) permiten generar hasta millones de secuencias, garantizando el estudio de la población real de variantes que constituyen la QS. Además, gracias a este sistema de secuenciación los análisis de genómica, metagenómica, transcriptómica y de interacción son exhaustivos, baratos, rutinarios y generalizados, sin requerir esfuerzos significativos a nivel de escala de producción (119).

En este trabajo de tesis doctoral la QS viral del VHB se analizó a través de la plataforma de NGS MiSeq Illumina (Illumina, San Diego, USA). El procedimiento de secuenciación y sus características se detallan en el siguiente apartado.

1.7.4 MiSeq Illumina

Entre las últimas tecnologías de secuenciación disponibles en el mercado, la secuenciación por síntesis (SBS) de la plataforma Illumina representa la nueva promesa entre las técnicas de NGS. Esta plataforma está bien consolidada y presenta un muy elevado rendimiento, pues ofrece un flujo de trabajo sencillo y es muy poco susceptible a errores de homopolímeros (120). Esta permite una secuenciación masiva paralela usando un método patentado que detecta nucleótidos individuales a medida que se incorporan a las hebras de ADN durante la fase de extensión de la secuencia (121).

El funcionamiento de esta tecnología se basa en la generación de *clusters* (agrupaciones de secuencias idénticas amplificadas de forma clonal) a partir de librerías (colecciones de fragmentos genéticos amplificados o amplicones) previamente preparadas. Cada amplicón de la librería presenta en ambos extremos (5' y 3') unos adaptadores específicos (P5 y P7) que permitirán su hibridación con el manto de oligonucleótidos (*primers*) de la *flowcell*. La *flowcell* (Figura 17 A) es un dispositivo de vidrio grueso con canales en el que cada canal está recubierto aleatoriamente con un manto de oligonucleótidos complementarios a los adaptadores presentes en las muestras a secuenciar (Figura 17 B). Las muestras se amplifican de forma clonal creando los *clusters,* que se identifican como puntos brillantes en una imagen (Figura 17 C), representando cada uno de ellos miles de copias de la misma cadena de ADN en una zona de 1-2 micrones.

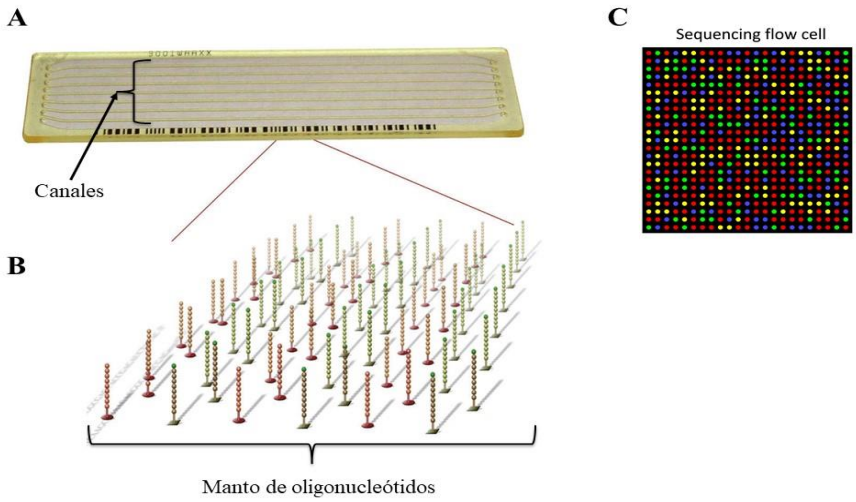


Figura 17: Representación de: A: *Flowcell*, se evidencian los diferentes canales que contiene. B: Manto de oligonucleótidos complementarios a los adaptadores de secuenciación que recubre cada canal de la *flowcell*. C: Imagen del proceso de secuenciación en la que cada punto brillante corresponde a un *cluster*. Imagen modificada de Lakdawalla A et al (122) y de Lowe R et al (123).

El proceso de secuenciación incluye diferentes fases.

Las librerías de ADN de cadena simple (amplicones previamente desnaturalizados), que ya presentan los adaptadores de secuenciación en sus extremos, hibridan con el manto de oligonucleótidos de la *flowcell*. Esto permite la reacción de extensión a través de la polimerasa (Figura 18, cuadrante A) formando moléculas de ADN de doble cadena. Estos productos de doble cadena se desnaturalizan y quedan unidas covalentemente a la superficie de la *flowcell* solamente aquellas hebras recién sintetizadas. Estas hebras se doblan formando puentes por hibridación con los oligonucleótidos complementarios adyacentes de la *flowcell* y éstos son extendidos por la polimerasa, obteniendo puentes bicatenarios como resultado (Figura 18, cuadrante B). Seguidamente los puentes se desnaturalizan, resultando en dos hebras monocatenarias neosintetizadas unidas covalentemente a los 2 respectivos oligonucleótidos (Figura 18, cuadrante C).

En este punto las hebras vuelven a doblarse, repitiendo este ciclo de amplificación hasta que se forman múltiples puentes (Figura 18, cuadrante D). Una vez formados todos los puentes, estos se desnaturalizan y las cadenas *reverse* son escindidas, quedando solamente cadenas *forward*. El extremo 3' tanto de las cadenas *forward* como de los oligonucleótidos del manto se bloquean para evitar el cebado indeseado del ADN. El *primer* de secuenciación entonces se hibrida con la secuencia adaptadora y empieza la secuenciación de las cadenas *forward* (Figura 18, cuadrante E).

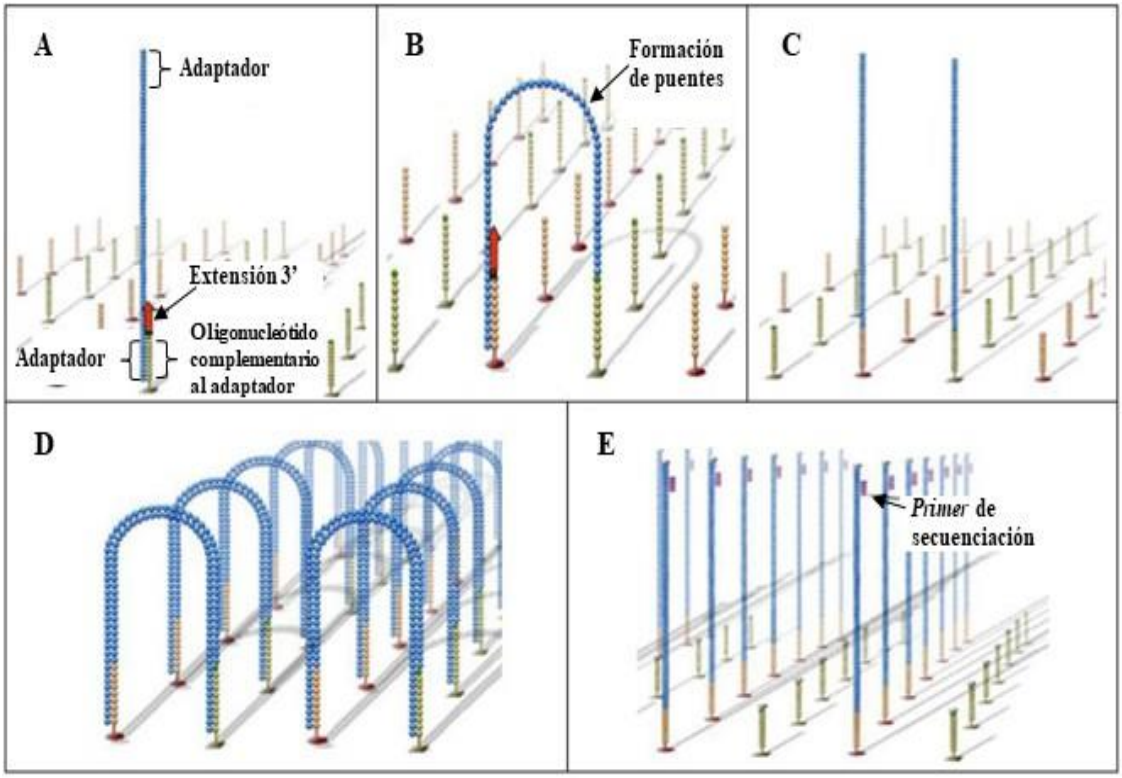


Figura 18: Secuenciación a través de la plataforma MiSeq Illumina (1). A: hibridación y extensión de las secuencias por la polimerasa y formación de ADN de doble cadena. B: desnaturalización, eliminación de las cadenas que han servido de molde y formación de puentes monocatenarios con su consiguiente extensión por la polimerasa (formación de puentes bicatenarios). C: desnaturalización de los puentes bicatenarios. D: Repetición del proceso de amplificación y formación de múltiples puentes. E: desnaturalización de los puentes, escisión de cadenas *reverse*, bloqueo en 3' de las cadenas *forward* y de los oligonucleótidos del manto y secuenciación de estas cadenas gracias al *primer* de secuenciación. Imágenes modificadas del archivo *Illumina Sequencing Overview* (124).

La secuenciación de los *clusters* se lleva a cabo gracias a un proceso de secuenciación por síntesis. Los 4 nt distintos se encuentran en la reacción unidos a fluorocromos de diferentes colores para su identificación. Cuando el nt se incorpora, el fluorocromo se libera y se excita gracias a una lámpara de fluorescencia, emitiendo una luz que será captada por la cámara y registrada (Figura 19). Illumina lleva a cabo esta secuenciación en dos fragmentos de 300 nt (*Read* 1 y *Read* 2) que posteriormente serán solapados parcialmente en el análisis bioinformático con tal de obtener la secuencia de la cadena original.

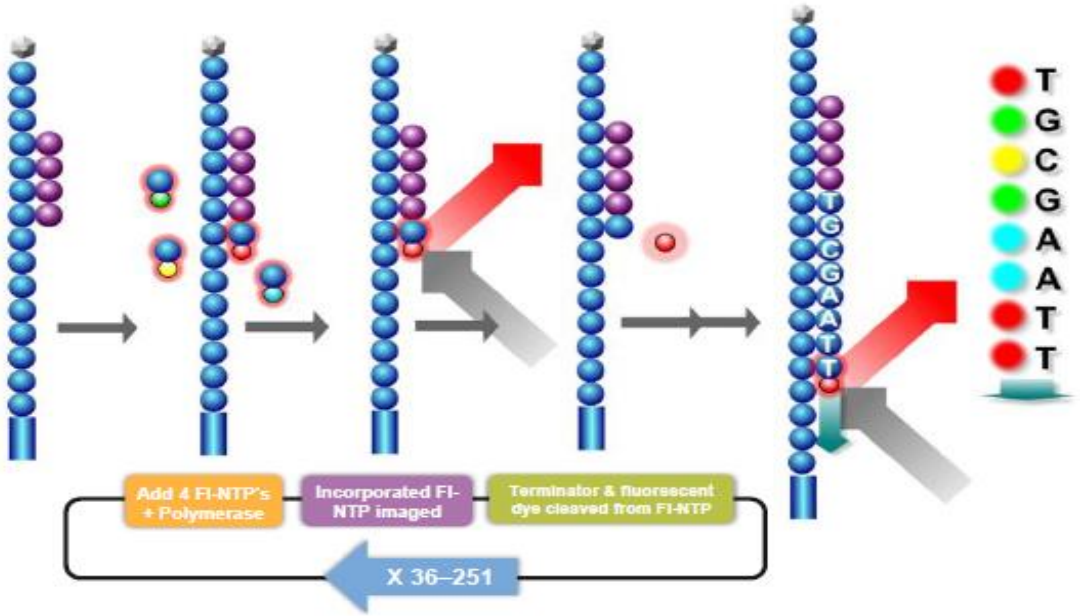


Figura 19: Pasos del proceso de secuenciación por síntesis. Los diferentes nt (A, T, C y G) se encuentran unidos a fluorocromos de distintos colores. Se añaden en tandas de 4 (uno de cada) a la reacción junto a una polimerasa. El nt que se una a la secuencia liberará al fluorocromo emitiendo una luz de un determinado color que será captada y registrada por la cámara. El proceso se repetirá tantas veces como bases haya que secuenciar. A: adenina; T: timina; C: citosina; G: guanina. Imagen extraída del archivo *Illumina Sequencing Overview* (124).

Una vez secuenciadas las cadenas *forward* se procede a la secuenciación de sus respectivas cadenas *reverse*. Este proceso, que se muestra en la Figura 20, se conoce como "*Paired-end sequencing*". Las cadenas ya secuenciadas se eliminan y los extremos 3' de las cadenas *forward* y de los oligonucleótidos del manto se desbloquean (Figura 20, cuadrante A). Las cadenas *forward* se doblan y se forman puentes de nuevo, que se extienden a 3' del oligonucleótido del manto (Figura 20, cuadrante B). Los puentes se desnaturalizan y las cadenas *forward,* usadas como plantilla, se escinden y se eliminan quedando solamente las cadenas *reverse* (Figura 20, cuadrante C). De nuevo, se bloquea el extremo 3' de estas cadenas *reverse* y de los oligonucleótidos del manto para evitar el cebado indeseado del ADN. El *primer* de secuenciación se hibrida con la secuencia adaptadora y se da otro ciclo de secuenciación, esta vez sobre las cadenas *reverse* (Figura 20, Cuadrante D).

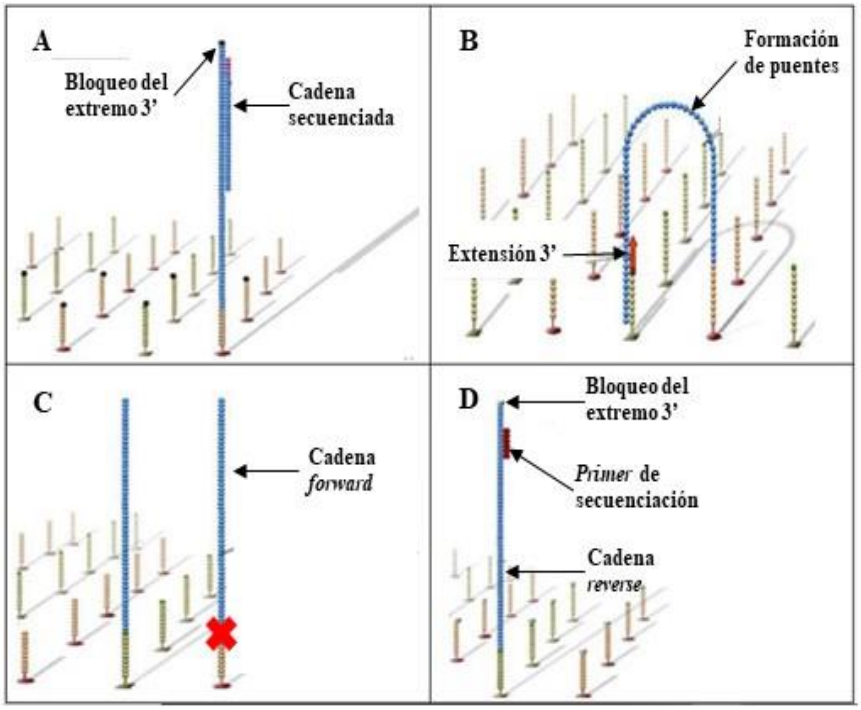


Figura 20: Secuenciación a través de la plataforma MiSeq Illumina (2). A: las cadenas recién secuenciadas se eliminan, el extremo 3' de las cadenas *forward* y de los oligonucleótidos del manto se desbloquean. B: las cadenas *forward* crean puentes que se extienden en 3'. C: desnaturalización de los puentes y eliminación de cadenas *forward* quedando solamente las *reverse*. D: bloqueo del extremo 3' de las cadenas *reverse* y de los oligonucleótidos del manto y secuenciación de estas cadenas gracias al primer de secuenciación. Imágenes modificadas del archivo *Illumina Sequencing Overview* (124).

Finalmente se obtiene un conjunto de secuencias (*reads*) que se analizaran a través de un seguido de filtros bioinformáticos, como se detallará en el apartado 4 (Materiales y métodos).

HIPÓTESIS

2. HIPÓTESIS

La proteína HBc del VHB, codificada por el gen *HBC*, es una proteína estructural y multifuncional esencial para el virus. Forma la cápside vírica y es una pieza fundamental clave en diferentes etapas del ciclo viral como la liberación del ADN del virus en el carioplasma nuclear o la formación y mantenimiento estructural del ADNccc. Además, gracias a las características de su dominio CTD, tiene la capacidad de modular diferentes procesos virales y celulares. Por estos motivos, esta proteína podría jugar un rol muy importante en la progresión de la enfermedad hepática.

Este proyecto de tesis doctoral se ha llevado a cabo con la hipótesis de que, dada la importancia de esta proteína, tanto su secuencia aminoacídica como la secuencia nucleotídica del gen que la codifica tendrían que mantenerse altamente conservadas al tener un rol fundamental en la viabilidad del virus. La identificación de regiones hiperconservadas (tanto a nivel de nt como de aa) comunes entre diferentes cuadros clínicos de la enfermedad hepática causada por la cronicidad del VHB indicaría la relevancia de estas regiones para la replicación viral. Estas regiones hiperconservadas comunes podrían ser de gran utilidad para el desarrollo de tratamientos antivirales basados en terapia génica al poder dirigir el tratamiento específicamente contra estas regiones esenciales para el virus. Al mismo tiempo, las regiones aminoacídicas hiperconservadas comunes podrían ser utilizadas como diana tanto a nivel terapéutico como de diagnóstico mediante la producción de anticuerpos altamente específicos.

Asimismo, la presencia de regiones que se encontraran diferentemente conservadas entre los distintos cuadros clínicos analizados contribuiría a la identificación de regiones que podrían tener un papel importante en la diferente evolución del daño hepático y que podrían ser útiles como factores pronósticos del avance de la enfermedad hepática. No solamente esto, sino que la detección diferencial de ciertas mutaciones en las secuencias, así como la variación de la complejidad de la quasiespecies entre estos cuadros clínicos, también podrían ser importantes factores pronósticos.

OBJETIVOS

3. OBJETIVOS

Considerando la importancia de la proteína HBc en la replicación viral y su posible rol en la progresión de la enfermedad hepática, este trabajo de tesis doctoral tiene como objetivo principal analizar mediante secuenciación masiva la conservación, la presencia de mutaciones y la complejidad de las quasiespecies del gen *HBC* y de su correspondiente secuencia aminoacídica (HBc) en pacientes con hepatitis crónica por VHB en diferentes estados de lesión hepática.

Para ello, se plantean como objetivos secundarios:

- Detectar regiones altamente conservadas, independientemente del estado de la enfermedad hepática, que puedan servir como dianas terapéuticas para una terapia génica dirigida.

- Identificar variables entre los diferentes grupos clínicos analizados que se puedan usar como factores pronósticos del avance de la enfermedad hepática. Esto consiste en analizar en las secuencias de *HBC* y HBc la presencia tanto de regiones conservadas o variables como de mutaciones específicas de un grupo clínico, así como analizar cómo varía la complejidad de las quasiespecies virales entre diferentes fases de la enfermedad hepática en un mismo grupo de pacientes.

MATERIALES Y MÉTODOS

4. MATERIALES Y MÉTODOS

4.1 Pacientes y muestras

Los pacientes con hepatitis crónica por VHB incluidos en los estudios se seleccionaron de la población general que acude al Hospital Universitario Vall d'Hebron en Barcelona. Se consideraron solo pacientes monoinfectados por el VHB (que dieron negativos para los virus de la hepatitis D, C y VIH) con una carga viral del VHB >3 log UI/mL (límite de sensibilidad de nuestro protocolo de amplificación).

Los marcadores virológicos del VHB (HBsAg, HBeAg y anti-HBe) se testaron en plasma a través de ensayos de quimioluminiscencia en el instrumento COBAS 8000 analyzer (Roche Diagnostics, Rotkreuz, Switzerland). El ADN del VHB se cuantificó a través de PCR a tiempo real con un límite de detección de 10 IU/mL en el instrumento COBAS 6800 (Roche Diagnostics, Rotkreuz, Switzerland).

Los pacientes fueron estratificados, de acuerdo con las bases de las últimas directivas europeas (EASL 2017 (31)), en 3 grupos en función del estado de la enfermedad hepática en el que se encontraban (determinado por biopsia hepática y/o por diagnóstico de imagen): pacientes con hepatitis crónica por VHB sin daño hepático (grupo CHB), con hepatitis crónica por VHB y cirrosis hepática (grupo LC) o con hepatitis crónica por VHB y carcinoma hepatocelular (grupo HCC). De estos últimos grupos (LC y HCC) se escogieron las muestras lo más próximas posible al diagnóstico de la complicación hepática (cirrosis y cáncer hepático respectivamente).

En el primer estudio de esta tesis doctoral hubo representación de los 3 grupos clínicos comentados y se analizó una muestra por paciente. En el segundo estudio solo hubo representación de los grupos CHB y HCC, pero en este caso se analizaron 2 muestras por paciente en tiempos diferentes (2 subgrupos por grupo: T0 y T1), entre las que había una diferencia mínima de tiempo de un año.

4.2 Amplificación del gen *HBC*

El ADN del VHB se extrajo a partir de 200 µL de suero de cada paciente usando el kit de extracciones QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) según el protocolo indicado por el fabricante. En cada extracción se incluyó un control "blanco" de extracción en el que se sustituyó el suero por agua estéril.

El gen *HBC* se amplificó a partir del ADN viral extraído a través de un proceso de 3 PCRs en cadena (Figura 21). En la primera reacción de PCR (PCR 1, o PCR externa) se amplificó una extensa región del genoma del VHB (nt 1774-2930) que englobaba al gen *HBC* (entre los nt 1901-2464 en el genotipo A i entre los nt 1901-2458 en el resto de los genotipos). En los siguientes pasos (PCR 2 y 3) el gen *HBC* se dividió en dos amplicones: un fragmento o amplicón 1, que abarcaba los nt 1863-2317 y el segundo fragmento o amplicón 2, que abarcaba los nt 2205-2483. Estos dos amplicones se sobreponen parcialmente entre sí en una zona terminal de 112 nt de longitud (región solapante: nt 2205-2317).

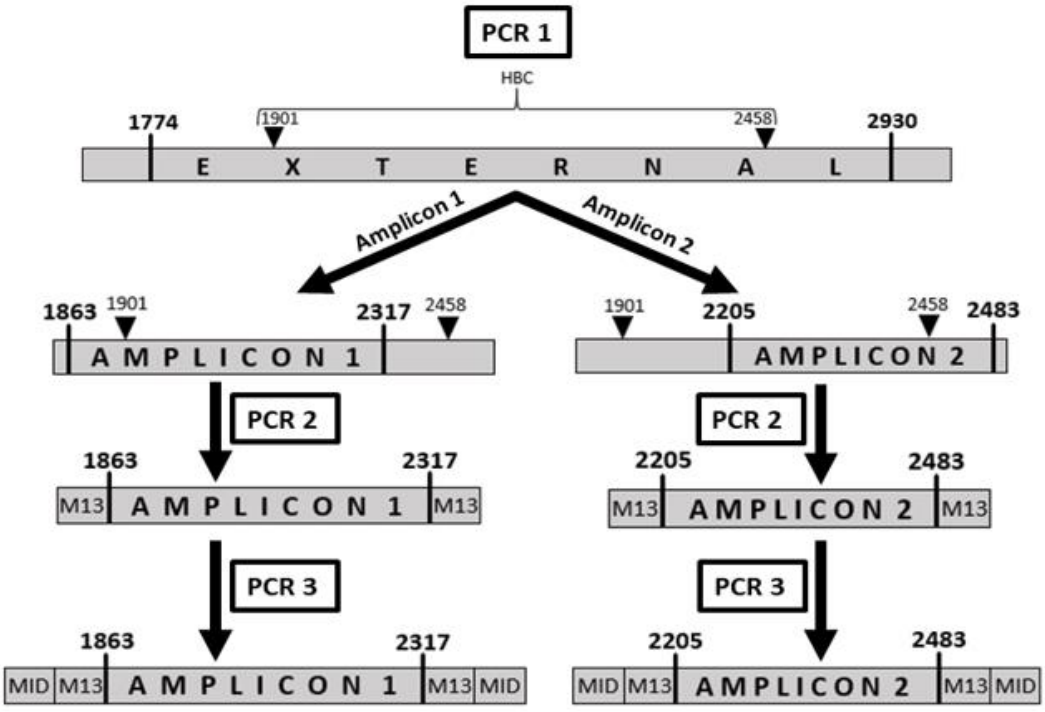


Figura 21: Resumen esquemático de los 3 pasos de PCR en cadena desempeñados para amplificar el gen *HBC*. En el primer paso (PCR1), se amplifica una extensa región que engloba al gen *HBC*. En el siguiente paso (PCR2), la región se divide en dos amplicones que se solapan en una porción de 112 nt de longitud. En este punto se les añaden colas M13. En el tercer paso (PCR 3) se agrega un identificador de muestra (MID).

En el segundo paso de las amplificaciones (PCR 2) se añadió una cola de secuencia M13 a los extremos de cada uno de los amplicones. Cabe destacar que la región de inicio del amplicón 2 difiere levemente entre los distintos genotipos virales por lo que en esta reacción de amplificación se utilizó un multiplexado de 4 *primers forward* (todos ellos a la misma concentración) con el fin de abarcar todos los genotipos virales (como se detalla en la Tabla 1). Las colas M13 añadidas se usaron como molde en la tercera y última PCR (PCR 3) que, a su vez, añadió una cola de 10 nt a cada extremo de los amplicones a modo de identificador específico de muestra (*primers* MID). La secuencia de los *primers* así como los protocolos de amplificación están reportados en la Tabla 1.

Pasos de amplificación	PCR	Primer	Secuencia del primer (5' -> 3')	Región amplificada	Protocolo
1º	PCR1	Forward	TAGGAGGCTGTAGGCATA	1774-2930	95ºC 5min; (95ºC 20s, 49ºC 20s, 72ºC 15s) x 35 ciclos; 72ºC 3min
		Reverse	GGAAAGAATCCCAGAGG		
2º	PCR2 A.1	Forward A.1	GTTGTAAAACGACGGCCAGTTTCAAGCCTCCAAGCTGT	1863-2317	95ºC 2min; (95ºC 20s, 58ºC 20s, 72ºC 15s) x 35 ciclos; 72ºC 3min
		Reverse A.1	CACAGGAAACAGCTATGACCGATAGGGGCATTTGGTGGTCT		
	PCR2 A.2	Forward 1 A.2	GTTGTAAAACGACGGCCAGTGGTTTCATATTTCTTGCC	2205-2483	95ºC 2min; (95ºC 20s, 50ºC 20s, 72ºC 15s) x 35 ciclos; 72ºC 3min
		Forward 2 A.2	GTTGTAAAACGACGGCCAGTGGTTTCACATTTCCTGTC		
		Forward 3 A.2	GTTGTAAAACGACGGCCAGTGGTTTCACATTTCTTGTC		
		Forward 4 A.2	GTTGTAAAACGACGGCCAGTGGTTTCACATTTCCTGCC		
		Reverse A.2	CACAGGAAACAGCTATGACCTCCCACCTTATGAGTCCAAG		
3º	PCR3	Forward	GTTGTAAAACGACGGCCAGT+10 nt MID (específicos)		95ºC 2min; (95ºC 20s, 60ºC 20s, 72ºC 15s) x 20 ciclos; 72ºC 3min
		Reverse	CACAGGAAACAGCTATGACC+10 nt MID (específicos)		

Tabla 1: *Primers* y protocolos de PCR. La tabla muestra las secuencias de los *primers forward* y *reverse* usados en cada paso de amplificación. Los nt subrayados indican las secuencias M13. Se reportan los protocolos de amplificación y las posiciones nucleotídicas de los extremos de los amplicones obtenidos (región amplificada). A.1: amplicón 1; A.2: amplicón 2; MID: identificador multiplex.

Una vez hecho este seguido de PCRs se obtiene pues, el gen *HBC* de cada muestra amplificado en 2 regiones solapantes tal y como se muestra en la Figura 22.

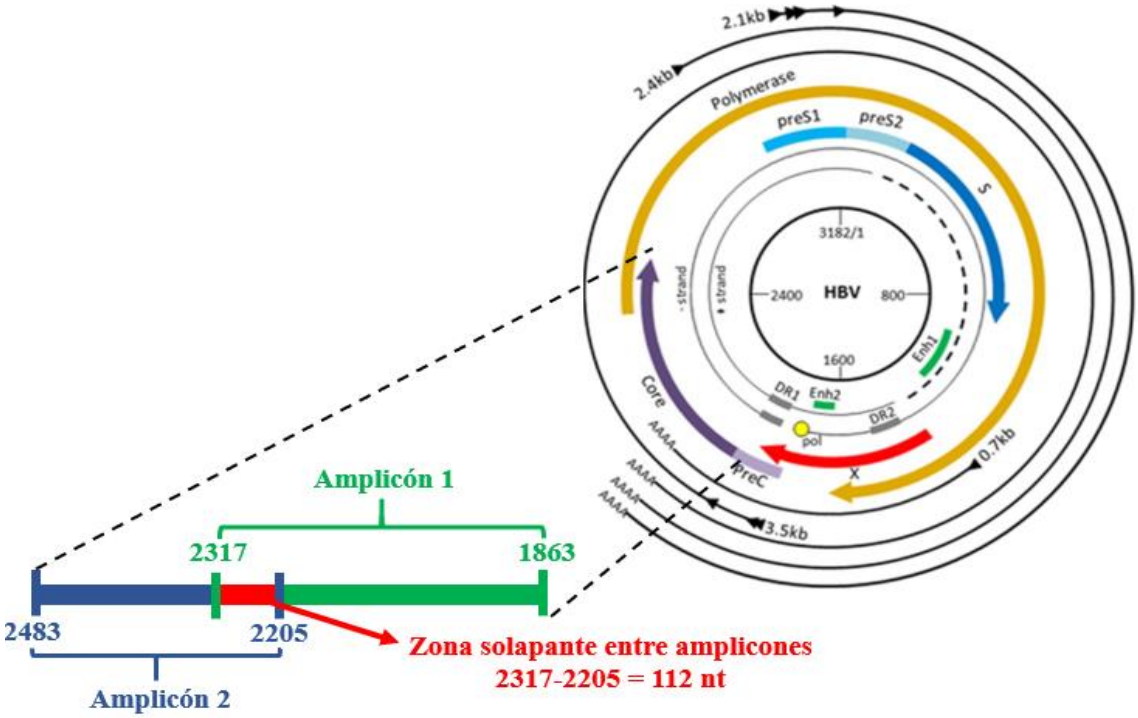


Figura 22: Representación de la región de estudio en relación con el genoma completo del VHB. Se especifican las posiciones de los dos amplicones (en verde el amplicón 1 y en azul el amplicón 2) y de la zona solapante entre ambos (en rojo). Imagen extraída de Minor MM et al (47).

Hay que considerar que, para la plataforma Illumina MiSeq (Illumina, San Diego, USA), se recomienda el uso de fragmentos a secuenciar que no superen los 600 pb, ya que los fragmentos de gran tamaño no hibridan correctamente con el manto de oligonucleótidos de la *flowcell* y podrían ser susceptibles a fraccionamiento, cosa que comprometería su correcta secuenciación. El gen *HBC* está compuesto por 557 nt (posiciones 1901 – 2458) pero hay que tener en cuenta que a lo largo de los pasos de amplificación se añaden 180 nt en total por cada amplicón (90 nt por cada extremo repartidos en 20 nt de cola M13, 10 nt del identificador MID, y 60 nt del adaptador de secuenciación de MiSeq Illumina). Por esta razón y con el fin de garantizar una adecuada eficiencia en la secuenciación, el gen se analizó repartido en dos amplicones de menor tamaño parcialmente solapantes como se ha indicado en este apartado.

Todas las PCRs realizadas a lo largo de este proceso se llevaron a cabo usando la polimerasa de alta fidelidad Pfu Ultra II DNA (Stratagene, Agilent Technologies, Santa Clara, United States). En cada paso se pusieron los controles de amplificación pertinentes (el control blanco de extracción y agua estéril como controles negativos y el plásmido TriEx-HBV con el genoma completo del VHB como control positivo).

Para cada muestra, la amplificación de ambos amplicones se corroboró mediante electroforesis en gel de agarosa al 2% en el que se comprobaron tanto la correcta amplificación de los amplicones (presencia de una banda limpia del tamaño pertinente) como la ausencia de contaminaciones (ausencia de banda en los varios controles negativos).

4.3 Preparación de librerías y secuenciación por NGS

4.3.1 Purificación de amplicones

Los amplicones se purificaron con KAPA Pure beads (Roche Diagnostics, Rotkreuz, Switzerland) a través de un protocolo automatizado en el instrumento TECAN (*laboratory protocols automation machine*; Tecan Trading AG, Switzerland). Estas son pequeñas esferas magnéticas cargadas positivamente que atraen y ligan a ellas fragmentos de ADN gracias a la carga negativa de los fosfatos presentes en el material genético. Los dos amplicones se purificaron por selección positiva, es decir, que durante los lavados con etanol las moléculas de ADN de interés (de mayor tamaño) se mantuvieron en el tubo gracias a la interacción entre las esferas y un imán. Por esta razón, la adecuada funcionalidad de estas esferas magnéticas depende de la proporción en la que se usan en relación con el tamaño de los fragmentos genéticos a purificar, puesto que un exceso de cargas positivas en la mezcla podría ocasionar la purificación de fragmentos no deseados, mientras que un déficit de cargas positivas podría ocasionar la inadecuada purificación de la totalidad del material genético de interés. En este estudio se usaron estas esferas magnéticas a 0,8X en proporción a las muestras.

La calidad de la purificación se verificó electroforéticamente usando la plataforma Agilent 2200 TapeStation System y el kit D1000 ScreenTape (Agilent Technologies, Waldbronn, Germany). El *software* incorporado Agilent 2100 Bioanalyzer permite una visualización óptima del resultado de la electroforesis, mientras que el uso de una recta patrón estándar de ADN proporcionada con el kit D1000 ScreenTape hacen que este sistema ofrezca una información muy detallada de la presencia y longitud de los fragmentos presentes en cada muestra. Este paso es importante para asegurar que se hayan eliminado todos los posibles

fragmentos cortos no deseados que, si no se eliminaran, podrían hacer disminuir la eficiencia de la secuenciación de los amplicones de interés.

4.3.2 Cuantificación y normalización de amplicones y formación de la librería

Los amplicones, ya correctamente purificados, se cuantificaron a través de un ensayo fluorimétrico (Quant-iT PicoGreen dsDNA Assay Kit, Thermo Fisher Scientific-Life Technologies) que se basa en añadir un fluoróforo intercalante (Quant-iT™ PicoGreen® dsDNA reagent) que permite la cuantificación ultrasensible de las moléculas de ADN de doble cadena. Los amplicones se normalizaron con tampón EB (10 mM Tris-HCl, pH 8,0–8,5) todos ellos a la misma concentración de 0,5 ng/µL para que al juntar todos los amplicones en la librería estos estuvieran igualmente representados en la mezcla.

Teniendo en cuenta que la secuenciación por MiSeq Illumina es más efectiva sobre fragmentos cortos, poniendo en la misma proporción los dos amplicones el amplicón más corto (amplicón 2) se secuenciaría con mayor eficiencia. Por esta razón, la proporción entre ambos amplicones se adaptó para que el amplicón más largo (amplicón 1) se encontrara en mayor proporción (2,5 amplicón 1 : 1 amplicón 2).

4.3.3 Preparación de la librería

Llamamos librería a la colección de fragmentos genéticos amplificados que serán secuenciados conjuntamente. Los pasos requeridos para prepararla para su secuenciación por MiSeq Illumina se detallan a continuación.

- Reparación y adenilación de extremos y ligación de adaptadores

Para que la librería se pueda usar en los siguientes pasos, los extremos de los amplicones tienen que pasar de cohesivos a romos. Este paso de reparación se llevó a cabo incubando la librería junto a un mix de enzimas (End Repair & A-tailing Enzyme Mix*)* en un termociclador a 20°C durante 30 minutos y seguidamente a 65°C otros 30 minutos. Este proceso no solo deja todos los extremos romos, sino que también les añade una adenina en posición 3’ (Figura 23 A). La reacción de ligación entre la adenina añadida y los adaptadores de secuenciación (a concentración de 15 µM) se llevó a cabo incubando la librería junto con el adaptador y una enzima ligasa a 20ºC durante 15 minutos (Figura 23 B).

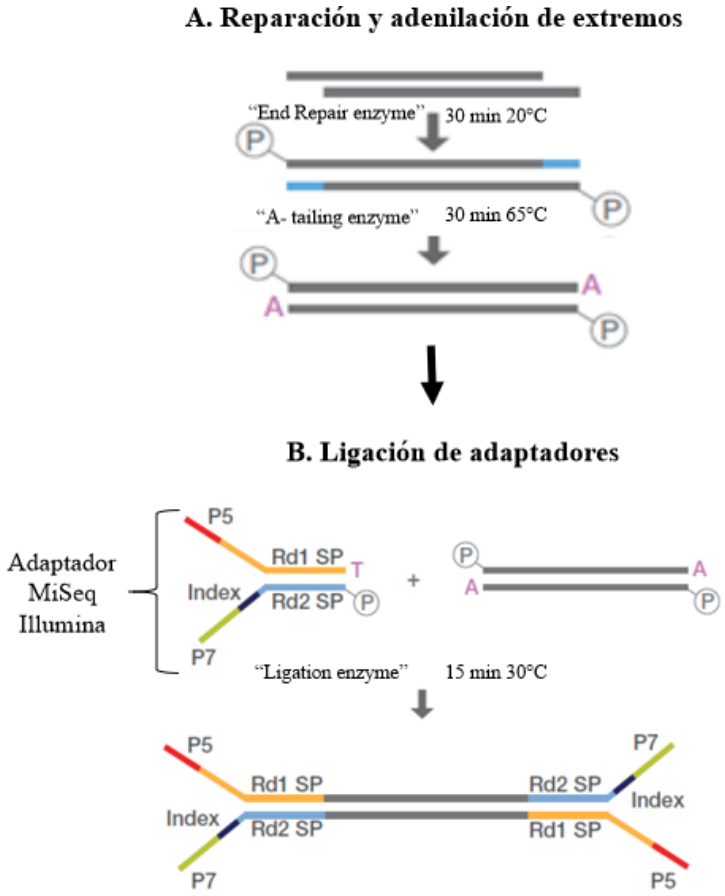


Figura 23: Representación esquemática de los primeros pasos de preparación de la librería. A: reparación y adenilación de extremos. B: ligación de adaptadores. Las enzimas y características de las reacciones están reportadas. A= adenina; P= grupo fosfato.

Los pasos que siguen para la preparación de la librería se esquematizan en la Figura 24.

- Lavados post-ligación

La librería se purificó nuevamente a través de KAPA Pure Beads (proporción 0,8X), eliminando de esta forma el exceso de adaptadores no ligados. La mezcla de las esferas magnéticas con la librería se incubó 10 minutos a 22ºC agitándose a 1400 rpm. Posteriormente el tubo con la mezcla se puso en un imán (DynaMag™-2 Magnet) y se lavó dos veces consecutivas con etanol al 80% con el fin de eliminar los fragmentos más cortos (selección positiva de los amplicones). La librería se eluyó de las esferas con tampón EB (incubándose a 37ºC durante 5 minutos para facilitar esta disociación) y seguidamente se hizo un último paso por el imán (selección y retención de las esferas ya libres de material genético de interés).

- Amplificación intermedia y cuantificación

Antes de secuenciar la librería se hizo una amplificación intermedia para amplificar los amplicones que se pudieran encontrar a baja concentración. En este paso se usaron los kits KAPA HiFi HotStart Ready Mix (2x) y KAPA Library Amplification Primer Mix (10x) en un termociclador con el siguiente protocolo: 1 ciclo de 45s a 98ºC, 5 ciclos de 15s a 98ºC, 30s a 60ºC y 30s a 72ºC y un último ciclo de 1 minuto a 72ºC. Los productos se lavaron nuevamente con KAPA Pure Beads (0,8X) como se ha descrito anteriormente.

La librería se cuantificó (a una dilución 1:10000 en tampón EB) por qPCR con el reactivo KAPA SYBR FAST qPCR Master Mix (2X) (con Primer Premix (10X)) en el instrumento LightCycler 4800. La recta patrón a partir de la cual se calculó la concentración de la librería se obtuvo cargando 6 diluciones seriales del ADN estándar proporcionado en el KAPPA Library Quantification Kit. Tanto la librería como los estándares se analizaron por triplicado.

- Pool de librerías o máster pool

El máster pool es el conjunto de librerías distintas que se preparan y cargan simultáneamente en el aparato MiSeq con el fin de aportar variabilidad en la fase de secuenciación. Cada librería se puso en el máster pool a una concentración de 4 nM (previamente diluida con tampón EB). Las diluciones 1:400 y 1:4000 (hechas con tampón EB) del máster pool se cuantificaron nuevamente por qPCR en el instrumento LightCycler 4800.

Al terminar la cuantificación, se ajustó la concentración del máster pool hasta 4 nM con tampón EB.

- Desnaturalización del máster pool

Para garantizar que los amplicones hibriden con el manto de oligonucleótidos de la *flowcell* del instrumento MiSeq Illumina, es necesario que los fragmentos genéticos bicatenarios se desnaturalicen para formar fragmentos de cadena simple. Por lo tanto, el máster pool y el reactivo PhiX (biblioteca de control de ADN genómico del fago phi que se añade para aumentar la diversidad de la secuenciación) se trataron con NaOH a 0,2N y el exceso de pH básico se neutralizó con el compuesto Tris-HCL a 200mM (pH 7-8). El máster pool y el PhiX ya desnaturalizados se diluyeron hasta 12-15 pM con el tampón de hibridación HT1 y finalmente se mezclaron en una proporción de 4:1 (máster pool : PhiX) en un total de 600 µL.

El producto obtenido, llamado pool de carga, se sometió a 95ºC durante 2 minutos para asegurar la desnaturalización y seguidamente se cargó en el instrumento MiSeq Illumina para su secuenciación.

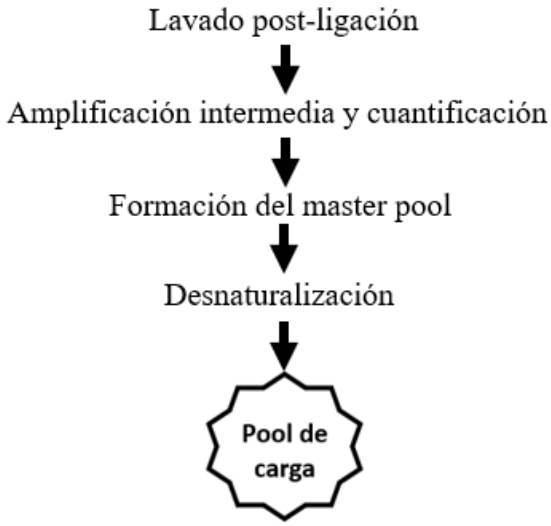


Figura 24: Esquema del flujo de trabajo de la preparación de la librería. Se mencionan los pasos seguidos a partir de la ligación de adaptadores hasta llegar al pool de carga.

4.4 Análisis bioinformático de los datos de secuenciación obtenidos: Filtros de calidad

A las lecturas de las secuencias obtenidas una vez finalizada la secuenciación (llamadas *reads*) se les aplicaron un seguido de filtros bioinformáticos desarrollados por nuestro grupo (125) en el programa de código abierto R (126), utilizando las librerías Bioconductor (127) y Biostrings (128). Los filtros bioinformáticos a los que se hace referencia, esquematizados en la Figura 25, se detallan a continuación:

1. Filtrado de calidad de las secuencias: se eliminaron aquellos *reads* que contenían indeterminaciones o los que no se habían secuenciado en su totalidad (de menor longitud). También se evaluaron los parámetros generales del instrumento MiSeq Illumina relacionados con la calidad de la secuenciación.
2. Colapso de los haplotipos: se basa en sobreponer los *reads* obtenidos en sentido *forward* con los obtenidos en sentido *reverse* de una misma secuencia para que

solapen parcialmente entre ellos y así poder comprobar la veracidad de ambas lecturas. Para este paso se utilizó el entorno de programación FLASH (129) imponiendo un mínimo de 20 pb solapantes entre los *reads forward* y *reverse,* con un 10% de bases desapareadas como máximo de tolerancia. Esto permitió eliminar aquellos *reads* con una calidad y veracidad de secuenciación baja.

3. Eliminación de *reads* erróneos: se descartaron todos los *reads* en los que más del 5% de sus bases obtenían un valor inferior a 30 en la puntuación de Phred (130). Este parámetro es particularmente eficaz a la hora de discriminar entre bases erróneas y correctas. Un valor superior a 30 en este parámetro corresponde con una precisión en la secuenciación del 99.9%.
4. Demultiplexado de los *reads*: este paso consiste en la asignación de los *reads* a cada amplicón de cada paciente mediante la detección de las secuencias específicas de identificación (MID). Los adaptadores de Illumina se usaron para distinguir entre las diferentes librerías del pool y los primers M13 para distinguir entre las dos hebras. En este paso se aceptaron hasta tres desapareamientos entre bases ya que las secuencias de estos *primers* son de 20 nt o más, tanto para M13 como para los adaptadores). En la asignación del MID se permitió solo un desapareamiento entre bases debido a que son secuencias muy cortas (10 nt). Finalmente, los MID, M13 y adaptadores se recortaron, obteniendo un archivo "fasta" para cada combinación de MID – adaptador – hebra. En estos archivos los *reads* se colapsaron en haplotipos con sus frecuencias correspondientes.
5. Eliminación de artefactos respecto al haplotipo de referencia: Por artefactos se entienden aquellos haplotipos que no han cubierto el amplicón completo, haplotipos con más de 2 indeterminaciones, haplotipos con 3 *gaps* o más y haplotipos con más de 99 bases diferentes respecto a la secuencia consenso. Estos artefactos se identificaron en cada uno de los archivos fasta alineando los haplotipos con la secuencia consenso (el haplotipo más abundante en la muestra). Las indeterminaciones y los gaps aceptados se repararon según la secuencia consenso.
6. Intersección de haplotipos: se seleccionaron aquellos haplotipos con una abundancia igual o superior al 0,1% en ambas hebras y se descartaron todos los haplotipos únicos para una sola hebra. La cobertura de los haplotipos que pasaron el filtro de calidad se calculó como la suma de los *reads* restantes en ambas cadenas.

7. Selección final de los haplotipos que forman la QS: solo los haplotipos con abundancia igual o superior al 0,25% y con un buen solapamiento entre amplicón 1 y 2 en la zona de 112 nt fueron considerados en la definición de la QS viral y fueron objeto de los análisis que se describen a continuación.

Figura 25: Filtros de calidad del análisis bioinformático. El diagrama de flujo muestra todos los pasos de filtros de calidad al que se someten los datos de secuenciación obtenidos de MiSeq Illumina con tal de obtener el muestreo de las QS virales del VHB de cada paciente.

Los protocolos i métodos bioinformáticos aplicados en los diferentes experimentos y estudios fueron revisados por Mercedes Guerrero-Murillo del departamento de Microbiología del Hospital Universitario Vall d'Hebron y por el Dr. Josep Gregori, perteneciente al grupo de investigación de hepatitis virales del VHIR, al grupo de investigación CIBERehd y a Roche Diagnostics SL.

4.5 Genotipado de los haplotipos

El análisis del genotipo viral se hizo mediante un análisis filogenético basado en distancias genéticas (DB rule) (131,132), en el que se tiene en consideración la variabilidad tanto dentro de un mismo genotipo como entre diferentes genotipos.

Se seleccionaron 106 secuencias completas del VHB representativas de los diferentes genotipos y subgenotipos virales A-J (13 secuencias del genotipo A, 20 del B, 24 del C, 17 del D, 8 del E, 10 del F, 6 del G, 5 del H, 2 del I y 1 del genotipo J) obtenidas de NCBI GenBank (National Center for Biotechnology Information, USA), de las cuales se extrajo la región genómica de interés (gen *HBC*). En estas secuencias se calcularon las distancias genéticas máximas entre secuencias del mismo genotipo, así como las distancias genéticas mínimas entre secuencias de diferentes genotipos. La secuencia de un haplotipo se consideró del mismo genotipo que una de las secuencias de referencia al observar una coincidencia mínima del 96% entre ambas.

Las distancias genéticas necesarias para el genotipado se obtuvieron de acuerdo con el modelo Kimura-80 (133).

Los números de acceso a estas secuencias en NCBI GenBank se detallan en el Anexo 2: Tabla suplementaria.

4.6 Análisis de la conservación

El grado de conservación de las secuencias tanto a nivel de nt como de aa se determinó calculando el contenido de información de cada posición en cada uno de los haplotipos detectados en cada paciente.

Con tal llevar a cabo el análisis de la conservación sobre las secuencias de *HBC* completas (y no separadas por amplicones), el contenido de información se calculó para cada posición exclusiva del amplicón 1 y el amplicón 2, mientras que para las posiciones incluidas en la región de 112 nt solapante se determinó el contenido de información, para cada posición de esta, como la media de los valores de contenido de información obtenidos en amplicón 1 y amplicón 2.

El cálculo del contenido de información se basa en la entropía de Shannon, que se define como el número de decisiones binarias (número de preguntas cuya respuesta es sí / no) requerido para encontrar el elemento correcto en un conjunto de N elementos y se representa como:

$$CI_j = \log_2(N) - \sum_{i=1}^{j} p_{ij} log_2(p_{ij}) *$$

**Donde j es la posición en el alineamiento, pj es la frecuencia del haplotipo en la quasispecie viral en la posición j y N es el número de posibilidades (4 nt o 22 aa). El contenido de información (CI) varía desde 0 (indicando máxima incertidumbre o variabilidad) hasta 2 en nt o 4,32 en aa (indicando máxima información o conservación).*

La conservación en la región de interés se evaluó aplicando un análisis por *sliding windows* (ventanas deslizantes), que consiste en representar gráficamente el promedio del contenido de información de ventanas de 25 nt o 10 aa avanzando en pasos de 1 nt o aa en la secuencia (Figura 26). Este análisis permite identificar las regiones más conservadas, es decir, las que presentan un valor de contenido de información más alto. Las secuencias aminoacídicas de los haplotipos se obtuvieron a partir de la traducción de sus respectivas secuencias nucleotídicas usando la pauta de lectura del gen *HBC*.

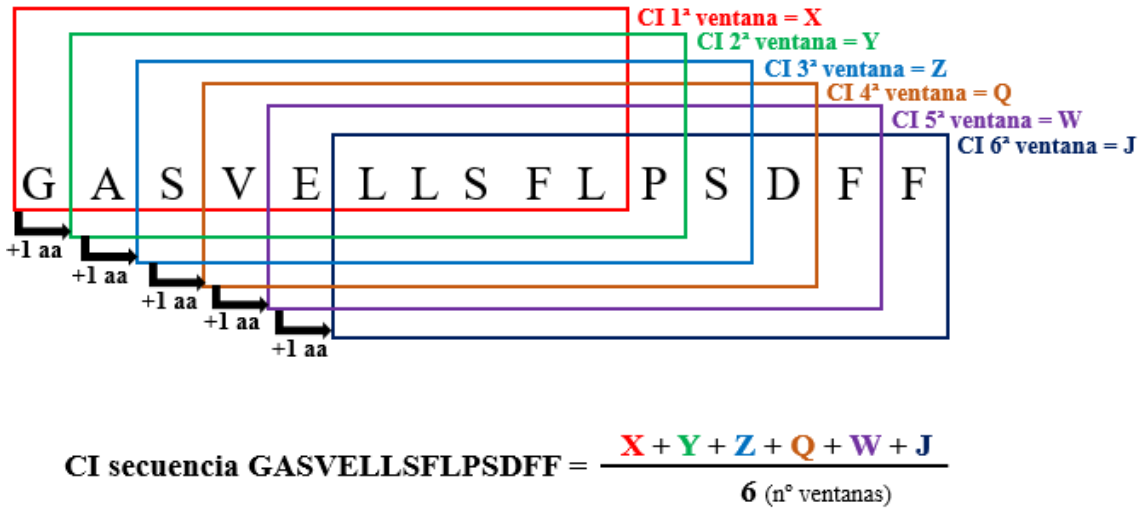


Figura 26. Representación esquemática del proceso de análisis del contenido de información por *sliding windows* de una secuencia aleatoria aminoacídica (ventanas de 10 aa). X corresponde al contenido de información calculado para la primera ventana de 10 aa, Y al calculado para la segunda ventana de 10 aa y así sucesivamente para Z, Q, W y J. CI= contenido de información.

En el primer estudio, este análisis se aplicó tanto en la población total de pacientes incluidos en el estudio sin tener en cuenta su estado clínico como específicamente en los pacientes pertenecientes a cada grupo clínico. En el primer caso (población total de estudio, sin tener en cuenta el estado clínico) se implementaron dos escenarios. En el primer escenario el contenido de información se calculó a partir de todos los haplotipos presentes en las QS de todos los pacientes sin tener en cuenta sus frecuencias relativas (análisis por haplotipos), por lo que cada haplotipo tuvo el mismo peso en el análisis. En el segundo escenario el contenido de información se calculó teniendo en cuenta la frecuencia relativa de cada uno de los haplotipos (análisis por frecuencia de haplotipos). Las frecuencias relativas, de hecho, nos informan sobre la abundancia de los haplotipos y esto está altamente relacionado con la *fitness* relativa de cada uno de ellos (111). En el segundo caso del primer estudio (pacientes pertenecientes a cada grupo clínico) solamente se implementó el escenario de análisis por haplotipos (sin tener en cuenta la frecuencia relativa de los haplotipos).

En el segundo estudio (2 grupos a 2 tiempos, es decir, 4 subgrupos) este análisis se aplicó únicamente sobre los pacientes pertenecientes a cada subgrupo clínico. Como en el segundo caso del primer estudio, en este segundo estudio el análisis por *sliding windows* de cada subgrupo de pacientes en específico se realizó aplicando solamente el escenario de análisis por haplotipos (sin tener en cuenta la frecuencia relativa de los haplotipos). Además, en el segundo estudio se calculó la diferencia de conservación entre los 2 subgrupos de cada grupo. Es decir, la diferencia de conservación entre las muestras de los mismos pacientes a T0 y T1. Este análisis se realizó comparando los valores de contenido de información entre estas muestras.

Las regiones conservadas (tanto nucleotídicas como aminoacídicas) detectadas se representaron en forma de logos (R, paquete Rseq), una representación de la secuencia donde el tamaño de cada letra nos indica el nivel de su conservación. En la Figura 27 se ejemplifica un logo de secuencia aminoacídica.

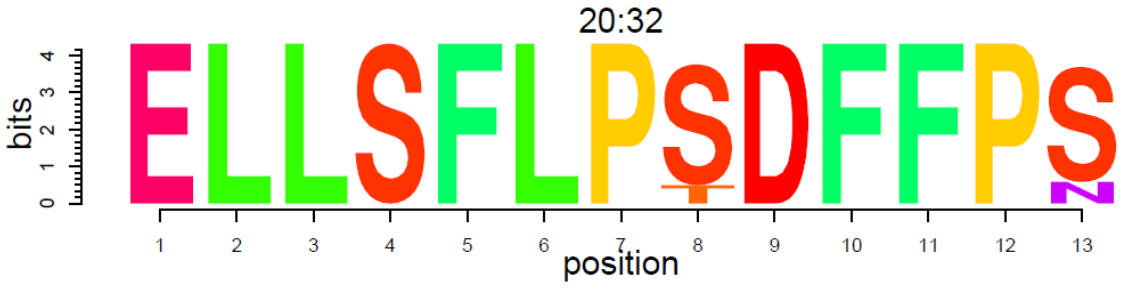


Figura 27. Ejemplo de logo de secuencia aminoacídica. Representación como logo de secuencia aminoacídica de una región de HBc. Las posiciones de la región están indicadas en la parte superior del logo. Los tamaños relativos de las letras en el logo indican sus frecuencias relativas en cada posición. La altura de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (4,32, 100% de conservación en aa).

4.7 Estudio de mutaciones

Con el fin de detectar mutaciones que pudieran relacionarse con un estado clínico especifico, los haplotipos de cada paciente se alinearon en su misma región con una secuencia consenso del genotipo correspondiente obtenida al alinear todas las secuencias de ese mismo genotipo del grupo de 106 secuencias utilizadas anteriormente en la determinación de los genotipos (como se ha descrito en el apartado 4.5: Genotipado de los haplotipos).

A nivel nucleotídico, se estudió la presencia de inserciones (Ins) y deleciones (Del) que, al traducirse, pudieran determinar una alteración en la secuencia proteica de HBc. Se reportó la presencia de inserciones o deleciones de longitud limitada (InsDels) que pudieran cambiar la pauta de lectura del gen *HBC* y también de macrodeleciones (deleciones de tamaño consistente).

A nivel aminoacídico se tuvieron en cuenta aquellas mutaciones no sinónimas que determinaran sustituciones del aa codificado en una posición específica respecto a la secuencia consenso del genotipo correspondiente.

En el análisis de las sustituciones aminoacídicas se descartaron los polimorfismos, es decir, cambios específicos en la secuencia del genoma viral que a lo largo del tiempo se han fijado en una parte de la población, y que por lo tanto no se han de considerar como mutaciones, sino que forman parte de la secuencia *wild-type* en aquella población en particular. Con este fin, los haplotipos de cada paciente se alinearon con la secuencia consenso de la población

total de cada estudio (obtenida a partir de todos los haplotipos detectados en todos los pacientes analizados para cada estudio). Sólo aquellos cambios que se confirmaron respecto a la secuencia consenso de la población total se consideraron mutaciones reales y como tales, se estudiaron.

4.8 Estudio de la complejidad de la quasiespecies

La complejidad de una QS se define como la propiedad que cuantifica la diversidad y la frecuencia de los haplotipos independientemente del tamaño de la población viral (111). Esta complejidad puede influir en el potencial patogénico y en la evolución clínica, es por eso por lo que su estudio es importante.

Este análisis se realizó sobre los datos del segundo estudio con el fin de detectar diferencias al comparar los valores de complejidad de la QS obtenidos para los diferentes subgrupos, que representan a pacientes en dos tiempos distintos de su enfermedad hepática causada por la hepatitis crónica B.

La complejidad de una QS se puede calcular con los índices de diversidad usados en el campo de la Ecología (111,113). Para ello, se recomienda un análisis multivariado, es decir, examinar el mismo conjunto de datos usando diferentes índices (111). Estos se clasifican en índices de incidencia, de abundancia y de función.

- Incidencia

Son los índices que corresponden al total de entidades en el alineamiento múltiple de los haplotipos. Son índices de riqueza, de número de especies en una comunidad (111).

En el proyecto se ha analizado el índice de incidencia de **número de haplotipos (*number of haplotypes,* nº Hpl)** obtenidos en cada muestra.

- Abundancia

Estos índices consideran las entidades observadas y también su frecuencia en la población. Miden diversidad o uniformidad de la distribución de haplotipos (111).

Los índices de abundancia más comúnmente usados y que se han estudiado en el proyecto son:

Entropía de Shannon normalizada (*Shannon entropy*, H_{SN}): mide la diversidad o uniformidad de una QS cuando es normalizada a la diversidad máxima. Permite medir la diversidad de la QS teniendo en cuenta la cantidad de haplotipos o variantes que se han encontrado y la frecuencia relativa entre ellos. Varía entre 0 (cuando todas las variantes son idénticas) y 1 (cuando todas son diferentes), pero no tiene en cuenta la heterogeneidad de las diferentes variantes de la QS viral porque es insensible al número de mutaciones (111).

$$H_{SN} = -\sum_{i=1}^{H} p_i log p_i / \log(H) *$$

**pi: frecuencia del haplotipo i en la población de la QS. H: número de Haplotipos en la población de la QS.*

Gini-Simpson (H_{GS}): dentro de un mismo genoma, mide la probabilidad de que dos individuos seleccionados al azar pertenezcan a haplotipos diferentes. Varía entre 0 (cuando todas las variantes son idénticas) y 1 (cuando todas son diferentes) (111).

$$H_{GS} = 1 - \sum_{i=1}^{H} {p_i}^2 *$$

** pi: frecuencia del haplotipo i en la población de QS.*

- Función

Los índices funcionales se basan en las diferencias entre haplotipos dentro de la QS viral y pueden tener en cuenta o no la frecuencia de cada uno de ellos. Miden la heterogeneidad intrapoblacional, es decir, cuan son de diferentes miembros de la misma población. Los índices funcionales se pueden basar en la incidencia (total de diferencias) o en la abundancia (frecuencia de las diferencias) (111). En el proyecto se han estudiado 3 índices funcionales:

Frecuencia de mutación (*Mutation Frequency*, Mf): índice funcional de incidencia. Mide la heterogeneidad genética respecto a la secuencia consenso de la QS. Es decir, mide la fracción de nts en la alineación de los haplotipos que difieren del haplotipo dominante de la QS. Cuanto mayor es el valor obtenido para este índice más diferentes son los individuos de la población con respecto a esta secuencia consenso (111).

$$Mf = \frac{1}{H}\sum_{i=1}^{H} d_{1i} *$$

**d1i: Proporción de mutaciones en el haplotipo i relacionadas con la secuencia consenso.*

Frecuencia media de mutaciones por molécula (*Average mutation frequency by molecule*, Mfm): índice funcional de abundancia. Mide la media de mutaciones detectadas por nt secuenciado respecto al haplotipo dominante teniendo en cuenta la frecuencia de estas mutaciones. Cuanto mayor es el valor, mayor es la tasa de mutación a nivel molecular para la QS viral (111).

$$\mathrm{Mfm} = \sum_{i=1}^{H} \hat{p}_i d_{1i} *$$

**pi: frecuencia del haplotipo i en la población de QS, d1i: Proporción de mutaciones en el haplotipo i relacionadas con la secuencia consenso.*

Diversidad nucleotídica (*Nucleotide diversity,* π): índice funcional de abundancia. Mide la heterogeneidad genética global gracias al cálculo de la media del número de mutaciones existentes entre cada posible pareja de haplotipos de la QS. Esto corresponde al cálculo del promedio de nt diferentes entre dos secuencias cualesquiera de la QS viral (111).

$$\pi = \sum_{i=1}^{H} \sum_{j=1}^{H} d_{ij} p_j *$$

** pi: frecuencia del haplotipo i en la QS viral, pj: frecuencia del haplotipo j en la QS viral, dij: distancia de Hamming (número de mutaciones que diferencia el haplotipo i del j.*

Como se ha comentado al inicio del apartado, estos índices se calcularon, analizaron y compararon para los datos obtenidos en el segundo estudio de esta tesis doctoral.

4.9 Análisis estadístico

Las diferencias de conservación entre los diferentes grupos clínicos observadas en los análisis de *sliding windows* se analizaron aplicando el test Wilcoxon–Mann–Whitney.

Las frecuencias de las diferentes mutaciones de cada paciente se calcularon sumando la frecuencia relativa de los haplotipos que presentaban tales alteraciones. Las frecuencias de cada grupo se obtuvieron calculando la mediana con su rango intercuartil (IQR 25-75) de todos los pacientes pertenecientes a cada grupo y se compararon implementando una prueba de Kruskal-Wallis asociada al test posthoc para comparaciones múltiples de Dunn.

Los diferentes índices de complejidad de la QS analizados se calcularon para cada subgrupo clínico del segundo estudio. Las frecuencias de estos índices para cada subgrupo se obtuvieron calculando la mediana con su rango intercuantil (IQR 25-75) de todos los pacientes pertenecientes a cada subgrupo y se compararon implementando una prueba de Kruskal-Wallis asociada al test posthoc para comparaciones múltiples de Dunn.

Todos los análisis se han llevado a cabo en el sistema R (versión 3.2.3) y los p-valores han sido corregidos con el método de Bonferroni. Los valores <0,05 se han considerado estadísticamente significativos.

RESULTADOS

5. RESULTADOS

5.1 Primer estudio: conservación y variabilidad en pacientes con hepatitis crónica B en diferentes estados clínicos

5.1.1 Pacientes de estudio y características

En el primer estudio de este proyecto de tesis doctoral, se reclutaron inicialmente 45 pacientes con hepatitis crónica por VHB en diferentes estados de la enfermedad hepática. De estos, solo en 38/45 ambos amplicones fueron amplificados y secuenciados con la adecuada calidad, por lo que se incluyeron en el estudio. Concretamente, en 6 de los 7 pacientes descartados no se consiguió amplificar ambos amplicones, mientras que en el séptimo paciente los amplicones no pasaron los filtros bioinformáticos.

Estos 38 pacientes se agruparon según sus características clínicas (como anteriormente se ha explicado en el apartado 4.1 de Materiales y métodos: Pacientes y muestras): 16 pacientes con hepatitis crónica por VHB sin daño hepático (grupo CHB) y 22 pacientes con hepatitis crónica por VHB y diagnóstico, por imagen o por biopsia, de daño hepático (5 pacientes con cirrosis hepática, grupo LC, y 17 con carcinoma hepatocelular, grupo HCC).

De cada uno de estos pacientes se analizó una muestra de suero. En el caso de los pacientes incluidos en los grupos HCC y LC, se consideró la muestra con la fecha más próxima al diagnóstico de la lesión hepática en concreto. Las características clínicas y virológicas de estos pacientes se muestran en la Tabla 2.

Mediana (IQR 25-75)	CHB (n=16)	HCC (n=17)	LC (n=5)
Edad	38,5 (33,5-46,5)	67 (58-69)	56 (48-66)
Carga viral (log IU/mL)	6,8 (5,7-8,0)	5,5 (4,7-6,7)	5,7 (4,8-6,2)
ALT	56,5 (41,25-180,5)	70 (47-212)	46 (43-79)
AST	56 (34,75-124)	120,5 (59-163,5)	66,45 (48,675-84,225)
Plaquetas (10^9/L)	183 (161,5-226)	136 (98,5-255)	81,5 (61,25-101,75)
Proporción			
Género (masculino)	11 / 16	15 / 17	3 / 5
HBeAg (positivo)	8 / 16	3 / 17	0 / 5

Tabla 2. Características clínicas y virológicas de los pacientes con hepatitis crónica por VHB incluidos en cada grupo del primer estudio. ALT = alanina aminotransferasa; AST = aspartato aminotransferasa; IQR = rango intercuartil.

5.1.2 Resultados de la secuenciación NGS

Una vez aplicados los filtros bioinformáticos de calidad pertinentes se obtuvieron un total de 45.214.965 *reads* para el amplicón 1 y 62.354.415 *reads* para el amplicón 2. Estos resultados se traducen a una mediana [IQR 25-75] por paciente de 133.156,5 [85.961,25-605.212] y 66.571 [25.958,5-2.301.225] *reads* para los amplicones 1 y 2 respectivamente.

5.1.3 Resultados del genotipado

Como se reporta anteriormente en el apartado 1.1.4 de Introducción: Genotipos del VHB, la evolución clínica y las características virológicas están altamente relacionadas con los diferentes genotipos virales, por lo que un adecuado genotipado del virus a través del análisis por secuenciación masiva es esencial en el estudio del virus y de su variabilidad.

Los resultados del análisis filogenético (obtenidos como se detalla en el apartado 4.5 de Materiales y métodos: Genotipado de los haplotipos) se muestran en la Tabla 3.

Se detectaron 5 genotipos (A, C, D, E y F) y 2 mezclas de genotipos (D/E y D/A) en los pacientes analizados. El genotipo D fue el más prevalente, habiéndose detectado en 17 de los 38 pacientes. El genotipo C se detectó en 8/38 pacientes, el genotipo A en 5/38. Los menos representados fueron los genotipos E y F que se detectaron en 2/38 pacientes

respectivamente. La mezcla de genotipos D/E se detectó en 3/38 pacientes mientras que la mezcla de genotipos D/A solamente en 1/38.

Porcentaje de pacientes (n)			
Genotipo	**CHB (n=16)**	**HCC (n=17)**	**LC (n=5)**
A	18.75 (3)	5.88 (1)	20.0 (1)
C	37.5 (6)	5.88 (1)	20.0 (1)
D	25.0 (4)	64.71 (11)	40.0 (2)
E	6.25 (1)	0.0 (0)	20.0 (1)
F	6.25 (1)	5.88 (1)	0.0 (0)
D/E	6.25 (1)	11.77 (2)	0.0 (0)
D/A	0.0 (0)	5.88 (1)	0.0 (0)

Tabla 3. Distribución de los genotipos en los grupos clínicos del primer estudio. La tabla muestra el porcentaje de pacientes con cada genotipo dentro de los grupos clínicos. El número de pacientes con cada genotipo está reportado entre paréntesis (n). D/E y D/A indican la mezcla de los dos genotipos.

El hecho de que se haya detectado en un mismo paciente la presencia de más de un genotipo (mezclas de genotipos D/E y D/A) es indicativo de un fenómeno de recombinación intergenotípica. Éste fenómeno resulta de la coinfección de un mismo huésped con distintas poblaciones virales de diferentes genotipos o subgenotipos que intercambian su material genético dentro de los hepatocitos infectados.

5.1.4 Análisis de la conservación

Con el fin de evidenciar, tanto en el gen (*HBC)* como en la secuencia proteica (HBc) de estudio regiones altamente conservadas en todos los pacientes y también regiones que estuvieran diferentemente conservadas entre los distintos grupos clínicos de estudio, la conservación se estudió aplicando un análisis de *sliding windows* a nivel nucleotídico y aminoácido como se ha explicado anteriormente en el apartado 4.6 de Materiales y métodos: Análisis de la conservación.

5.1.4.1 Detección de regiones hiperconservadas en la población total del estudio.

Para identificar regiones hiperconservadas tanto en el gen *HBC* como en su correspondiente secuencia proteica, HBc, la conservación se estudió teniendo en cuenta los dos escenarios descritos en el apartado 4.6 de Materiales y métodos: Análisis de la conservación. Es decir, o considerando las mismas frecuencias para todos los haplotipos (análisis por haplotipos) o teniendo en cuenta la frecuencia relativa de cada uno de los haplotipos (análisis por frecuencia de haplotipos).

Para este análisis se consideró el alineamiento múltiple de todos los haplotipos de los 38 pacientes sin considerar su estado clínico.

Regiones hiperconservadas en el gen *HBC*

No se detectaron diferencias entre analizar el contenido de información a nivel nucleotídico sin tener en cuenta (análisis por haplotipos, línea lila en la Figura 28) o teniendo en cuenta la frecuencia relativa de los haplotipos (análisis por frecuencia de haplotipos, línea naranja en la Figura 28).

Se evidenciaron 3 regiones nucleotídicas hiperconservadas: entre los nt 1900-1929 (región que incluye el codón de inicio de expresión del gen *HBC*), entre los nt 2249-2284 (región en la que encontramos epítopos T CD8+ al ser traducida a aa) y entre los nt 2364-2398 (que equivale a una región rica en arginina del CTD cuando se traduce a aa) (Figura 28).

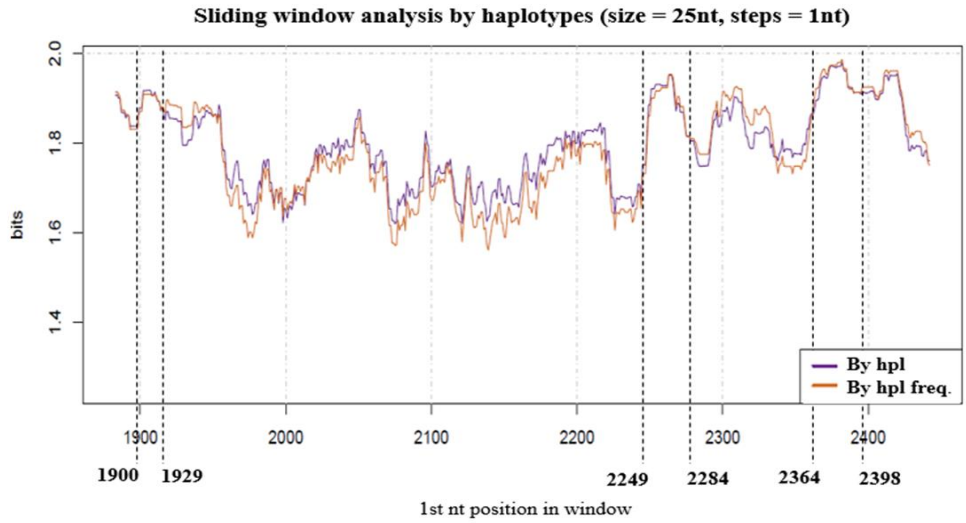


Figura 28. Análisis del contenido de información en la población total de estudio a nivel de nt. El análisis de *sliding windows* del gen *HBC* se ha realizado alineando los haplotipos de las QS de los 38 pacientes con y sin considerar la frecuencia relativa de los haplotipos. Cada punto en el gráfico representa el promedio del contenido de información (en bits) de las ventanas de 25 nt con desplazamiento de 1 nt hacia adelante entre cada ventana. La línea lila muestra el análisis por haplotipos (By hpl.), mientras que la línea naranja representa el análisis de haplotipos teniendo en cuenta sus frecuencias relativas (By hpl freq.). Las líneas negras discontinuas indican las 3 regiones hiperconservadas comunes observadas y se reportan sus posiciones.

Las tres regiones hiperconservadas detectadas mostraron valores de contenido de información altos (mayoritariamente en torno a 2 bits, lo que corresponde al 100% de conservación en secuencias de nt), como se muestra en la Figura 29.

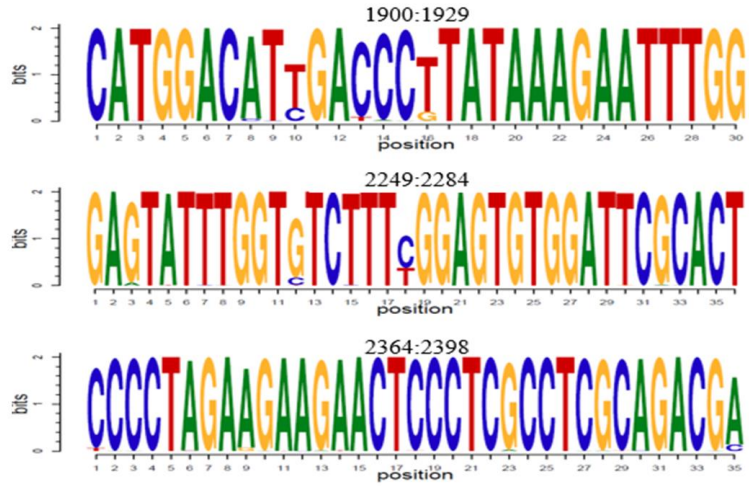


Figura 29. Logos de las regiones hiperconservadas en la población total a nivel de nt. Representación como logos de las regiones hiperconservadas comunes detectadas en la población total de estudio a nivel de secuencia nucleotídica. Los extremos de cada región están indicados en la parte superior de cada logo. Los tamaños relativos de las letras en los logos son representativos de sus frecuencias relativas en cada posición dentro del alineamiento múltiple de los haplotipos. La altura total de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (2 bits, correspondiente a 100% de conservación en secuencias de nt).

- Regiones hiperconservadas en la secuencia proteica de HBc

La presencia de regiones hiperconservadas a nivel de aa se analizó calculando el contenido de información sobre el alineamiento de todos los haplotipos de los 38 pacientes sin considerar su pertenencia a un grupo clínico concreto. Nuevamente, el análisis se hizo tanto teniendo en cuenta la frecuencia relativa de los haplotipos como sin tenerla en cuenta.

Los haplotipos aminoacídicos se obtuvieron de la traducción de sus respectivos haplotipos nucleotídicos usando la pauta de lectura del gen *HBC*.

Como ya se vio a nivel nucleotídico, a nivel aminoacídico tampoco se detectaron diferencias entre analizar la conservación por haplotipos o por frecuencia de haplotipos (Figura 30).

La secuencia aminoacídica de HBc mostró una elevada conservación a lo largo de toda su secuencia a excepción de una región central, entre los aa 50-100, región que engloba al MIR (aa 78-82; *Major Immunodominant Region*), en la que el contenido de información decrece hasta un máximo de 0,6 bits respecto a las posiciones de su alrededor (Figura 30).

Se detectaron 2 regiones hiperconservadas a nivel de aa: la primera entre los aa 117-120 y la segunda entre los aa 159-167 (Figura 30). Estas dos regiones aminoacídicas coinciden a nivel de posición con dos de las tres regiones nucleotídicas hiperconservadas previamente descritas. Concretamente la región aa 117-120 coincide con la región nt 2249-2284 (región relacionada con epítopos T CD8+) y la región aa 159-167 coincide con la región nt 2364-2398 (región rica en argininas).

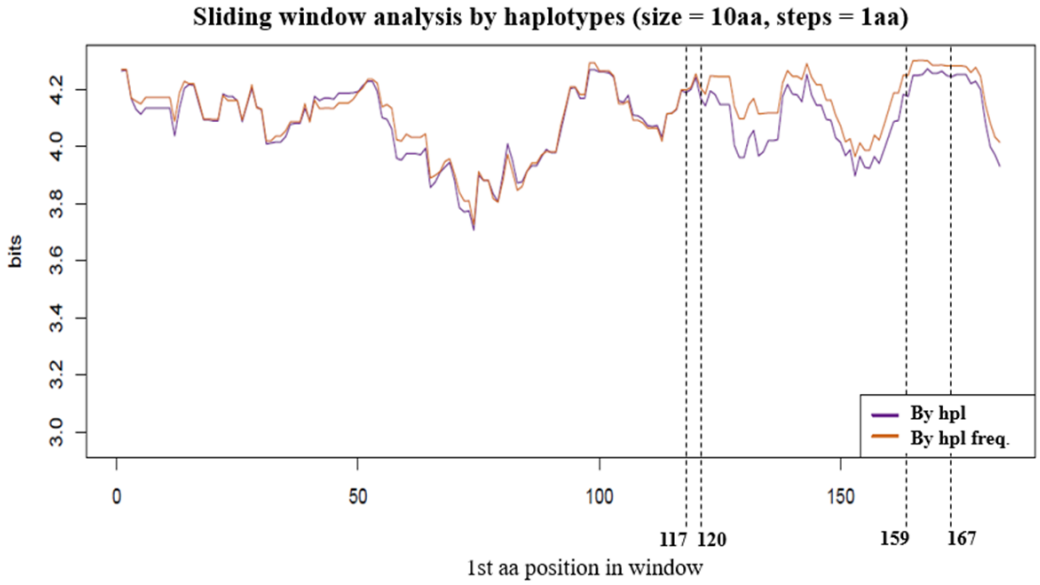


Figura 30. Análisis del contenido de información en la población total de estudio a nivel de aa. El análisis de *sliding windows* de la secuencia aminoacídica de HBc se ha realizado alineando los haplotipos de las QS de los 38 pacientes con y sin considerar la frecuencia relativa de los haplotipos. Cada punto en el gráfico representa el promedio del contenido de información (en bits) de las ventanas de 10 aa con desplazamiento de 1 aa hacia adelante entre cada ventana. La línea lila muestra el análisis por haplotipos (By hpl.), mientras que la línea naranja representa el análisis de haplotipos teniendo en cuenta sus frecuencias relativas (By hpl freq.). Las líneas negras discontinuas indican las 2 regiones aminoacídicas hiperconservadas comunes observadas y se reportan sus posiciones.

Como muestra la Figura 31, todas las posiciones de estas dos regiones aminoacídicas hiperconservadas mostraron valores de contenido de información en torno a 4,32 bits (100% de conservación en secuencias de aa).

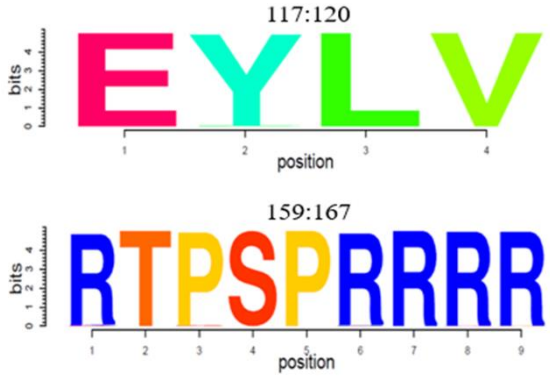


Figura 31. Logos de las regiones hiperconservadas en la población total a nivel de aa. Representación como logos de las regiones hiperconservadas comunes detectadas en la población total de estudio a nivel de secuencia aminoacídica. Los extremos de cada región están reportados en la parte superior de cada logo. Los tamaños relativos de las letras en los logos indican sus frecuencias relativas en cada posición dentro del alineamiento múltiple de los haplotipos. La altura de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (4,32, 100% de conservación en secuencias de aa).

5.1.4.2 Análisis de la conservación entre grupos: Regiones conservadas específicas de grupo.

En el estudio de la conservación específica de cada grupo se tuvo en cuenta sólo el escenario en el cual todos los haplotipos se consideraban con igual frecuencia (análisis por haplotipos, sin tener en cuenta su frecuencia relativa). De hecho, al igualar las frecuencias de todos los haplotipos se obtiene una información más detallada de la conservación de las QS circulantes al dar más peso a secuencias minoritarias. Además, como se explica en el apartado anterior, no se detectaron diferencias a la hora de analizar la conservación considerando o no la frecuencia relativa de los haplotipos tanto a nivel nucleotídico como aminoacídico.

- Diferencias de conservación entre los grupos clínicos en el gen *HBC*

Al comparar el contenido de información del gen *HBC* obtenido para cada uno de los grupos clínicos de estudio se observó que los pacientes de los grupos HCC y LC presentaban un patrón de conservación muy similar entre sí, a diferencia de los pacientes pertenecientes al grupo CHB, los cuales mostraban un nivel de conservación inferior. Concretamente, el grupo CHB se encontraba menos conservado en 5 regiones nucleotídicas específicas: nt 1946-1992, 2060-2095, 2145-2175, 2230-2250, y 2270-2293 (p-valor <0.05, marcadas en rojo en la Figura 32).

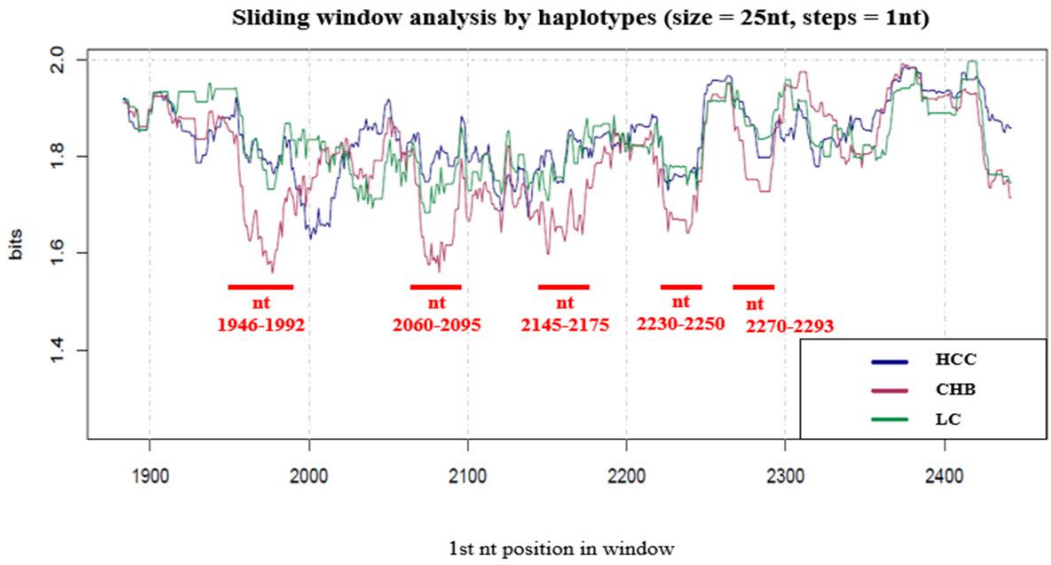


Figura 32. Conservación del gen *HBC* en los diferentes grupos clínicos. Análisis de *sliding windows* por haplotipos del gen *HBC* en los alineamientos múltiples pertenecientes a cada grupo clínico (HCC en azul, CHB en rojo y LC en verde). Las 5 regiones en las que el grupo CHB muestra niveles más bajos de conservación (p-valor<0.05 hecho por el test de Wilcoxon-Mann-Whitney) se muestran en rojo y se reportan sus posiciones.

Se detectaron algunas regiones nucleotídicas conservadas específicas de grupo (Figura 33). En concreto, se identificó una región nucleotídica conservada específica del grupo CHB entre los nt 2306-2334 (región que al ser traducida incluye a los 5 primeros aa de la región bisagra entre NTD y CTD). En los pacientes LC se identificaron otras dos regiones nucleotídicas conservadas específicas: la región nt 1935-1976, que al ser traducida incluye una región involucrada en el ensamblaje de la cápside y en la producción de viriones (134) y la región nt 2402-2435 que al ser traducida corresponde a una región rica en argininas del CTD.

Como muestran los logos de secuencia (Figura 33), la mayoría de las posiciones nucleotídicas de estas 3 regiones conservadas específicas de grupo presentaban valores de contenido de información en torno a 2 bits (100% de conservación en secuencias de nt).

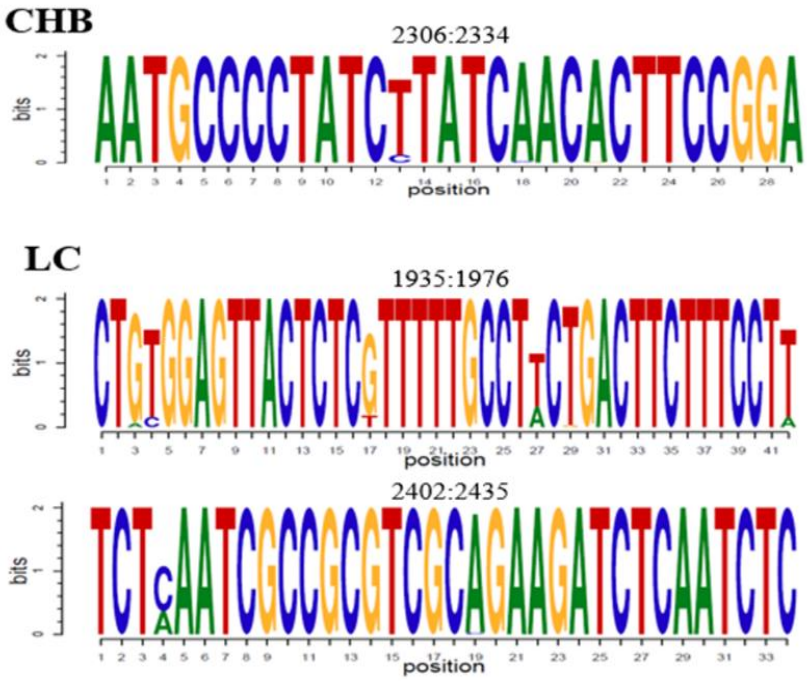


Figura 33. Regiones nucleotídicas conservadas específicas de los grupos clínicos. Representación como logos de secuencia nucleotídica de las regiones conservadas específicas identificadas en cada grupo clínico. Los extremos de cada región están indicados en la parte superior de cada logo. Los tamaños relativos de las letras en los logos son representativos de sus frecuencias relativas en cada posición dentro del alineamiento múltiple de los haplotipos. La altura total de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (2 bits, correspondiente a 100% de conservación en nt).

- Diferencias de conservación entre los grupos clínicos en la secuencia proteica de HBc

Al comparar el contenido de información a nivel de secuencia aminoacídica entre los grupos clínicos, se observó un patrón de conservación similar entre los 3 grupos a lo largo de toda la secuencia a excepción de una región específica entre los aa 140-160 (marcada en verde en la Figura 34) en la que el grupo LC mostró un nivel de conservación inferior a los obtenidos para los grupos CHB y HCC (p-valor <0.05).

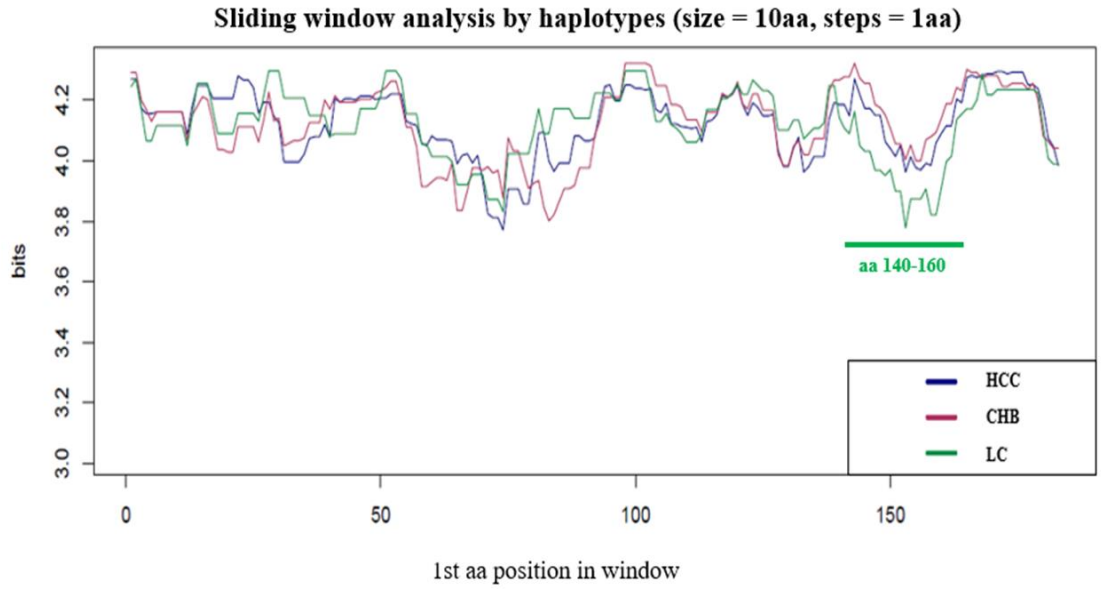


Figura 34. Conservación de la secuencia aminoacídica de HBc en los diferentes grupos clínicos. Análisis de *sliding windows* por haplotipos de la secuencia de aa de HBc en los alineamientos múltiples pertenecientes a cada grupo (HCC en azul, CHB en rojo y LC en verde). La región en la que el grupo LC muestra niveles más bajos de conservación (p-valor<0.05 hecho por el test de Wilcoxon-Mann-Whitney) se muestra en verde y se reportan sus posiciones.

Como ya se observó a nivel nucleotídico, también a nivel aminoacídico se detectaron regiones conservadas específicas de un grupo clínico concreto. Se evidenció una región aminoacídica conservada específicamente en el grupo CHB (aa 98-103) y otras dos en el grupo LC (aa 28-30, región que incluye una zona involucrada en el ensamblaje de la cápside y en la producción de viriones (134) y aa 51-54) (Figura 35).

Todas las posiciones de estas tres regiones aminoacídicas mostraron valores de contenido de información en torno a 4,32 bits (100% de conservación en secuencias aminoacídicas) como muestra la Figura 35.

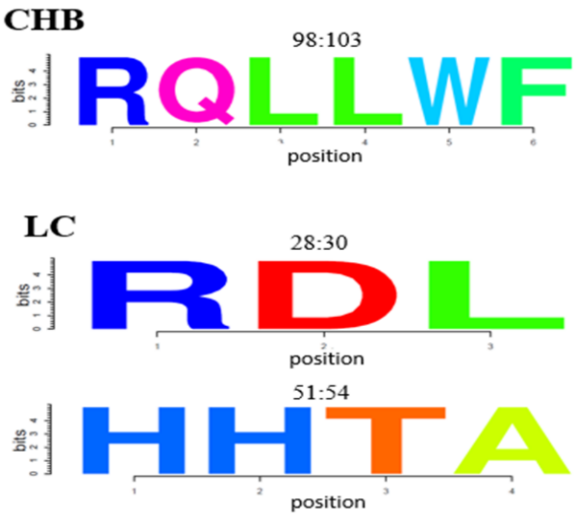


Figura 35. Regiones aminoacídicas conservadas específicas de los grupos clínicos. Representación como logos de secuencia aminoacídica de las regiones conservadas específicas de grupo detectadas. Los extremos de cada región están indicados en la parte superior de cada logo. Los tamaños relativos de las letras en los logos indican sus frecuencias relativas en cada posición dentro del alineamiento múltiple de los haplotipos. La altura de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (4,32, 100% de conservación en aa).

5.1.5 Estudio de mutaciones en los diferentes grupos clínicos

El estudio de mutaciones (detallado en el apartado 4.7 de Materiales y métodos: Estudio de mutaciones) se llevó a cabo con el fin de identificar cambios (tanto nucleotídicos como aminoacídicos) en los distintos grupos clínicos que pudieran afectar a la secuencia proteica de HBc y por lo tanto relacionarse con una diferente evolución clínica.

5.1.5.1 Mutaciones nucleotídicas

A nivel nucleotídico el análisis se hizo con el fin de buscar inserciones o deleciones que, al ser traducidas, pudieran alterar la secuencia de HBc.

- Inserciones y deleciones puntuales (InsDels)

En 11 de los 38 pacientes incluidos en el estudio se detectó por lo menos una InsDel de 1 nt que afectaba concretamente a las posiciones nt 1951 o 2085, con la inserción o deleción de un residuo de Timina (T) o Guanina (G) en cada posición respectivamente. Los pacientes en los que se detectaron estas InsDels, así como las frecuencias y los porcentajes de los haplotipos que las contenían se reportan en la Tabla 4.

Debido al limitado número de pacientes del estudio no se hallaron diferencias estadísticas significativas al comparar la frecuencia de las InsDels entre los grupos clínicos de estudio.

Grupo clínico (n/total)	Paciente	Posición de la InsDel (nt insertado/delecionado)	Frecuencia relativa	% de haplotipos mutados
CHB (8/16)	1	1951 (1 nt: T)	8.36	8.7%
	2	2085 (1 nt: G)	17.12	40%
	3	2085 (1 nt: G)	3.19	5%
	4	1951 (1 nt: T)	0.37	5.9%
	9	1951 (1 nt: T)	2.02	8.82%
	10	2085 (1 nt: G)	1.34	50%
	12	2085 (1 nt: G)	1.04	10%
	13	1951 (1 nt: T)	2.79	22.22%
HCC (2/17)	28	2085 (1 nt: G)	0.78	4%
	33	2085 (1 nt: G)	2.42	4.8%
LC (1/5)	34	2085 (1 nt: G)	17.42	19.2%

Tabla 4. InsDels detectadas. La tabla muestra la frecuencia relativa de las InsDels detectadas en las posiciones nt 1951 y 2085, junto con el porcentaje de haplotipos mutados por cada paciente. Solo los pacientes portadores de estas mutaciones se han incluido en la tabla. T = timina; G = guanina.

En todos los casos esta inserción o deleción de 1 nt en estas posiciones nucleotídicas concretas provocaba un cambio en la pauta de lectura del gen ocasionando diferentes truncamientos de la secuencia de HBc debido a la aparición de un codón STOP prematuro.

Concretamente, en la posición nt 1951, la inserción de una T producía el truncamiento de la secuencia proteica en el aa 21 mientras que la deleción de una T en esa misma posición producía el truncamiento en el aa 41. En el caso de la posición nt 2085, la inserción de una G producía el truncamiento de la secuencia proteica en el aa 74 mientras que el truncamiento se producía en la posición aa 64 si se trataba de la deleción de una G en nt 2085. En los 4 truncamientos comentados en este párrafo, las posiciones aminoacídicas indicadas (21, 41, 64 y 74) corresponden al último aa traducido correctamente antes de la aparición del codón STOP prematuro. En todos los casos indicados, el truncamiento generaba la aparición de una proteína carente del dominio CTD (aa 150-183).

- Macrodeleciones

En 3/38 pacientes (concretamente un paciente de cada grupo clínico) se evidenciaron macrodeleciones (deleciones de alrededor de 50 o más nt) en diferentes posiciones del gen *HBC*.

Estas macrodeleciones afectaban a más del 25% de los haplotipos en cada respectivo paciente: al 43,5% de los haplotipos del paciente CHB (paciente nº 1), al 26.78% del paciente HCC (paciente nº23) y al 49.97% de los haplotipos del paciente LC (paciente nº 34). En general estas mutaciones nucleotídicas presentaban una frecuencia relativa elevada en estos pacientes. Concretamente, de 22,46 en el paciente CHB, de 22,61 en el paciente HCC y de 41,53 en el paciente LC. Las posiciones, frecuencias y porcentaje de haplotipos mutados se detallan en la Tabla 5.

Paciente; grupo clínico	**Último nt antes de la macrodeleción**	**Nº de nts delecionados**	**Aa eliminados o truncamientos al traducir la secuencia**	**Frecuencia relativa**	**% de haplotipos mutados**
Pt 1; CHB	2160	57	88 - 106	4.58	4.35%
	2151	78	84 - 109	0.79	4.35%
	2151	81	85 - 111	0.98	4.35%
	2135	87	79 - 107	5.29	4.35%
	2146	90	84 - 113	4.04	4.35%
	2149	105	86 - 120	1.28	4.35%
	2140	114	81 - 118	1.11	4.35%
	2149	114	84 - 121	0.84	4.35%
	2140	123	81 - 121	1.73	4.35%
	2088	144	64 - 111	1.82	4.35%
Pt 23; HCC	2163	129	89 - 131	22.61	26.78%
Pt 34; LC	2005	46	*49	31.31	26.92%
	2000	52	*47	6.07	7.69%
	2164	87	89 - 117	1.12	3.84%
	2123	103	*81	1.39	3.84%
	2130	101	*98	0.72	3.84%
	2123	63	*94	0.92	3.84%

Tabla 5. Macrodeleciones detectadas en el gen *HBC*. La tabla muestra la frecuencia relativa de las macrodeleciones detectadas. Las posiciones nucleotídicas afectadas, el número de nt eliminados en cada caso, los aa eliminados al traducirse la secuencia y el porcentaje de haplotipos mutados por cada paciente están reportados. La aparición de codón STOP prematuro se muestra con un asterisco y el número de la última posición aminoacídica codificada. Solo los pacientes portadores de estas mutaciones se han incluido en la tabla. Pt: paciente.

Cabe señalar que, diferentemente de lo que se vio con las InsDels, en general estas macrodeleciones no alteraban la pauta de lectura del gen *HBC*, por lo que se producían proteínas HBc carentes de la región aminoacídica codificada por la secuencia nucleotídica delecionada. Esto valía para todos los casos a excepción del paciente LC, donde 5/6 macrodeleciones se traducían a truncamientos, ocasionados por la aparición de codones STOP prematuros.

Debido a que este tipo de mutación afectaba solo a un paciente de cada grupo, no se halló ninguna diferencia estadísticamente significativa. En todos los casos las macrodeleciones afectaban a la región nucleotídica codificante del dominio NTD, resultando en formas alteradas de la secuencia proteica de HBc carentes de cierta región del NTD, pero con la región bisagra y el CTD intactos (a excepción de las 5 secuencias truncadas del paciente LC, en las que estos dos dominios se eliminaban).

5.1.5.2 Mutaciones aminoacídicas: sustituciones de aa

Los haplotipos aminoacídicos de cada paciente se alinearon con una secuencia consenso específica del genotipo del paciente en cuestión con el objetivo de identificar sustituciones aminoacídicas que se pudieran relacionar con un determinado cuadro clínico.

En 15/16 pacientes CHB, en 16/17 pacientes HCC y en 3/5 pacientes LC se detectaron sustituciones aminoacídicas que involucraban a 59 posiciones de los 183 aa de la secuencia proteica de HBc (Tabla 6).

Región de HBc (extremos aa)	Sustituciones aminoacídicas detectadas			
NTD (1-139)	I3L/T	P5T/L/S/H	T12S	V13L/A/T
	E14Q/D	S21P/L/H/A/G/T	S26N/A/T	I27V
	D32N/A	A34T/V	S35L/A/K	Y38H/F
	R39K/G	E40D	P45T/A/S	C48W/Y
	S49T/A	P50H/A	I59T/F/V/C/L	L60V
	G63E/V	E64K/D/N	M66T/I	T67S/A/N
	A69V/G	V74A/E/T/G	S74G	E77Q/D
	P79Q	A80S/V/E/Q/T/G	D83E	L84A/Q/R
	V85I/A	S87G/N/T	N92T/H	M93V/A
	I105T/V	T109M	R112K/S	E113D/Q/K/S
	T114I	I116V/L	R127H	P130L/T/Q/A/S/I
	A131P	P135Q/S/L/T		
Región bisagra (140-149)	T147C/A	V149I		
CTD (150-183)	R151G/C/Q	G153C/F/H/Y	S155T/A/F/L	R159K/G
	R164S/H	Q177K	A180G/Q/G	E180K/D/G/Q
	S181P/F/L	Q182K/C/P/H	S183P	

Tabla 6. Sustituciones aminoacídicas detectadas en la secuencia de HBc en la población total de estudio. La tabla muestra las 59 posiciones en las que se han detectado sustituciones aminoacídicas siendo la primera letra el aa *wild-type* de HBc (es decir, el aa en la secuencia consenso del genotipo), el número la posición aminoacídica involucrada y la segunda letra o grupo de letras el/los nuevos aa codificados. La primera columna indica el dominio de HBc (con las posiciones de sus extremos en la secuencia proteica de HBc reportadas) en el que se han identificado las diferentes sustituciones: NTD (dominio N terminal), Región bisagra y CTD (dominio C terminal).

En la región NTD (de 139 aa) se detectaron 46 posiciones modificadas de un total de 139 posiciones aminoacídicas que conforman la región, mientras que en la región bisagra (de 10 aa) y CTD (de 34 aa) los cambios afectaban respectivamente a 2 y a 11 posiciones. Al comparar la proporción de pacientes de cada grupo que presentaban cambios en cada uno de los dominios no se detectó ninguna diferencia estadísticamente significativa. Concretamente, 15/16 pacientes CHB, 16/17 HCC y 3/5 LC presentaban mutaciones en el NTD, 3 CHB, 5 HCC y 1 LC en la región bisagra, y 9 CHB, 13 HCC y 1 LC en el dominio CTD. No se detectó ninguna diferencia al comparar estas proporciones.

Al comparar la frecuencia observada de cada sustitución entre los diferentes grupos clínicos, se identificó la sustitución P79Q, que está presente en varios pacientes representativos de los 3 grupos de estudio (concretamente en 3/16 pacientes CHB, 9/17 pacientes HCC y 1/5 pacientes LC). En esta sustitución aminoacídica la prolina original de la posición 79 se sustituye por una glutamina. Esta sustitución se observó mayoritariamente en el grupo HCC (frecuencia mediana (IQR) de 15,82 (0-78,9)), mostrando significación estadística en la

diferencia respecto a los otros dos grupos (frecuencia mediana (IQR) de 0 (0-0) tanto en LC como en CHB, p-valor = 0,012 entre HCC y CHB) (Figura 36).

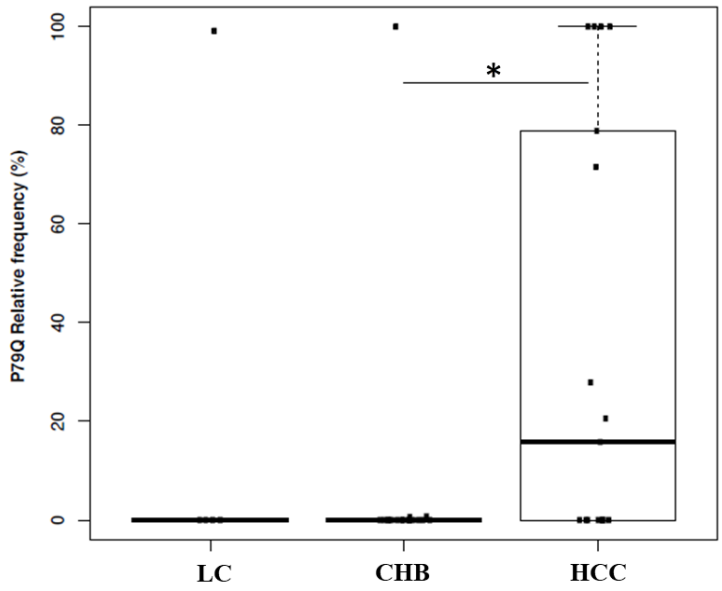


Figura 36. Frecuencia relativa (%) de la sustitución aminoacídica P79Q en los 3 grupos clínicos de estudio. Cada punto en el gráfico representa un paciente. El p-valor (corregido por Bonferroni) se ha calculado mediante la prueba de Kruskal-Wallis con la prueba de comparación múltiple posthoc Dunn. El p-valor estadísticamente significativo (<0.05) se representa con un asterisco (*).

5.2 Segundo estudio: conservación, variabilidad y complejidad de *HBC* en la progresión de la enfermedad hepática

5.2.1 Pacientes de estudio y características

En este segundo estudio del proyecto de tesis doctoral se analizó la evolución de la QS del VHB en el gen *HBC* en pacientes con hepatitis crónica antes y después de desarrollar hepatocarcinoma (HCC). Un grupo de pacientes con hepatitis crónica sin daño hepático (CHB), en el que no hubo progresión de la enfermedad, se utilizó como control. Inicialmente se seleccionaron 12 pacientes (6 CHB y 6 HCC) y se analizaron dos muestras por paciente (como se detalla en el apartado 4.1 de Materiales y métodos: Pacientes y muestras). Finalmente, los amplicones de ambas muestras pudieron ser amplificados y secuenciados con la calidad necesaria en solamente 9 de los 12 pacientes.

Estos 9 pacientes se agruparon según sus características clínicas (como se detalla en el apartado 4.1 de Materiales y métodos: Pacientes y muestras): 4 pacientes con hepatitis crónica por VHB sin daño hepático (grupo CHB) y 5 pacientes con hepatitis crónica por VHB y diagnóstico, por imagen o por biopsia, de carcinoma hepatocelular (grupo HCC). De cada uno de ellos se analizaron 2 muestras (T0 y T1) a modo de estudio longitudinal resultando en un total de 18 muestras.

Las muestras T0 y T1 de cada paciente presentaban una diferencia de tiempo mínima de un año entre ellas. En el caso del grupo HCC, las muestras T1 correspondían a las muestras con lesión tumoral, siendo estas las muestras con la fecha más próxima al diagnóstico del cáncer. Por otro lado, tanto las muestras a T0 del grupo HCC como ambas muestras (T0 y T1) del grupo control (CHB) correspondían a muestras sin daño hepático. Las características clínicas y virológicas de los pacientes de cada subgrupo se muestran en la Tabla 7.

Mediana (IQR 25-75)	**CHB T0 (n=4)**	**CHB T1 (n=4)**	**HCC T0 (n=5)**	**HCC T1 (n=5)**
Edad	44 (40-46,75)	45 (41-47,5)	67 (61-68)	69 (64-69)
Carga viral (log IU/mL)	7,7 (6,5-8,0)	7,7 (6,9-8,1)	3,1 (2,3-4,4)	4,7 (4,3-5,0)
ALT	73,5 (50,25-136)	142 (49-348,5)	24,8 (23-33,6)	61 (41,7-107)
AST	38,5 (33,75-79,75)	103,5 (34,75-201)	32,8 (32,3-35)	55 (53-129)
Plaquetas (10^9/L)	224 (212,5-257,25)	217 (202-237,5)	149 (129-165,4)	136 (131-140)
Proporción				
Género (masculino)	4/4	4/4	4/5	4/5
HBeAg (positivo)	3/4	3/4	1/5	1/5

Tabla 7: Características clínicas, bioquímicas y virológicas de los pacientes con hepatitis crónica por VHB incluidos en cada subgrupo del segundo estudio. ALT = alanina aminotransferasa; AST = aspartato aminotransferasa; IQR = rango intercuartil.

5.2.2 Resultados de la secuenciación NGS

Después de aplicar los filtros bioinformáticos de calidad pertinentes se obtuvieron un total de 21.967.662 *reads* para el amplicón 1 y 20.326.126 *reads* para el amplicón 2. Estos resultados se traducen a una mediana [IQR 25-75] por paciente de 144.978,5 [78.692,75-239.805,25] y 128.506,5 [57.558,25-194.366,25] *reads* para los amplicones 1 y 2 respectivamente.

5.2.3 Resultados del genotipado

Como se reporta anteriormente en el apartado 1.1.4 de Introducción: Genotipos del VHB, la evolución clínica y las características virológicas están altamente relacionadas con los diferentes genotipos virales, por lo que un adecuado genotipado del virus a través del análisis por secuenciación masiva es esencial en el estudio del virus y de su variabilidad.

Los resultados del análisis filogenético (obtenidos como se detalla en el apartado 4.5 de Materiales y métodos: Genotipado de los haplotipos) se muestran en la Tabla 8.

Contrariamente a lo detectado en el primer estudio, en este segundo no se evidenciaron mezclas de genotipos, sino que en cada paciente la QS era 100% de un genotipo en concreto. Los cuatro pacientes del grupo CHB estaban infectados por virus de genotipos diferentes, concretamente de los genotipos A, C, D y E, mientras que los 5 pacientes HCC estaban infectados por virus de genotipo D todos ellos. Los genotipos correspondientes se detectaron tanto en las muestras T0 como en las T1 de cada paciente.

Porcentaje de pacientes (n)		
Genotipo	**CHB (n=4)**	**HCC (n=5)**
A	25 (1)	0 (0)
C	25 (1)	0 (0)
D	25 (1)	100 (5)
F	25 (1)	0 (0)

Tabla 8. Distribución de los genotipos en los grupos clínicos del segundo estudio. La tabla muestra el porcentaje de pacientes con cada genotipo viral dentro de los grupos clínicos. El número de pacientes con cada genotipo dentro del grupo está reportado entre paréntesis (n).

5.2.4 Conservación y variabilidad en la progresión de la enfermedad hepática

La conservación tanto en el gen (*HBC)* como en la secuencia aminoacídica (HBc) de estudio se analizó en los cuatro subgrupos (HCC T0 y T1, y CHB T0 y T1) con el fin de evidenciar regiones diferentemente conservadas entre antes y después de desarrollar la lesión tumoral. La conservación se estudió aplicando un análisis de *sliding windows* a nivel nucleotídico y aminoacídico como se ha explicado anteriormente en el apartado 4.6 de Materiales y métodos: Análisis de la conservación, sobre los alineamientos múltiples de los haplotipos correspondientes a las muestras de cada uno de los 4 subgrupos. En este análisis en ningún caso se tuvieron en cuenta las frecuencias relativas de los haplotipos (todo se hizo por análisis por haplotipos), puesto que al igualar las frecuencias de todos los haplotipos se obtiene una información más detallada de la conservación de la QS circulantes al dar más peso a secuencias minoritarias.

5.2.4.1 Conservación del gen *HBC* en la progresión de la enfermedad hepática.

En el grupo CHB, en el que no se evidenciaba evolución de la lesión hepática entre ambas muestras de cada paciente, el patrón de conservación se mantenía constante en el año transcurrido entre T0 y T1. Diferentemente, los pacientes HCC T0 (cuando todavía no se evidenciaban signos de daño hepático) mostraron un nivel de conservación ligeramente más elevado que los subgrupos de CHB. No obstante, fueron los pacientes HCC T1 (es decir, aquellos que ya presentaban la lesión tumoral) los que mostraron un mayor grado de conservación, reflejándose en unos niveles de contenido de información notablemente más elevados que en los otros 3 subgrupos (Figura 37).

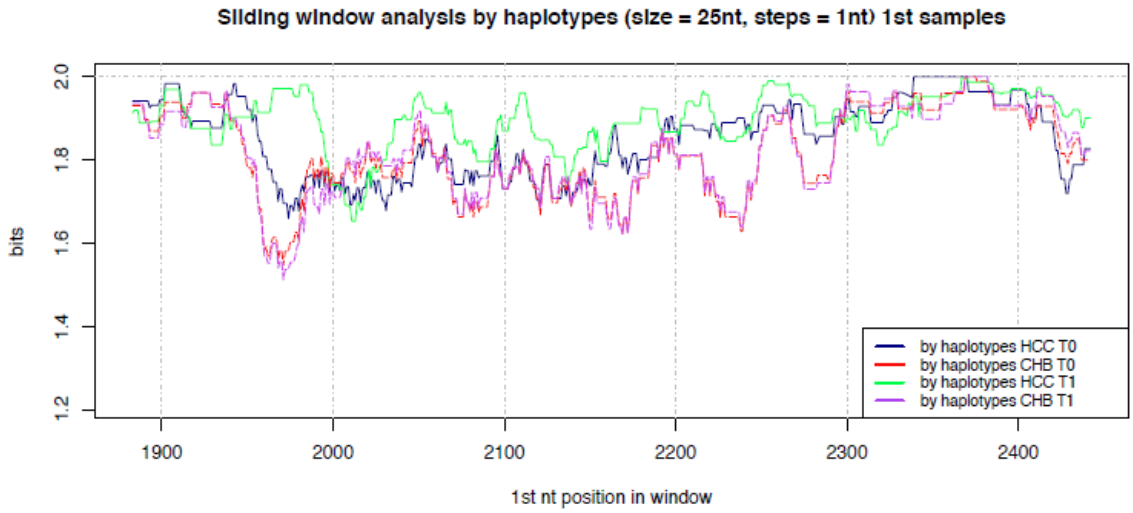


Figura 37. Conservación del gen *HBC* en los diferentes subgrupos clínicos. Análisis de *sliding windows* por haplotipos de la secuencia nucleotídica del gen *HBC* en los alineamientos múltiples de los pacientes de cada subgrupo. CHB se muestra con líneas punteadas, identificándose en rojo el subgrupo T0 y en lila el T1 mientras que HCC se muestra con líneas continúas identificándose en azul el subgrupo T0 (antes del desarrollo del HCC) y en verde el T1 (HCC ya diagnosticado).

Con el objetivo de evidenciar las regiones diferentemente conservadas en el grupo de HCC durante la progresión del daño hepático, se calculó la diferencia en el contenido de información a lo largo de toda la secuencia nucleotídica entre HCC T0 y HCC T1. Como se ha mencionado anteriormente, en las muestras T1 el nivel de conservación era mayor a lo largo de toda la secuencia con respecto a las muestras T0, pero lo era sobre todo en la región del gen entre los nt 1901-2318, región que codifica para el dominio proteico NTD, donde la diferencia de contenido de información (contenido de información T0 - contenido de

información T1) llegaba a valores negativos de hasta -0,3 bits (p-valor = 2,2E-16, Figura 38).

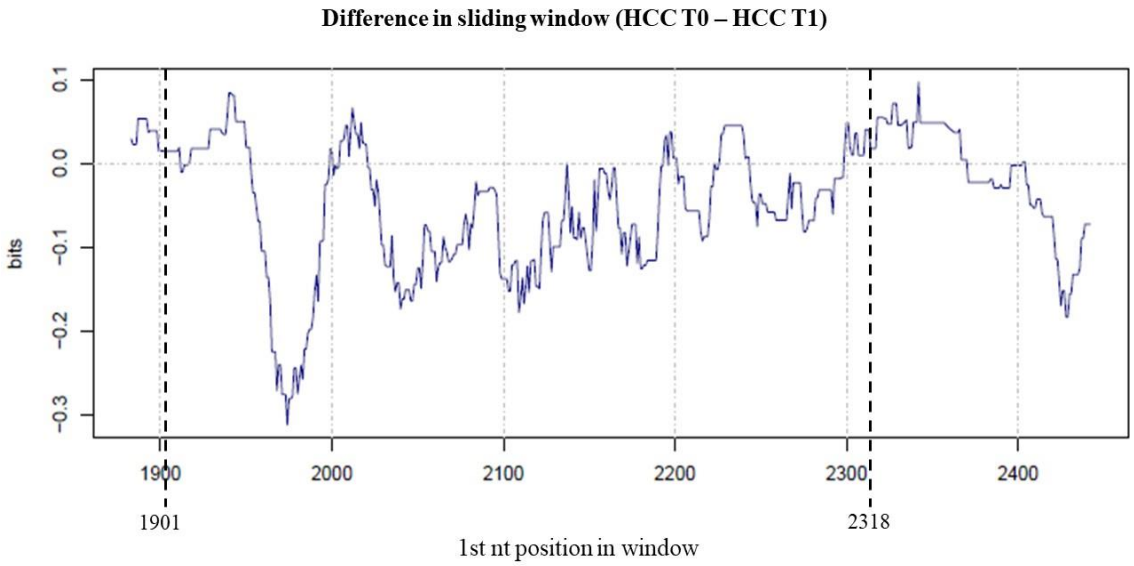


Figura 38. Diferencias en el contenido de información del gen *HBC* entre las muestras T0 y T1 de los pacientes del grupo HCC**.** El gráfico muestra la diferencia de contenido de información en el grupo HCC entre la muestra T0 (antes de desarrollar lesión tumoral) y la muestra T1 (diagnosis de HCC). Los valores negativos son indicativos de una mayor conservación al tiempo T1, mientras que los valores positivos indican un mayor contenido de información en las muestras T0. La región donde se encontró mayor diferencia a nivel de conservación nucleotídica está marcada y delimitada por líneas negras discontinuas y sus posiciones están reportadas en la parte inferior de estas (p-valor <0.05 por test de Wilcoxon-Mann-Whitney).

5.2.4.2 Conservación de la secuencia aminoacídica de HBc en la progresión de la enfermedad hepática.

Como también se ha visto a nivel nucleotídico, la secuencia aminoacídica de la proteína HBc se mantenía conservada con un patrón bastante similar entre CHB T0 y T1. De la misma forma, en ambos subgrupos de HCC la secuencia se mantenía conservada a lo largo de toda su longitud, con valores parecidos a los detectados en ambos subgrupos del grupo control, CHB (Figura 39).

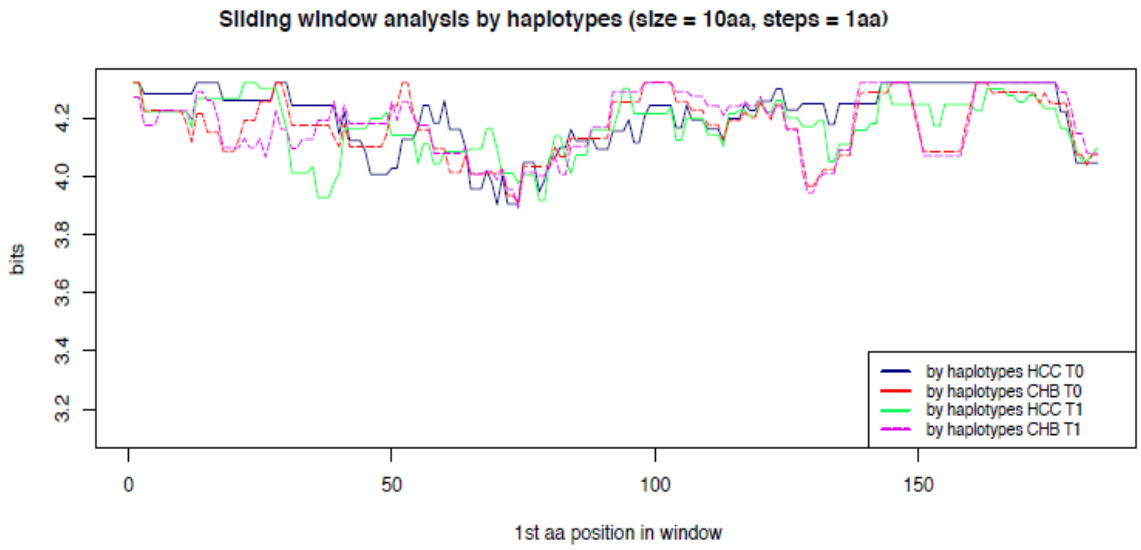


Figura 39. Conservación de la secuencia aminoacídica de HBc en los diferentes subgrupos clínicos. Análisis de *sliding windows* por haplotipos de la secuencia aminoacídica de HBc en los alineamientos múltiples pertenecientes a cada subgrupo (CHB (líneas discontinuas) T0 en rojo y T1 en lila, HCC (líneas continuas) T0 en azul y T1 en verde).

Al analizar las diferencias de conservación en la secuencia aminoacídica de HBc en el grupo HCC entre las muestras T0 y T1, se observaron ciertas regiones en las cuales, esta vez, las muestras al tiempo T0 presentaban una mayor conservación que las muestras al tiempo T1 (valores de diferencia T0-T1 positivos), hasta llegar a una diferencia máxima de 0,3 bits. En concreto se identificaron tres regiones de la secuencia de HBc en las que se detectó (con diferencia estadísticamente significativa) una mayor conservación en el subgrupo T0 que en el T1: aa 30-50, 133-142 y 145-165 (p-valores de 0,029, 0,0003 y 1,484E-09 respectivamente para la diferencia entre ambos subgrupos en estas 3 regiones, Figura 40).

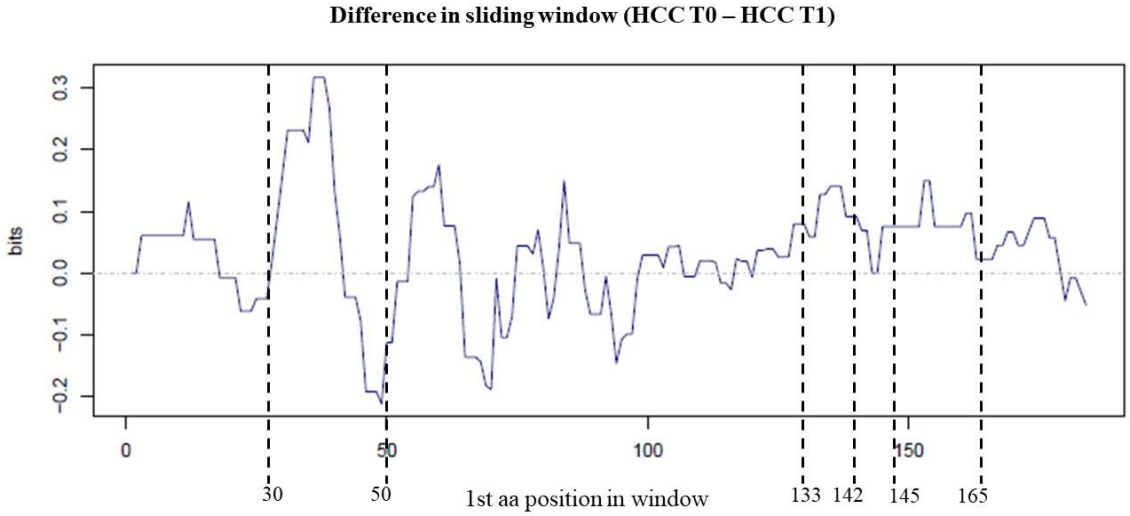


Figura 40. Diferencias en el contenido de información de la secuencia aminoacídica de HBc entre las muestras T0 y T1 de los pacientes del grupo HCC. El gráfico muestra la diferencia de contenido de información en el grupo HCC entre la muestra T0 (antes de desarrollar lesión tumoral) y la muestra T1 (diagnosis de HCC). Los valores negativos son indicativos de una mayor conservación al tiempo T1, mientras que los valores positivos indican un mayor contenido de información en las muestras T0. Las regiones donde se encontró mayor diferencia a nivel de conservación aminoacídica están marcadas y delimitadas por líneas negras discontinuas y sus posiciones están reportadas en la parte inferior de estas (p-valor <0.05 por test de Wilcoxon-Mann-Whitney).

Es importante destacar que estas tres regiones aminoacídicas diferentemente conservadas no presentaban grandes diferencias en el contenido de información entre ambos subgrupos de HCC (Figura 41). De hecho, alrededor del 90% de las posiciones aminoacídicas en estas regiones se encuentran altamente conservadas (en torno a 4 bits) tanto en HCC T0 como en HCC T1. Cabe destacar que, entre los dos subgrupos, en dos de las posiciones aminoacídicas de la región aa 30-50 cambiaban parcialmente aminoácidos polares (tirosina (Y) y arginina (R) en las posiciones 38 y 39, en HCC T0) por no polares (fenilalanina (F) y glicina (G) respectivamente, en HCC T1).

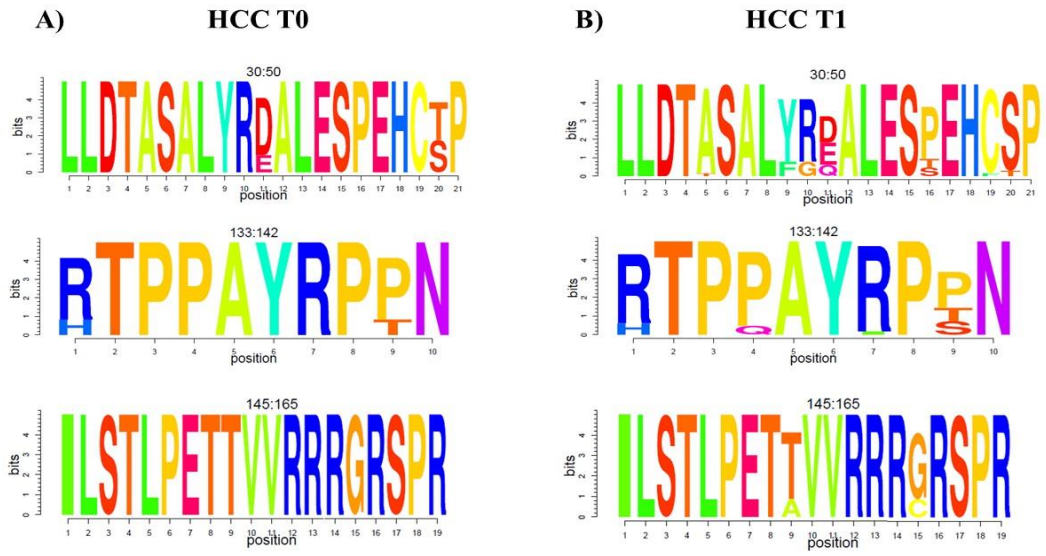


Figura 41. Regiones aminoacídicas diferentemente conservadas en los pacientes HCC entre los dos tiempos (T0 y T1). Representación como logos de secuencia aminoacídica de las tres regiones en las que los pacientes HCC al tiempo T1 estaban menos conservados que al tiempo T0. El panel A corresponde a los logos de estas regiones para el subgrupo HCC T0. El panel B corresponde a los logos de estas regiones para el subgrupo HCC T1. Los extremos de cada región están indicados en la parte superior de cada logo. Los tamaños relativos de las letras en los logos indican sus frecuencias relativas en cada posición dentro del alineamiento múltiple de los haplotipos. La altura de cada letra o pila de letras representa el contenido de información de cada posición medido en bits (eje Y): desde la mínima (0) hasta la máxima conservación (4,32, 100% de conservación en aa).

5.2.5 Estudio de mutaciones en los diferentes subgrupos

El estudio de mutaciones (detallado en el apartado 4.7 de Materiales y métodos: Estudio de mutaciones) se llevó a cabo con el fin de identificar cambios (tanto nucleotídicos como aminoacídicos) que pudieran afectar a la secuencia proteica de HBc y que fueran diferentemente seleccionadas a lo largo de la progresión de la enfermedad hepática.

5.2.5.1 Mutaciones nucleotídicas

Como se ha detallado en el apartado 4.7 de Materiales y métodos: Estudio de mutaciones, en este estudio se analizó la presencia tanto de inserciones y deleciones puntuales de nt (InsDels) como de macrodeleciones de nt con el objetivo de identificar diferencias entre los subgrupos que se pudieran asociar al progreso de la enfermedad hepática. No obstante, en este estudio se detectó un número muy reducido de este tipo de mutaciones y que, además,

presentaban una frecuencia muy baja, por lo que no se determinó ninguna diferencia entre los distintos subgrupos.

5.2.5.2 Sustituciones de aa en la secuencia de HBc.

La búsqueda de sustituciones aminoacídicas se llevó a cabo (como se detalla en el apartado 4.7 de Materiales y métodos: Estudio de mutaciones) también en esta segunda parte del proyecto de tesis doctoral con el fin de identificar cambios en la secuencia aminoacídica que pudieran haber sido seleccionados en la progresión del daño hepático y que pudieran asociarse con el desarrollo de una lesión tumoral.

Como en el caso del primer estudio de este proyecto, en este segundo estudio también se identificaron varias sustituciones aminoacídicas. Concretamente, esta vez se detectaron sustituciones que involucraban a 48 posiciones (reportadas en la Tabla 9) de las 183 posiciones aminoacídicas de la secuencia proteica de HBc.

Región de HBc (extremos aa)	**Sustituciones aminoacídicas detectadas**			
NTD (1-139)	P5T/L/S/H S26N/A/T R39G H51N G63E/V T70S P79Q L84A/Q/R M93V/A I116V/L P135Q/S/T	T12S V27I E40D L55I E64K/D/N G73D A80S/V/Q/G S87G/N/T F97I R127H	S21P/L/G/T S35L/A/K P45T/A/S I59T/F/V/L T67S/A/N V74A/E/G/S S81A T91S/V I105V P130L/T/A/S	F24Y Y38F S49T/A L60V A69V/G E77Q/D D83E N92T/H E113D/Q/K A131P
Región bisagra (140-149)	T147C/A			
CTD (150-183)	R151G/C/Q E180K/D/Q/A	G153C/F/H/Y S181P/F	Q177K	R179P

Tabla 9. Sustituciones aminoacídicas detectadas en la secuencia de HBc en la población total del segundo estudio. La tabla muestra las 48 posiciones en las que se han detectado sustituciones aminoacídicas siendo la primera letra el aa *wild-type* de HBc (es decir, el aa en la secuencia consenso del genotipo), el número la posición aminoacídica involucrada y la segunda letra o grupo de letras el/los nuevos aa codificados. La primera columna indica el dominio de HBc (con las posiciones de sus extremos en la secuencia proteica de HBc reportadas) en el que se han identificado las diferentes sustituciones: NTD (dominio N terminal), Región bisagra y CTD (dominio C terminal).

Cabe mencionar que, de las 48 posiciones en las que se han detectado sustituciones aminoacídicas en este segundo estudio, 39 se habían detectado también en el primer estudio.

Al comparar la proporción de pacientes de cada subgrupo que presentaban estas sustituciones en cada uno de los dominios no se detectó ninguna diferencia estadísticamente significativa.

Por otro lado, al comparar entre los grupos y subgrupos clínicos la frecuencia observada para cada sustitución en cada uno de ellos, se identificó otra vez la sustitución P79Q (cambio de prolina por glutamina en la posición aminoacídica 79). De nuevo, esta sustitución se encontraba más representada en el grupo HCC (T0 + T1) que en el grupo CHB (T0 + T1), mostrando significación estadística en la diferencia, con una mediana (IQR) de 18,19 (0-100) en el grupo HCC *versus* 0 (0-0) en el grupo CHB (p-valor = 0,025, Figura 42)

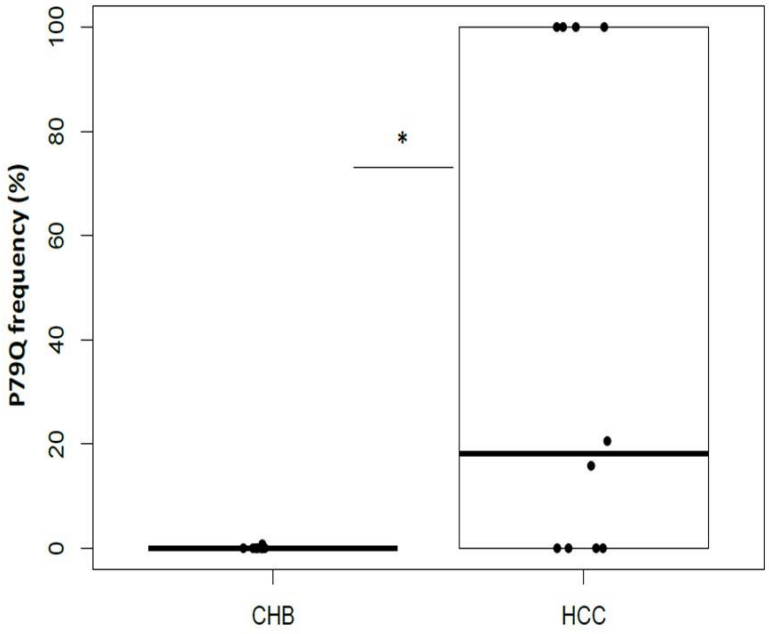


Figura 42. Frecuencia relativa (%) de la sustitución aminoacídica P79Q en la comparativa de los dos grupos clínicos de estudio (CHB y HCC). El diagrama de cajas muestra la comparativa de mediana (IQR) de la frecuencia relativa de la sustitución P79Q entre los grupos CHB y HCC. Cada punto en el gráfico representa una muestra. El p-valor (corregido por Bonferroni) se ha calculado mediante la prueba de Kruskal-Wallis. El p-valor estadísticamente significativo (<0.05) se representa con un asterisco (*).

Con tal de analizar específicamente el grupo HCC, se comparó la frecuencia relativa de las sustituciones detectadas entre las muestras T0 y T1 de este grupo para ver si alguna sustitución se asociaba al progreso de la enfermedad hepática dentro de este mismo grupo de pacientes. Se volvió a detectar la sustitución P79Q y, esta vez, se detectó específicamente en las muestras con lesión tumoral, T1, mostrando significación estadística en la diferencia al compararse con el subgrupo T0 (sin lesión tumoral). Concretamente, se detectó esta sustitución con una frecuencia mediana (IQR) de 100 (20,56-100) en HCC T1 y de 0 (0-0) en HCC T0 (p-valor = 0,045) (Figura 43).

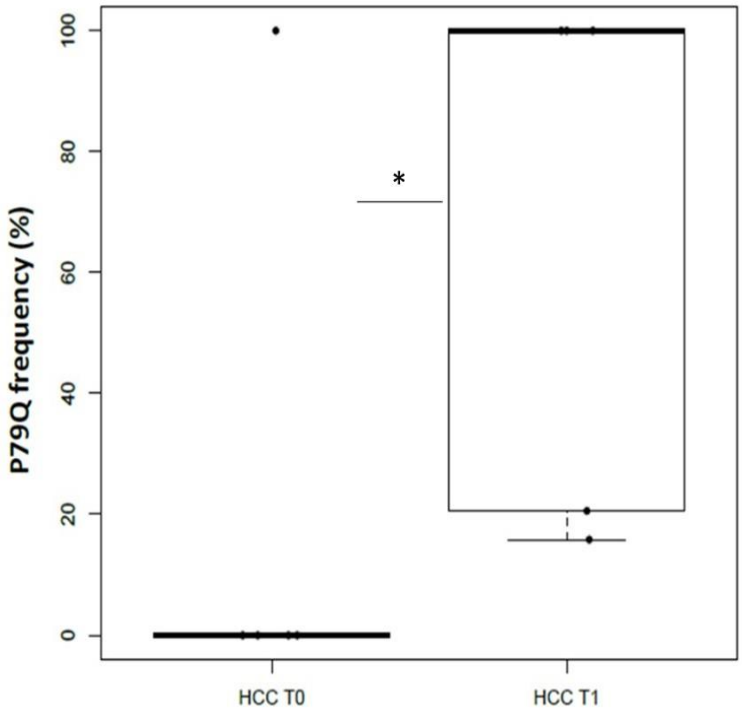


Figura 43. Frecuencia relativa (%) de la sustitución aminoacídica P79Q en la comparativa de los dos subgrupos pertenecientes al grupo clínico HCC (HCC T0 y HCC T1). El diagrama de cajas muestra la comparativa de mediana (IQR) de la frecuencia relativa de la sustitución P79Q entre los dos subgrupos de HCC. Cada punto en el gráfico representa una muestra. El p-valor (corregido por Bonferroni) se ha calculado mediante la prueba de Kruskal-Wallis. El p-valor estadísticamente significativo (<0.05) se representa con un asterisco (*).

5.2.6 Estudio de la complejidad de la quasiespecies

La complejidad de la QS se estudió analizando los índices detallados en el apartado 4.8 de Materiales y métodos: Estudio de la complejidad de la quasiespecies. Este análisis se realizó en cada uno de los 4 subgrupos clínicos de estudio con el fin de detectar cambios en términos de complejidad de la QS entre estos, con especial interés en el subgrupo HCC T1 para poder asociar estos cambios al avance de la enfermedad hepática hacia el desarrollo del tumor. Se analizaron índices de incidencia, de función y de abundancia.

La Tabla 10 resume los valores de mediana (IQR) obtenidos para cada uno de los índices en el análisis de la región del amplicón 1 (nt 1863-2317) para los distintos subgrupos clínicos. Tanto a nivel de índices de incidencia (*number of haplotypes*) como de abundancia (*Shannon entropy* y *Gini-Simpson*) no se detectaron diferencias entre los subgrupos. En el caso de los índices funcionales, no se detectó ninguna diferencia entre los subgrupos al comparar los valores de *Mutation Frequency* y *Nucleotide diversity*, pero sí que se detectó diferencia en terminos de *Average mutation frequency by molecule* al comparar los valores obtenidos para este índice funcional entre los distintos subgrupos (p-valor = 0,008 por test de Kruskal Wallis, Tabla 10).

AMPLICÓN 1					
	CHB		**HCC**		
ÍNDICE	**T0**	**T1**	**T0**	**T1**	**p-valor**
Number of haplotypes	5(3,75-6,75)	10(7,5-11,75)	5(4-6)	6(4-28)	n.s.
Mutation Frequency	4(2-7,75)	12,5(8,75-21,25)	4(3-5)	12(4-59)	n.s.
Shannon entropy	0,48(0,11-0,88)	0,87(0,6-1,02)	0,19(0,17-0,28)	0,63(0,21-2,49)	n.s.
Gini-Simpson	0,23(0,03-0,43)	0,36(0,21-0,45)	0,07(0,06-0,11)	0,33(0,07-0,78)	n.s.
Average mutation frequency by molecule	0,0025 (0,002-0,003)	0,0005(0,0003-0,001)	0,0024(0,0024-0,0024)	0,0043(0,003-0,005)	**0,008**
Nucleotide diversity	0,0006(0,00006-0,00138)	0,0009(0,0006-0,0019)	0,0002(0,0001-0,0003)	0,0008(0,0004-0,005)	n.s.

Tabla 10. Complejidad de la quasiespecies de los distintos subgrupos clínicos a los tiempos T0 y T1 en el amplicón 1. La tabla muestra los valores de mediana (IQR) obtenidos para los índices de complejidad analizados en los distintos tiempos de cada grupo clínico en el amplicón 1. La significación estadística (p-valor) estudiada aplicando el test de Kruskal-Wallis está reportada. El p-valor estadísticamente significativo (<0.05, en el índice *Average mutation frequency by molecule*) está reportado en negrita. Los p-valores >0.05 se consideran no significativos estadísticamente (n.s.).

Al analizar con más detalle la tendencia de este índice (*Average mutation frequency by molecule*) en la región del amplicón 1 y comparar individualmente los valores obtenidos para este entre los cuatros subgrupos, se detectó que el subgrupo HCC T1 (lesión tumoral) presentaba un valor 8,6 veces más alto que el subgrupo CHB T1 (sin daño hepático), mostrando significación estadística en la diferencia: mediana (IQR) del índice *Average mutation frequency by molecule* de 0,0043 (0,003-0,005) en HCC T1 y de 0,0005 (0,0003-0,001) en CHB T1, p-valor = 0,039, ajustado con Bonferroni (Figura 44).

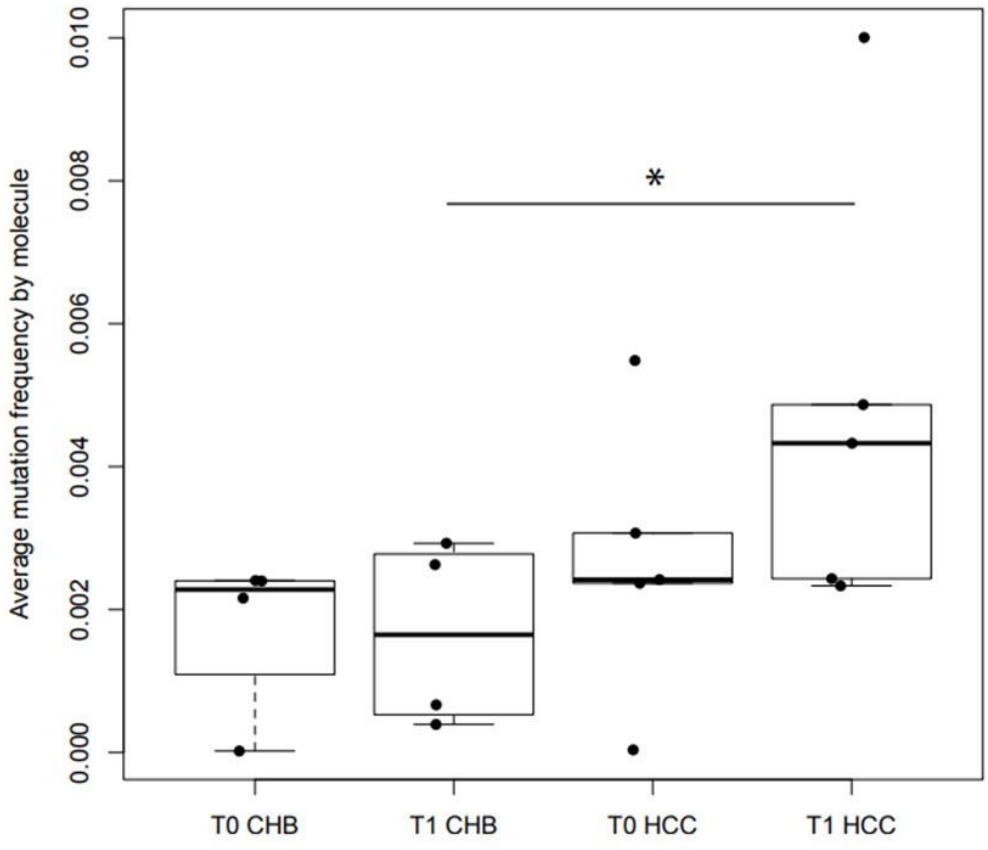


Figura 44. Comparación del índice *Average mutation frequency by molecule* entre los subgrupos clínicos de estudio en la región del amplicón 1. Cada punto en el gráfico representa una muestra. El p-valor (corregido por Bonferroni) se ha calculado mediante la prueba de Kruskal-Wallis con la prueba de comparación múltiple posthoc Dunn. El p-valor estadísticamente significativo (<0.05 entre los subgrupos CHB T1 y HCC T1) se representa con un asterisco (*).

A diferencia de lo detectado en el amplicón 1, en el análisis de la complejidad de las QS en la región del amplicón 2 (nt 2205-2483) no se detectaron diferencias a la hora de comparar los valores obtenidos de los diversos índices analizados entre los distintos subgrupos. La Tabla 11 resume los valores de mediana (IQR) obtenidos de cada uno de los índices en el análisis del amplicón 2 para los distintos subgrupos clínicos.

AMPLICÓN 2					
	CHB		HCC		
ÍNDICE	**T0**	**T1**	**T0**	**T1**	**p-valor**
Number of haplotypes	4(3,75-5)	5,5(4-6,25)	3(2-4)	8(4-8)	n.s.
Mutation Frequency	3(2,75-4)	4,5(3-5,5)	2(1-4)	7(3-44)	n.s.
Shannon entropy	0,2(0,07-0,38)	0,26(0,17-0,44)	0,13(0,04-0,39)	0,44(0,08-0,73)	n.s.
Gini-Simpson	0,09(0,02-0,18)	0,1(0,06-0,2)	0,05(0,01-0,13)	0,16(0,03-0,43)	n.s.
Average mutation frequency by molecule	0,0042(0,0041-0,0043)	0,002(0,001-0,005)	0,0035(0,001-0,0042)	0,0042(0,0012-0,0121)	n.s.
Nucleotide diversity	0,0004(0,0001-0,0008)	0,0004(0,0003-0,0009)	0,0002(0,0001-0,0018)	0,0019(0,0001-0,0042)	n.s.

Tabla 11. Complejidad de la quasiespecie de los distintos grupos clínicos a los tiempos T0 y T1 en el amplicón 2. La tabla muestra los valores de mediana (IQR) obtenidos para los índices de complejidad analizados en los distintos tiempos de cada grupo clínico en el amplicón 2. La significación estadística (p-valor) estudiada aplicando el test de Kruskal-Wallis está reportada. Los p-valores >0.05 se consideran no significativos estadísticamente (n.s.).

DISCUSIÓN

6. DISCUSIÓN

La infección crónica por el VHB afecta a más de 257 millones de personas en todo el mundo con una tasa de muerte del 15-25% debido a las complicaciones hepáticas que derivan de su cronicidad como la cirrosis o el carcinoma hepatocelular (HCC) (135). Se ha estimado que entre 2015 y 2030 ocurrirán más de 60 millones de nuevas infecciones crónicas, con una estimación de 17 millones de muertes debido a la continua transmisión del virus sobre todo en regiones donde la infección es endémica (136).

A pesar de disponer de una vacuna preventiva eficaz y de unas estrategias de tratamiento que permiten controlar la infección, esta se considera un problema de salud global, más teniendo en cuenta que el riesgo de desarrollar HCC en pacientes tratados es de hasta 1,4% en pacientes no cirróticos, llegando hasta más de 5% en pacientes que ya presentan cirrosis (137). Debido a que las estrategias de tratamiento no pueden interferir directamente con el reservorio natural del VHB, el ADN circular covalentemente cerrado (ADNccc), este material genético permanecerá en el núcleo de las células, permitiendo por tanto la continua expresión de los genes virales, aunque esto no se traduzca a la producción de nuevas partículas virales debido a la presencia de inhibidores de la retrotranscripción. Por esta razón, la persistencia del ADNccc y la continua expresión intracelular de antígenos virales hasta en ausencia de carga viral podrían alterar la expresión celular y favorecer la oncogénesis (70). Aun así, se ha calculado que gracias a un adecuado tratamiento preventivo prenatal y perinatal y una diagnosis y tratamiento exhaustivos se podrían evitar más de 7 millones de muertes hasta el 2030, incluyendo hasta 1,5 millones de muertes por HCC (136). Así pues, la identificación de nuevos marcadores de seguimiento y diagnóstico y de nuevas dianas terapéuticas es algo muy apremiante si se quiere llegar a estos resultados.

La proteína HBc (Core o HBcAg, codificada por el gen *HBC*) del VHB es una proteína estructural esencial en la morfología del virus. La cápside icosaédrica que protege al genoma viral y su polimerasa está constituida por 180-240 moléculas de esta proteína (41). Junto a esta actividad estructural, la proteína HBc desempeña todo un conjunto de actividades funcionales que van desde el control de la expresión y formación del ADNccc hasta la interferencia directa con la actividad celular (como se ha detallado en el apartado 1.6.3 de la Introducción: HBc: una proteína funcional en la replicación viral y en la regulación celular) (41). Estos hechos ponen a esta proteína en el punto de mira de cara a una posible diana terapéutica y/o diagnóstica. Dado su rol en la replicación viral, la inhibición de la expresión

intracelular de esta proteína podría ser una valiosa estrategia terapéutica. Asimismo, esta proteína podría tener un papel muy relevante en la progresión de la enfermedad hepática, por lo que su estudio podría contribuir en el conocimiento de los mecanismos de la evolución clínica, pudiendo llegar a la detección de factores pronósticos.

Partiendo de estos hechos, en este proyecto de tesis doctoral se han analizado las quasiespecies (QS) virales del gen *HBC* (y de la secuencia proteica de HBc) en grupos clínicos de pacientes con diferentes estados de la enfermedad hepática causada por la hepatitis crónica B. El análisis ha sido realizado mediante la técnica de Next-Generation Secuencing (NGS), concretamente a través de la plataforma MiSeq Illumina, puesto que esta técnica permite secuenciar completamente las QS que constituyen la población viral (111), obteniendo así unos resultados más fiables y completos.

Concretamente, esta tesis se compone de dos estudios. En el primer estudio se analizaron un total de 38 pacientes con hepatitis crónica B distribuidos en tres grupos en función de la etapa de la enfermedad hepática: sin daño hepático (grupo CHB), con cirrosis hepática (grupo LC) y con carcinoma hepatocelular (grupo HCC). Este primer estudio se focalizó en la identificación de regiones hiperconservadas (independientemente del cuadro clínico o genotipo viral) que pudieran servir como posibles dianas para una terapia basada en el silenciamiento génico. Al mismo tiempo se estudió la presencia de mutaciones o cambios en la conservación (tanto a nivel de nucleótidos (nt) como aminoácidos (aa)) específicos para los distintos grupos incluidos en el estudio y que por lo tanto pudieran estar relacionados con los distintos cuadros clínicos.

En el segundo estudio se incluyeron 9 pacientes con hepatitis crónica B distribuidos en dos grupos clínicos en función de la etapa de la enfermedad hepática (grupos CHB y HCC) y de cada paciente se analizaron dos muestras secuenciales con una diferencia mínima de tiempo de 1 año entre ellas (T0 y T1). En el caso del grupo HCC, las muestras correspondían a una muestra antes (HCC T0) y otra después (HCC T1) de desarrollar la lesión tumoral. Ambas muestras del grupo CHB (CHB T0 y CHB T1) correspondían a muestras sin signos de daño hepático, como el caso del subgrupo HCC T0 (antes de desarrollar el tumor). Este estudio se llevó a cabo con el objetivo de evaluar alteraciones o diferencias tanto en la detección de mutaciones como en la conservación, variabilidad y complejidad de la QS viral durante el

curso de la enfermedad hepática y que por lo tanto se pudieran relacionar con la progresión clínica y ser utilizados como factores pronósticos del avance de la enfermedad.

6.1 Regiones hiperconservadas en pacientes con hepatitis crónica B: búsqueda de nuevas dianas terapéuticas

Considerando la gran variabilidad del VHB, la detección de regiones hiperconservadas independientemente del cuadro clínico o del genotipo viral sería un punto clave en la determinación de nuevos sistemas tanto terapéuticos como de diagnóstico. En el primer estudio se analizó la conservación de la QS viral de pacientes con hepatitis crónica B en diferentes etapas de la enfermedad hepática. Los pacientes estaban infectados por diferentes genotipos virales. Se identificaron 5 genotipos (A, C, D, E, F) y dos mezclas de genotipos (D/E y D/A), probablemente indicativas de eventos de recombinación intergenotípica (99).

El análisis de la conservación de la población total del primer estudio mostró que el gen *HBC* se encuentra bastante conservado en la QS viral a lo largo de toda su secuencia nucleotídica, cosa que pone en evidencia su importancia en la replicación viral. Además, en esta secuencia, se detectaron regiones concretas que mostraban estar altamente conservadas en todos los pacientes del primer estudio, independientemente del genotipo viral por el que estuvieran infectados y de los distintos cuadros clínicos que presentaban. La primera de estas regiones hiperconservadas se detectó entre los nt 1900-1929. Esta región incluye el codón de inicio de traducción del gen (nt 1901-1903). La segunda de estas regiones (nt 2249-2284, aa 116-128 al ser traducida) corresponde a una región en la que al ser traducida se encuentran dos epítopos de linfocitos T CD8+ (138). En la tercera región hiperconservada detectada, entre los nt 2364-2398 (aa 154-166 al ser traducida) encontramos parte de las zonas ricas en arginina del CTD, esenciales para la localización subcelular (96).

Estas regiones altamente conservadas en el gen *HBC* podrían servir como dianas terapéuticas pangenotípicas y panclínicas para una estrategia basada en el silenciamiento de este gen. Los ARN pequeños de interferencia (siRNA) son una de las propuestas más prometedoras con este fin (82), ya que permitirían el silenciamiento de la expresión viral y de la producción de proteínas virales que pudieran interferir con la actividad celular y ser determinantes la progresión de la enfermedad hepática. Por ejemplo, un estudio realizado con ARC-520 (una

combinación de siRNAs dirigidos al gen *HBX* (139)) en combinación con entecavir mostró una disminución de HBsAg en pacientes HBeAg negativos (83–85). En el caso del gen *HBC* su silenciamiento garantizaría, no solo la inhibición de la producción de la proteína HBc (inhibiendo por lo tanto la formación de cápsides virales), sino que también tendría un efecto directo sobre la producción de ARNpg (y consecuentemente en la de ADNrc, pues el ARNpg sirve de molde para su producción). No obstante, teniendo en cuenta estos hechos y a pesar de los avances en los estudios con siRNAs (140) actualmente no hay ninguna molécula efectiva de este tipo dirigida específicamente contra *HBC*.

Al traducir las secuencias nucleotídicas a sus respectivas secuencias aminoacídicas se detectaron 2 regiones de aa hiperconservadas comunes en todos los pacientes del primer estudio. Estas fueron las regiones entre los aa 117-120 y aa 159-167. Estas, al ser traducidas, coincidían parcialmente con dos de las tres regiones nucleotídicas hiperconservadas detectadas, respectivamente con la segunda (nt 2249-2284) y tercera (nt 2364-2398) región nucleotídica citadas anteriormente. Así pues, la región aa 117-120 también corresponde parcialmente a epítopos de linfocitos T CD8+ y la región aa 159-167 engloba parte de las zonas ricas en arginina del CTD. Como se ha detallado en el apartado 1.6.3 de la Introducción: HBc: una proteína funcional en la replicación viral y en la regulación celular, el dominio CTD juega un papel clave en la funcionalidad de HBc, siendo determinantes para ello las diferentes regiones ricas en arginina que encontramos en este dominio. Estas garantizan la adecuada localización subcelular de la proteína actuando como señales de localización nuclear (NLS) o de retención citoplasmática (CRS) (96). La segunda región hiperconservada detectada a nivel de secuencia aminoacídica (aa 159-167), concretamente, englobaba a la tercera región rica en arginina del CTD, que actúa como NLS (situada entre los aa 164-167). Dada la importancia de esta región aminoacídica hiperconservada en la localización nuclear de HBc, su elevado grado de conservación tanto genético como proteico es indicativo de su relevancia a nivel funcional. Por esta razón, podría ser una valiosa diana para sistemas de diagnóstico basados en moléculas altamente específicas como anticuerpos o aptámeros.

Los aptámeros también están emergiendo como una opción muy prometedora por lo que hace al diagnóstico y tratamiento de diferentes enfermedades (141). Estas moléculas, denominadas también anticuerpos químicos, consisten en cadenas simples de ADN o ARN

con una elevada especificidad y afinidad para su diana y que no presentan toxicidad ni inmunogenicidad (142). Se han realizado varios estudios con aptámeros dirigidos al VHB en los que se medía la afinidad de unión de los diversos aptámeros probados a sus respectivos ligandos (142,143) y que arrojan resultados positivos y esperanzadores por lo que hace a esta técnica. Así pues, las regiones hiperconservadas detectadas podrían también servir como dianas para una estrategia basada en aptámeros o incluso para elaborar un nuevo sistema de detección del VHB, como ya se ha hecho con el VHC (144) y con virus sincitiales (145).

6.2 Conservación en los distintos grupos clínicos

El análisis de la conservación se aplicó también exclusivamente sobre cada uno de los distintos grupos clínicos incluidos en el primer estudio con el objetivo de identificar regiones diferentemente conservadas, tanto en la secuencia del gen como en su correspondiente secuencia aminoacídica, y que por lo tanto pudieran relacionarse con un determinado estado clínico.

No se observó mucha diferencia al comparar los niveles de contenido de información de la secuencia nucleotídica entre los grupos clínicos HCC y LC, los cuales presentaban unos elevados niveles de contenido de información. Sin embargo, no se vio lo mismo en el grupo CHB, que era el que presentaba una conservación menor y más fluctuante. En concreto, este último grupo presentaba cinco regiones significativamente menos conservadas respecto a los otros dos grupos. De estas cinco, la región comprendida por los nt 1946-1992, traducida a los aa 15-30, incluye 3 epítopos de linfocitos T CD8+ (138), lo que podría sugerir un mecanismo de evasión inmune en los pacientes del grupo CHB basado en la variabilidad de esta región específica. El hecho de encontrar específicamente en el grupo CHB ciertas regiones menos conservadas a nivel nucleotídico puede relacionarse con la elevada tasa de replicación del VHB durante este estado clínico. Esta variabilidad podría verse incrementada por la presencia, en el tejido, de muchas células hepáticas disponibles que podrían infectarse, en contraste con los grupos LC y HCC, donde la deposición de tejido fibrótico y la replicación descontrolada de las células cancerígenas podrían limitar la replicación viral y en parte, consecuentemente, su variabilidad. Remarcablemente, esta menor conservación detectada en el grupo CHB a nivel nucleotídico no estaba acompañada por una reducida conservación a nivel de secuencia aminoacídica. A este nivel, la secuencia aminoacídica

estaba altamente conservada y de forma muy similar en todos los grupos clínicos, con la excepción de una región entre los aa 140-160, donde los pacientes del grupo LC presentaban niveles inferiores de contenido de información. Esta región engloba a otro epítopo de linfocitos T CD8+ (situado entre los aa 141-151) (138). Estudios futuros permitirían aclarar si la mayor variabilidad en esta región podría estar relacionada con un intento de evasión inmune y con el desarrollo de una lesión cirrótica.

Al mismo tiempo, la comparativa entre grupos ha permitido identificar algunas regiones, tanto a nivel de nucleótido como aminoacídico, donde los niveles de conservación eran grupo-específicos. En el grupo CHB se detectó una región nucleotídica (nt 2306-2334) y una aminoacídica (aa 98-103) conservadas exclusivamente para este grupo. Esta región nucleotídica, que se traduce a los aa 135-144, incluye los 5 primeros aa de la región "bisagra" situada entre los aa 140-149 y que une los dominios NTD y CTD. Esta es una región involucrada en el ensamblaje de la cápside (146,147) y en la síntesis del ADN viral (146), por lo que su elevada conservación en este grupo clínico sería afín a la necesidad de mantener este dominio intacto, sobre todo en esta etapa de la infección donde se da una elevada replicación viral.

En el grupo LC se detectaron 2 regiones nucleotídicas (nt 1935-1976 y 2402-2435) y 2 regiones aminoacídicas (aa 28-30 y 51-54) grupo-específicas. Algunas de estas regiones (concretamente la región nt 1935-1976, que se traduce a los aa 11-25, y la región aa 28-30) corresponden parcialmente a porciones de HBc (aa 14-18 y 23-39) involucradas en el ensamblaje de la cápside y en la envuelta de viriones (134). La segunda región nucleotídica (nt 2402-2435) al ser traducida (aa 167-178) englobaba al cuarto dominio rico en arginina del CTD (RRRR aa 172-175).

Con el objetivo de analizar la evolución del gen *HBC* y su proteína codificante en el curso de la progresión de la enfermedad hepática, en el segundo estudio se incluyeron 5 pacientes con hepatitis crónica por VHB a los cuales se les había diagnosticado HCC. De estos 5 pacientes se analizó una muestra a la diagnosis de tumor (subgrupo HCC T1), y otra anterior, por lo menos de un año atrás (muestras en las que estos 5 pacientes no presentaban aun la lesión tumoral, subgrupo HCC T0). Dos muestras, también con una diferencia mínima de un

año, de 4 pacientes con hepatitis crónica por VHB sin lesión tumoral (en ninguna de las 2 muestras, subgrupos CHB T1 y CHB T0) se usaron como control.

A nivel nucleotídico, ambos subgrupos de HCC mostraron una mayor conservación en líneas generales que los subgrupos de CHB a lo largo de toda la secuencia del gen *HBC*. Como se esperaba, la conservación en los CHB se mantenía constante en ambos tiempos. En contraste, en el grupo HCC el nivel de contenido de información aumentaba en las muestras posteriores a la diagnosis de la lesión tumoral (HCC T1), sobre todo en la región relativa al dominio NTD (nt 1901-2318).

Al analizar la secuencia aminoacídica, los cuatros subgrupos (tanto de CHB como de HCC, a los dos tiempos) mostraban unos patrones de conservación elevados y similares entre sí. No obstante, al analizar específicamente las diferencias entre los dos subgrupos de HCC se identificaron ciertas regiones aminoacídicas en las que esta vez era el subgrupo sin daño tumoral (HCC T0) el que presentaba una mayor conservación, sobre todo en tres regiones: aa 30-50, aa 133-142 y aa 145-165. De estas regiones, la primera engloba a un epítopo menor de células T situado entre los aa 28-47 (87). El hecho de que esta región esté menos conservada en las muestras con HCC ya diagnosticado podría sugerir una selección de variantes capaces de desviar la respuesta T-especifica. De las otras dos regiones más conservadas en las muestras T0 de HCC, la última región (entre los aa 145-165) corresponde a los dos primeros dominios ricos en arginina del CTD. Hay que destacar que, a pesar de la significación estadística hallada en estas tres regiones, ambos subgrupos de HCC no presentaban grandes diferencias, pues en la gran mayoría de las posiciones aminoacídicas de estas regiones se observaba un contenido de información elevado (en torno a 4 bits) en ambos subgrupos. Aun así, el contraste entre la mayor conservación nucleotídica y la menor conservación aminoacídica detectada en las muestras HCC T1 al compararlas con las HCC T0 se explica por el hecho de que los cambios a nivel nucleotídico en HCC T1 afectan al primer o segundo nt del codón en mayor proporción que en el grupo HCC T0 (en el que los cambios a nivel nucleotídico afectan en mayor medida al tercer nt del codón). Así pues, en el subgrupo HCC T1 gran parte de los cambios nucleotídicos se reflejan en cambios en los aa codificados, resultando por lo tanto en una menor conservación aminoacídica. Esto podría ocasionar la presencia y selección de variantes que favorecieran la persistencia de la infección. Es importante tener en cuenta este hecho ya que esta persistencia actúa en pro del

daño hepático y/o la carcinogénesis y en consecuencia, de la morbilidad de la infección (148).

La identificación de regiones conservadas exclusivamente en ciertos grupos clínicos sugiere la existencia de diferentes historias evolutivas que podrían tener efectos sobre la progresión de la enfermedad hepática. Sin embargo, se necesitan más estudios para comprobar esta posible asociación entre las regiones detectadas y las diferentes etapas clínicas, así como para investigar su papel en la progresión de la enfermedad.

6.3 Estudio de las mutaciones: P79Q como posible factor pronóstico de carcinoma hepatocelular

La identificación de sustituciones aminoacídicas como factores pronósticos de la evolución clínica es una valiosa herramienta que ayudaría en el seguimiento de los pacientes con hepatitis crónica por VHB (149). Las mutaciones T1753C y A1762T / G1764A (K130M / V131I en HBx) del BPC (*basal core promoter*), por ejemplo, se identificaron como posibles marcadores pronósticos para el HCC (109,110) y las sustituciones F24Y, E64D, E77Q, A80ITV, L116I y E180A en la secuencia aminoacídica de HBc se asociaron al desarrollo de cirrosis y también de HCC (150).

Con el fin de detectar mutaciones que pudieran relacionarse con un estado clínico específico y servir de factor pronóstico, en ambos estudios se analizó la presencia de mutaciones tanto a nivel nucleotídico como aminoacídico. No obstante, entre las mutaciones nucleotídicas detectadas (inserciones y deleciones puntuales o macrodeleciones) no hubo ninguna, en los dos estudios, que se asociara específicamente a un cuadro clínico. Sin embargo, de las sustituciones aminoacídicas que se detectaron en el primer estudio hubo una, la P79Q (prolina por glutamina en la posición 79), que se vio diferentemente representada entre los tres grupos clínicos del estudio. En concreto, más del 50% de pacientes pertenecientes al grupo HCC (9/17) presentaban esta sustitución y, aunque la frecuencia mediana con la que se detectó en este grupo no fue muy elevada (frecuencia mediana (IQR) de 15,82 (0-78,9)), mostró una diferencia estadísticamente significativa al compararla con la frecuencia mediana con la que se detectó esta sustitución en el grupo control CHB (frecuencia mediana (IQR) de 0 (0-0)).

Al repetir este análisis en el segundo estudio, comparando muestras de los mismos pacientes antes y después del desarrollo de HCC, los resultados obtenidos en el primer estudio se confirmaron. Concretamente, el 100% de las muestras del subgrupo HCC T1 (5/5) presentaban la sustitución P79Q, mientras que sólo se detectó en 1/5 muestras del subgrupo HCC T0 (muestras de los mismos pacientes previas al tumor), mostrando significación estadística en la diferencia entre ambos subgrupos, con una frecuencia mediana (IQR) de 100 (20,56-100) en el tiempo T1 *versus* 0 (0-0) en el tiempo T0 del grupo HCC.

Sustituciones aminoacídicas en la posición aa 79 ya se describieron y asociaron relativamente a la reactivación tumoral después de la resección del tumor hepático (151). Esta posición se encuentra dentro del MIR (*Major Immunodominant Region*, aa 78-82), es decir, en la principal región de reconocimiento de HBc por parte de células B (87). El MIR se localiza en las espículas que forman los dímeros de HBc en la cápside y concretamente, la posición aa 79 se encuentra en el punto más elevado de las espículas en el que la estructura de HBc pasa de una hélice alfa ascendente a otra descendente. Esta posición no solamente es de gran importancia inmunológica, sino que además estudios de microscopía crioelectrónica han mostrado que la posición aa 79, al estar en el extremo de las espículas, forma parte del área de contacto que se da entre HBc y las proteínas de la envuelta viral en los viriones (88).

En el caso de esta mutación aminoacídica identificada, cabe destacar las características de los aa involucrados en la sustitución. La prolina es un aa apolar cuya naturaleza (único aa proteinogénico con un ángulo restringido) le permite influir en la estructura proteica, aportando un punto de flexión a esta estructura en el punto donde se encuentra (152). El hecho de que en esta posición se encuentre una prolina es, por lo tanto, importante para garantizar la arquitectura necesaria para unir a dos hélices alfa paralelas en su parte superior (punta de las espículas). La glutamina, en cambio, es un aa polar sin carga y su estructura difiere mucho de la de la prolina, por lo que al encontrarse en una posición tan señalada como la aa 79 podría estar afectando a la estructura terciaria de HBc. Si sumamos este hecho a la importancia inmunológica de la posición se podría especular que la sustitución P79Q haya sido seleccionada probablemente debido a la presión inmunológica como mecanismo de evasión inmune, aunque podría afectar al contacto entre la cápside y la envuelta viral. El hecho que la sustitución esté más representada en los pacientes con lesión tumoral podría

sugerir su relación con el desarrollo de HCC. De todas formas, se requieren posteriores estudios *in vitro* para investigar y definir correctamente el papel de la mutación P79Q en la progresión de la enfermedad hepática.

6.4 Estudio de la complejidad de la quasiespecie

La complejidad de una QS informa de factores como el potencial patogénico, la respuesta al tratamiento antiviral, la evolución clínica y la seroconversión (112–117). Por esta razón, en el segundo estudio de la tesis doctoral se analizó la complejidad de la QS en muestras secuenciales de los mismos pacientes con el fin de observar cómo varía entre dos fases distintas de la enfermedad hepática (sin daño hepático y con lesión tumoral). Con este fin se calcularon diferentes índices de complejidad (de incidencia, de abundancia y de función, como se ha detallado en el apartado 4.8 de Materiales y métodos: Estudio de la complejidad de la quasiespecies).

En ninguno de los dos amplicones se detectaron diferencias significativas entre los 3 subgrupos que correspondían a muestras sin daño hepático (CHB T0, CHB T1 y HCC T0). Sin embargo, al comparar los datos de estos tres subgrupos con los datos obtenidos para las muestras al tiempo de la diagnosis de lesión tumoral de los pacientes HCC (subgrupo HCC T1), se detectó una complejidad 8,6 veces mayor en este último subgrupo en términos de *Average mutation frequency by molecule (*Mfm*)* en la región del amplicón 1 (nt 1863-2317) respecto al subgrupo CHB T1 (mediana de 0,0043 vs 0,0005 respectivamente en HCC T1 y CHB T1). Éste es un índice funcional de abundancia que mide la proporción de nts diferentes a nivel molecular (teniendo en cuenta el número de nts secuenciados), es decir, es una medida de la fracción de nts de la población viral que difieren del haplotipo dominante de la QS (153). Por lo tanto, cuanto más alto sea el valor obtenido para este índice mayor será la tasa de mutación a nivel molecular para la QS (111).

Cabe destacar que, como se ha discutido anteriormente en el apartado 6.2 de esta Discusión: Conservación en los distintos grupos clínicos, la QS del gen *HBC* del VHB en el subgrupo HCC T1 (con tumor hepático) se presentaba muy conservada a lo largo de toda su secuencia nucleotídica. Al mismo tiempo, como se ha comentado en el párrafo anterior, en la región del amplicón 1 de este subgrupo HCC T1 se observó una elevada complejidad en términos de frecuencia de mutaciones por molécula. Estos resultados, aparentemente contradictorios,

podrían explicarse por la forma en que se calcula cada uno de los dos factores (Mfm y conservación). En el cálculo del primero (Mfm), se evalúan los cambios globales de cada haplotipo respecto al haplotipo dominante, sin tener en cuenta si las posiciones que presentan mutaciones son las mismas entre un haplotipo y otro. En el caso de la conservación, se considera el contenido de información de ventanas de 25 nts que avanzan en pasos secuenciales de 1 nt entre una ventana y la siguiente. En este caso pues, se considera el contenido de información obtenido en general para cada posición nucleotídica. Por lo tanto, estos dos resultados podrían sugerir que este subgrupo (HCC T1) se caracteriza por haplotipos (sobre todo en la región del primer amplicón) que presentan muchas mutaciones nucleotídicas, pero en distintas posiciones entre ellos, por lo que la QS en cada posición en líneas generales se conserva, aunque la frecuencia de estas mutaciones sea alta, ya que las posiciones involucradas son muchas y distintas. Además, estos cambios estarían afectando en gran medida a los dos primeros nts de los codones, ocasionando sustituciones aminoacídicas al ser traducidos, lo que se refleja en una reducción de la conservación a nivel aminoacídico de HBc.

En el análisis de la complejidad del amplicón 2 (nt 2205-2483) no se detectó ninguna diferencia significativa entre los subgrupos al comparar los valores obtenidos para los diferentes índices de complejidad calculados. Es más, en este amplicón 2 la mediana obtenida para el *Average mutation frequency by molecule* para el grupo HCC T1 fue idéntica a la obtenida para el subgrupo sin daño hepático CHB T0 (0,0042).

La situación histológica asociada a una lesión tumoral es un hecho a tener en cuenta para esta argumentación. En esta fase la histología es heterogénea, con un patrón nodular en el que el nódulo puede actuar como una "isla" independiente del resto del hígado, quedando cada una de estas dos zonas sujetas a diferentes presiones de selección, a diferencia de lo que ocurre en la histología de una hepatitis crónica sin lesión donde todo el tejido infectado es homogéneo e igualmente accesible por la respuesta inmune. Esto podría favorecer una presión de selección que afectara a regiones específicas de las proteínas virales involucradas en el reconocimiento inmunológico, como podría haber sucedido con la mutación P79Q. Además, en un nódulo tumoral el metabolismo de los hepatocitos está alterado por la restricción del riego sanguíneo y su energía metabólica se estaría destinando sobre todo a procesos de división celular (efecto Warburg: las células tumorales, a pesar de consumir

menos oxígeno, metabolizan más glucosa que las células sanas, lo que actúa en pro del crecimiento y proliferación descontrolados) (154), minimizando el consumo de energía destinado a los sistemas de edición antiviral como ADAR o APOBEC (155,156), que podrían tener una actividad muy inferior a la basal. De hecho en el segundo estudio las mutaciones G-A, que pueden estar causadas por el sistema APOBEC (155), se encontraron en menor proporción en el subgrupo HCC T1 que en el HCC T0, informando de una posible edición antiviral reducida en los pacientes con lesión tumoral. De esta manera habría menos edición antiviral y por tanto menos cambios posición-específicos debidos a estos sistemas, por lo que la QS se vería más conservada y los cambios se estarían ocasionando mayoritariamente al azar (probablemente debido a errores de la retrotranscripción). Esta situación también ayudaría a explicar la elevada conservación observada a nivel nucleotídico en los pacientes con HCC, ya que en ambos estudios estos grupos (grupo HCC en el primer estudio, subgrupo HCC T1 en el segundo) se encontraban más conservados a nivel de nt que los grupos sin daño hepático (CHB en el primer estudio, CHB T0, CHB T1 y HCC T0 en el segundo).

No obstante, serán necesarios estudios con una población mayor y estudios *in vitro* para corroborar este hecho y para definir cómo la complejidad de la QS puede influir en el avance de la enfermedad hepática.

CONCLUSIONES

7. CONCLUSIONES

1. Las regiones hiperconservadas detectadas a nivel nucleotídico y aminoacídico evidencian su importancia funcional para el VHB y podrían ser dianas hacia las que dirigir nuevas estrategias panclínicas y pangenotípicas de terapia y diagnóstico.

2. La identificación de regiones conservadas exclusivas de un cuadro clínico concreto sugiere una relación entre estas regiones y la evolución del daño hepático. La relevancia de estas regiones en la capacidad replicativa del virus y el papel que cumplen en el progreso de la enfermedad los patrones de conservación diferenciales observados entre los distintos cuadros clínicos deberán analizarse en profundidad en futuros estudios.

3. La sustitución aminoacídica P79Q se ha identificado en ambos estudios en los pacientes con lesión hepática por HCC. Estos resultados son muy prometedores a la hora de postular a esta mutación como un posible factor pronóstico del desarrollo de HCC. No obstante, es necesario un estudio en el que se analice una población más numerosa, así como estudios *in vitro* para evaluar su función en la transformación celular.

4. Los pacientes del segundo estudio con lesión tumoral (HCC T1) han mostrado una "complejidad" de la QS mayor que los pacientes sin daño hepático, caracterizada por una elevada frecuencia de mutaciones por molécula, por una elevada conservación a nivel nucleotídico y una menor conservación a nivel de secuencia aminoacídica. Estos resultados podrían explicarse por la presencia de muchas mutaciones pero que afectan a diferentes posiciones nucleotídicas y que por lo tanto no afectan a la conservación especifica de cada posición, pero sí que podrían generar una ligera alteración de la conservación de la secuencia aminoacídica.

5. La presencia en los pacientes con HCC de una QS compleja y conservada al mismo tiempo podría estar relacionada con la propia arquitectura del tejido hepático en esta etapa de la enfermedad y, por consiguiente, con las diferencias respecto a un hígado sano en términos de respuesta inmune, tanto adaptativa como innata, intracelular.

6. Posteriores estudios con un grupo de pacientes más amplio serán necesarios para confirmar los resultados obtenidos. También serán necesarios experimentos *in vitro* para evaluar las posibilidades terapéuticas y diagnosticas de las regiones hiperconservadas identificadas, así como para examinar la validez de la conservación grupo-específica y la sustitución aminoacídica P79Q como posibles factores pronósticos de la evolución clínica de la enfermedad hepática.

LIMITACIONES DEL PROYECTO

8. LIMITACIONES DEL PROYECTO

La principal limitación de esta tesis doctoral ha sido el número de pacientes incluidos en los estudios debido a los límites de detección de los protocolos de PCR. Esto ha sido particularmente evidente en el grupo LC del primer estudio, en el que solo se pudieron incluir a 5 pacientes.

A esto hay que sumarle el hecho de que, aunque la plataforma MiSeq Illumina ofrezca una longitud de secuenciación relativamente larga, no ha sido suficiente como para cubrir el gen *HBC* completo con solo un amplicón. Esto obligó a estudiar el gen dividido en dos amplicones parcialmente solapantes, por lo que el rendimiento a la hora de secuenciar cada uno de los dos amplicones de cada muestra podría haber influido en los resultados obtenidos. Este hecho también ha influido en gran medida a la hora de incluir pacientes en los estudios, por lo que se incluyeron sólo aquellos pacientes en que ambos amplicones se amplificaron correctamente y que presentaban un solapamiento correcto de la región común. Este hecho fue todavía más importante en el segundo estudio, donde la principal premisa era la correcta secuenciación de ambos amplicones en ambos tiempos (T0 y T1) de cada paciente. Por esta razón, de los 12 pacientes inicialmente incluidos en este segundo estudio, solo se pudieron analizar cuatro del grupo CHB y cinco del grupo HCC.

LÍNEAS DE FUTURO

9. LÍNEAS DE FUTURO

Los resultados obtenidos en esta tesis doctoral podrían guiar la creación de una nueva estrategia de terapia génica basada en el silenciamiento genético a través de siRNAs. Para este fin, será necesario investigar más en profundidad las regiones hiperconservadas identificadas en el proyecto, analizando la conservación de estas regiones en grupos de pacientes más grandes y que incluyan también otros genotipos virales.

Asimismo, será necesario seguir estudiando las regiones conservadas detectadas exclusivamente en algún grupo clínico concreto para poder determinar su papel en el progreso de la enfermedad hepática. Para ello, se podrían introducir sustituciones aminoacídicas en estas regiones con conservación o variabilidad grupo-específica con el fin de evaluar si estos cambios afectan a la expresión viral *in vitro* y si inducen un mayor o menor daño celular. Este sistema de expresión *in vitro* de mutantes del VHB ya está puesto a punto por nuestro grupo.

Disponer de una población de estudio de mayor tamaño sería muy ventajoso dado que permitiría corroborar con más contundencia los resultados aquí expuestos.

La sustitución aminoacídica P79Q, que se detectó en presencia de HCC, podría ser un valioso factor pronóstico de tumor hepático que ayudaría en el seguimiento de los pacientes con hepatitis crónica debida a la infección por VHB. Para ese fin se llevarán a cabo estudios *in vitro* en los que se investigará cómo esta mutación puede afectar a la replicación viral y a la actividad y proliferación celular. Nuestro grupo ya ha producido el mutante correspondiente, por lo que en un futuro muy próximo se analizará a través de la transfección de este mutante en células susceptibles a infección para evaluar si la sustitución aminoacídica introducida en el mutante puede afectar a la expresión viral, así como para evaluar si se puede asociar a un mayor daño celular o a una alteración de la proliferación celular, lo que podría relacionar a esta sustitución P79Q con el desarrollo del tumor hepático.

Los datos obtenidos de complejidad y conservación en los pacientes HCC T1 (tumor hepático) del segundo estudio y su relación con la heterogeneidad histológica que muestra el tejido hepático al desarrollar un tumor podrán ser analizados más en profundidad gracias a biopsias hepáticas en las que se pueda estudiar exclusivamente la complejidad y

conservación del nódulo tumoral por una parte y por otra parte estudiarlas en el tejido hepático no tumoral con tal de comparar los datos obtenidos y corroborar la situación expuesta en la discusión. Asimismo, estos resultados se podrían corroborar con el análisis de la QS del ARN circulante del VHB, con el fin de compararlo con lo que se observa a nivel de ADN viral circulante e intrahepático, y así poder hipotetizar el origen de esta complejidad. Nuestro grupo ya está desarrollando un estudio en el que se comparará la complejidad de la QS del ARN y del ADN del VHB.

Así pues, los resultados que se podrán obtener con los posteriores estudios nos permitirán evaluar la utilidad tanto de la complejidad y la conservación grupo-específicas como de la sustitución aminoacídica detectadas como posibles factores pronósticos de la evolución de la enfermedad hepática, con el fin de limitar la necesidad de tener que estudiar la progresión de la enfermedad mediante la aplicación de una biopsia hepática.

BIBLIOGRAFÍA

10. BIBLIOGRAFÍA

1. Pourkarim MR, Amini-Bavil-Olyaee S, Kurbanov F, Van Ranst M, Tacke F. Molecular identification of hepatitis B virus genotypes/ subgenotypes: Revised classification hurdles and updated resolutions. Vol. 20, World Journal of Gastroenterology. WJG Press; 2014. p. 7152–68.

2. Ott JJ, Stevens GA, Groeger J, Wiersma ST. Global epidemiology of hepatitis B virus infection: New estimates of age-specific HBsAg seroprevalence and endemicity. Vaccine. 2012 Mar 9;30(12):2212–9.

3. WHO Hepatitis B. Hepatitis B [Internet]. Available from: https://www.who.int/es/news-room/fact-sheets/detail/hepatitis-b

4. Papastergiou V, Lombardi R, MacDonald D, Tsochatzis EA. Global epidemiology of hepatitis B virus (HBV) infection. Curr Hepat Rep. 2015 Jul 17;14(3):171–8.

5. Blumberg BS, Alter HJ, Visnich S. A "New" Antigen in Leukemia Sera. JAMA J Am Med Assoc. 1965 Feb 15;191(7):541–6.

6. Giles JP, McCollum RW, Berndtson LW, Krugman S. Relation of Australia-SH antigen to the willowbrook MS-2 strain. N Engl J Med. 1969 Jul 17;281(3):119–22.

7. Almeida JD, Rubenstein D, Stott EJ. NEW ANTIGEN-ANTIBODY SYSTEM IN AUSTRALIA-ANTIGEN-POSITIVE HEPATITIS. Lancet. 1971 Dec 4;298(7736):1225–7.

8. Dane DS, Cameron CH, Briggs M. VIRUS-LIKE PARTICLES IN SERUM OF PATIENTS WITH AUSTRALIA-ANTIGEN-ASSOCIATED HEPATITIS. Lancet. 1970 Apr 4;295(7649):695–8.

9. Galibert F, Mandart E, Fitoussi F, Tiollais P, Charnay P. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. Nature. 1979;281(5733):646–50.

10. Gerlich WH. Medical Virology of Hepatitis B: How it began and where we are now. Vol. 10, Virology Journal. 2013. p. 239.

11. Gust ID, Burrell CJ, Coulepis AG, Robinson WS, Zuckerman AJ. Taxonomic classification of human hepatitis B virus. Intervirology. 1986;25(1):14–29.

12. Alter MJ. Epidemiology and prevention of hepatitis B. Vol. 23, Seminars in Liver Disease. 2003. p. 39–46.

13. ASSCAT. Transmisión del virus de la hepatitis B (VHB) [Internet]. Available from: https://asscat-hepatitis.org/hepatitis-viricas/hepatitis-b/informacion-basica-sobre-la-hepatitis-b/transmision-del-virus-de-la-hepatitis-b-vhb/

14. Tatematsu K, Tanaka Y, Kurbanov F, Sugauchi F, Mano S, Maeshiro T, et al. A Genetic Variant of Hepatitis B Virus Divergent from Known Human and Ape Genotypes Isolated from a Japanese Patient and Provisionally Assigned to New Genotype J. J Virol. 2009 Oct 15;83(20):10538–47.

15. Kramvis A. Genotypes and genetic variability of hepatitis B virus. Intervirology. 2014;57(3–4):141–50.

16. Sunbul M. Hepatitis B virus genotypes: Global distribution and clinical importance. World J Gastroenterol. 2014 May 14;20(18):5427–34.

17. Kao JH. Molecular epidemiology of hepatitis B virus. Korean J Intern Med. 2011 Sep;26(3):255–61.

18. Luca AS, Ursu RG, Teu T, Luca CM PC. THE NEED FOR HBV GENOTYPING: A COST-EFFICIENT APPROACH. Rev Med Chir Soc Med Nat Iasi. 2015;119:982–7.

19. Kramvis A, Arakawa K, Yu MC, Nogueira R, Stram DO, Kew MC. Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. J Med Virol. 2008 Jan;80(1):27–46.

20. Huang H, Wang J, Li W, Chen R, Chen X, Zhang F, et al. Serum HBV DNA plus RNA shows superiority in reflecting the activity of intrahepatic cccDNA in treatment-naïve HBV-infected individuals. J Clin Virol. 2018 Feb 1;99–100:71–8.

21. Wang J, Yu Y, Li G, Shen C, Meng Z, Zheng J, et al. Relationship between serum HBV-RNA levels and intrahepatic viral as well as histologic activity markers in entecavir-treated patients. J Hepatol. 2018 Jan 1;68(1):16–24.

22. Liu Y, Jiang M, Xue J, Yan H, Liang X. Serum HBV RNA quantification: Useful for monitoring natural history of chronic hepatitis B infection. BMC Gastroenterol. 2019 Apr 16;19(1):53.

23. Hosaka T, Suzuki F, Kobayashi M, Fujiyama S, Kawamura Y, Sezaki H, et al. Impact of hepatitis B core-related antigen on the incidence of hepatocellular carcinoma in patients treated with nucleos(t)ide analogues. Aliment Pharmacol Ther. 2019 Feb 1;49(4):457–71.

24. Riveiro-Barciela M, Bes M, Rodríguez-Frías F, Tabernero D, Ruiz A, Casillas R, et al. Serum hepatitis B core-related antigen is more accurate than hepatitis B surface antigen to identify inactive carriers, regardless of hepatitis B virus genotype. Clin Microbiol Infect. 2017 Nov 1;23(11):860–7.

25. Elsevier. Hepatitis aguda | Medicina Integral [Internet]. Available from: https://www.elsevier.es/es-revista-medicina-integral-63-articulo-hepatitis-aguda-11321

26. Trépo C, Chan HLY, Lok A. Hepatitis B virus infection. Lancet. 2014 Dec 6;384(9959):2053–63.

27. Ganem D, Prince AM. Hepatitis B Virus Infection - Natural History and Clinical Consequences. N Engl J Med. 2004 Mar 11;350(11):1118–29.

28. Lee WM, Squires RH, Nyberg SL, Doo E, Hoofnagle JH. Acute liver failure: Summary of a workshop. Hepatology. 2008 Apr 19;47(4):1401–15.

29. Buti M, García-Samaniego J, Prieto M, Rodríguez M, Sánchez-Tapias JM, Suárez E, et al. Documento de consenso de la AEEH sobre el tratamiento de la infección por el virus de la hepatitis B (2012). Gastroenterol Hepatol. 2012 Aug;35(7):512–28.

30. Levrero M, Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma. J Hepatol. 2016 Apr;64(1):S84–101.

31. Lampertico P, Agarwal K, Berg T, Buti M, Janssen HLA, Papatheodoridis G, et al. EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. J Hepatol. 2017 Aug 1;67(2):370–98.

32. Hadziyannis SJ. Natural history of chronic hepatitis B in Euro-Mediterranean and African Countries. J Hepatol. 2011 Jul;55(1):183–91.

33. Hadziyannis SJ, Papatheodoridis G V. Hepatitis B e antigen-negative chronic hepatitis B: Natural history and treatment. Semin Liver Dis. 2006 May;26(2):130–41.

34. Venook AP, Papandreou C, Furuse J, Ladrón de Guevara L. The Incidence and Epidemiology of Hepatocellular Carcinoma: A Global and Regional Perspective. Oncologist. 2010 Nov;15(S4):5–13.

35. Park YN, Chae KJ, Kim YB, Park C TN. Apoptosis and proliferation in hepatocarcinogenesis related to cirrhosis. Cancer. 2001;92(11):2733–2738.

36. Rapti I, Hadziyannis S. Risk for hepatocellular carcinoma in the course of chronic hepatitis B virus infection and the protective effect of therapy with nucleos(t)ide analogues. World J Hepatol. 2015;7(8):1064–73.

37. Perrillo RP. Acute flares in chronic hepatitis B: The natural and unnatural history of an immunologically mediated liver disease. Gastroenterology. 2001 Mar;120(4):1009–22.

38. Neuveut C, Wei Y, Buendia MA. Mechanisms of HBV-related hepatocarcinogenesis. J Hepatol. 2010 Apr;52(4):594–604.

39. Ning-Fang M, Lau SH, Hu L, Xie D, Wu J, Yang J, et al. COOH-terminal truncated HBV X protein plays key role in hepatocarcinogenesis. Clin Cancer Res. 2008 Aug 15;14(16):5061–8.

40. Urban S, Bartenschlager R, Kubitz R, Zoulim F. Strategies to inhibit entry of HBV and HDV into hepatocytes. Gastroenterology. 2014 Jul;147(1):48–64.

41. Diab A, Foca A, Zoulim F, Durantel D, Andrisani O. The diverse functions of the hepatitis B core/capsid protein (HBc) in the viral life cycle: Implications for the development of HBc-targeting antivirals. Antiviral Res. 2018 Jan 1;149:211–20.

42. Araujo NM, Vianna COA, Moraes MTB, Gomes SA. Expression of hepatitis B virus surface antigen (HBsAg) from genotypes A, D and F and influence of amino acid variations related or not to genotypes on HBsAg detection. Brazilian J Infect Dis. 2009 Aug;13(4):266–71.

43. Schädler S, Hildt E. HBV life cycle: Entry and morphogenesis. Viruses. 2009 Sep 1;1(2):185–209.

44. Datta S, Chatterjee S, Veer V, Chakravarty R. Molecular Biology of the Hepatitis B Virus for Clinicians. J Clin Exp Hepatol. 2012 Dec;2(4):353–65.

45. Rodríguez-Frias F, Jardi R. Molecular virology of the hepatitis B virus. Enferm Infecc Microbiol Clin. 2008 May;26(SUPPL. 7):2–10.

46. Tang H, Oishi N, Kaneko S, Murakami S. Molecular functions and biological roles of hepatitis B virus x protein. Cancer Sci. 2006 Oct;97(10):977–83.

47. Minor MM, Slagle BL. Hepatitis B virus HBx protein interactions with the ubiquitin proteasome system. Viruses. 2014 Nov 24;6(11):4683–702.

48. Glebe D, König A. Molecular virology of hepatitis B virus and targets for antiviral intervention. Intervirology. 2014;57(3–4):134–40.

49. Caligiuri P, Cerruti R, Icardi G, Bruzzone B. Overview of hepatitis B virus mutations and their implications in the management of infection. World J Gastroenterol. 2016 Jan 7;22(1):145–54.

50. Torresi J. The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. J Clin Virol. 2002 Aug;25(2):97–106.

51. Park NH, Song IH, Chung YH. Chronic hepatitis B in hepatocarcinogenesis. Postgrad Med J. 2006 Aug;82(970):507–15.

52. Visvanathan K, Skinner NA, Thompson AJV, Riordan SM, Sozzi V, Edwards R, et al. Regulation of Toll-like receptor-2 expression in chronic hepatitis B by the precore protein. Hepatology. 2007 Jan;45(1):102–10.

53. Milich DR, Scho¨del F, Scho¨del S, Hughes JL, Jones JE, Peterson DL. The Hepatitis B Virus Core and e Antigens Elicit Different Th Cell Subsets: Antigen Structure Can Affect Th Cell Phenotype †. J Virol. 1997;71(3):2192–201.

54. Lang T, Lo C, Skinner N, Locarnini S, Visvanathan K, Mansell A. The Hepatitis B e antigen (HBeAg) targets and suppresses activation of the Toll-like receptor signaling pathway. J Hepatol. 2011 Oct;55(4):762–9.

55. Pumpens P, Grens E. HBV core particles as a carrier for B cell/T cell epitopes. Intervirology. 2001;44(2–3):98–114.

56. Caballero A. Complejidad de la cuasiespecie del virus de la hepatitis B en la región X/preCore: asociación con la evolución de la infección con y sin tratamiento antiviral. 2016;Tesis doct:UAB.

57. Sureau C, Salisse J. A conformational heparan sulfate binding site essential to infectivity overlaps with the conserved hepatitis B virus a-determinant. Hepatology. 2013 Mar;57(3):985–94.

58. Stoeckl L, Funk A, Kopitzki A, Brandenburg B, Oess S, Will H, et al. Identification of a structural motif crucial for infectivity of hepatitis B viruses. Proc Natl Acad Sci U S A. 2006 Apr 25;103(17):6730–4.

59. Rabe B, Glebe D, Kann M. Lipid-Mediated Introduction of Hepatitis B Virus Capsids into Nonsusceptible Cells Allows Highly Efficient Replication and Facilitates the Study of Early Infection Events. J Virol. 2006 Jun 1;80(11):5465–73.

60. Bock CT, Schwinn S, Locarnini S, Fyfe J, Manns MP, Trautwein C, et al. Structural organization of the hepatitis B virus minichromosome. J Mol Biol. 2001 Mar 16;307(1):183–96.

61. Levrero M, Pollicino T, Petersen J, Belloni L, Raimondo G, Dandri M. Control of cccDNA function in hepatitis B virus infection. J Hepatol. 2009 Sep;51(3):581–92.

62. Moraleda G, Saputelli J, Aldrich CE, Averett D, Condreay L, Mason WS. Lack of Effect of Antiviral Therapy in Nondividing Hepatocyte Cultures on the Closed Circular DNA of Woodchuck Hepatitis Virus. J Virol. 1997;71(12):9392–9.

63. Chan HLY, Thompson A, Martinot-Peignoux M, Piratvisuth T, Cornberg M, Brunetto MR, et al. Hepatitis B surface antigen quantification: Why and how to use it in 2011 - A core group report. J Hepatol. 2011 Nov;55(5):1121–31.

64. Bartenschlager R, Schaller H. Hepadnaviral assembly is initiated by polymerase binding to the encapsidation signal in the viral RNA genome. EMBO J. 1992 Sep;11(9):3413–20.

65. Le Pogam S, Chua PK, Newman M, Shih C. Exposure of RNA Templates and Encapsidation of Spliced Viral RNA Are Influenced by the Arginine-Rich Domain of Human Hepatitis B Virus Core Antigen (HBcAg 165-173). J Virol. 2005 Feb 1;79(3):1871–87.

66. Feng H, Hu KH. Structural characteristics and molecular mechanism of hepatitis B virus reverse transcriptase. Virol Sin. 2009 Dec 27;24(6):509–17.

67. Lambert C, Döring T, Prange R. Hepatitis B Virus Maturation Is Sensitive to Functional Inhibition of ESCRT-III, Vps4, and γ2-Adaptin. J Virol. 2007 Sep 1;81(17):9050–60.

68. Tuttleman JS, Pourcel C, Summers J. Formation of the pool of covalently closed circular viral DNA in hepadnavirus-infected cells. Cell. 1986 Nov 7;47(3):451–60.

69. Hu J, Liu K. Complete and incomplete hepatitis B virus particles: Formation, function, and application. Viruses. 2017 Mar 21;9(3):56.

70. Papatheodoridis G, Buti M, Cornberg M, Janssen H, Mutimer D, Pol S, et al. EASL clinical practice guidelines: Management of chronic hepatitis B virus infection. J Hepatol. 2012 Jul 1;57(1):167–85.

71. Tanaka E, Matsumoto A. Guidelines for avoiding risks resulting from discontinuation of nucleoside/nucleotide analogs in patients with chronic hepatitis B. Hepatol Res. 2014 Jan;44(1):1–8.

72. Santantonio TA. Chronic hepatitis B: Advances in treatment. World J Hepatol. 2014;6(5):284.

73. Zeisel MB, Lucifora J, Mason WS, Sureau C, Beck J, Levrero M, et al. Towards an HBV cure: State-of-the-art and unresolved questions-report of the ANRS workshop on HBV cure. Gut. 2015 Aug 1;64(8):1314–26.

74. Lok AS, Zoulim F, Dusheiko G, Ghany MG. Hepatitis B cure: From discovery to regulatory approval. Hepatology. 2017 Oct 1;66(4):1296–313.

75. Lin CL, Kao JH. Review article: novel therapies for hepatitis B virus cure – advances and perspectives. Aliment Pharmacol Ther. 2016 Aug 1;44(3):213–22.

76. Gane E, Verdon DJ, Brooks AE, Gaggar A, Nguyen AH, Subramanian GM, et al. Anti-PD-1 blockade with nivolumab with and without therapeutic vaccination for virally suppressed chronic hepatitis B: A pilot study. J Hepatol. 2019 Nov 1;71(5):900–7.

77. Testoni B, Durantel D, Zoulim F. Novel targets for hepatitis B virus therapy. Liver Int. 2017 Jan 1;37:33–9.

78. Lucifora J, Xia Y, Reisinger F, Zhang K, Stadler D, Cheng X, et al. Specific and nonhepatotoxic degradation of nuclear hepatitis B virus cccDNA. Science (80-). 2014 Mar 14;343(6176):1221–8.

79. Hong X, Kim ES, Guo H. Epigenetic regulation of hepatitis B virus covalently closed circular DNA: Implications for epigenetic therapy against chronic hepatitis B. Hepatology. 2017 Dec 1;66(6):2066–77.

80. Vaillant A. Nucleic acid polymers: Broad spectrum antiviral activity, antiviral mechanisms and optimization for the treatment of hepatitis B and hepatitis D infection. Antiviral Res. 2016 Sep 1;133:32–40.

81. Cai CW, Lomonosova E, Moran EA, Cheng X, Patel KB, Bailly F, et al. Hepatitis B virus replication is blocked by a 2-hydroxyisoquinoline-1,3(2H, 4H)-dione (HID) inhibitor of the viral ribonuclease H activity. Antiviral Res. 2014 Aug;108(1):48–55.

82. Soriano V, Barreiro P, Benitez L, Peña JM, de Mendoza C. New antivirals for the treatment of chronic hepatitis B. Expert Opin Investig Drugs. 2017 Jul 3;26(7):843–51.

83. Maepa MB, Roelofse I, Ely A, Arbuthnot P. Progress and prospects of anti-HBV gene therapy development. Int J Mol Sci. 2015 Jul 31;16(8):17589–610.

84. Liang TJ, Block TM, McMahon BJ, Ghany MG, Urban S, Guo JT, et al. Present and future therapies of hepatitis B: From discovery to cure. Hepatology. 2015 Dec 1;62(6):1893–908.

85. Gish RG, Yuen MF, Chan HLY, Given BD, Lai CL, Locarnini SA, et al. Synthetic RNAi triggers and their use in chronic hepatitis B therapies with curative intent. Antiviral Res. 2015 Jul 11;121:97–108.

86. Ning X, Basagoudanavar SH, Liu K, Luckenbaugh L, Wei D, Wang C, et al. Capsid Phosphorylation State and Hepadnavirus Virion Secretion. J Virol. 2017 May 1;91(9).

87. Vanlandschoot P, Cao T, Leroux-Roels G. The nucleocapsid of the hepatitis B virus: A remarkable immunogenic structure. Antiviral Res. 2003 Oct;60(2):67–74.

88. Seitz S, Urban S, Antoni C, Böttcher B. Cryo-electron microscopy of hepatitis B virions reveals variability in envelope capsid interactions. EMBO J. 2007 Sep 19;26(18):4160–7.

89. Zlotnick A, Venkatakrishnan B, Tan Z, Lewellyn E, Turner W, Francis S. Core protein: A pleiotropic keystone in the HBV lifecycle. Antiviral Res. 2015 Jul 11;121:82–93.

90. Venkatakrishnan B, Zlotnick A. The Structural Biology of Hepatitis B Virus: Form and Function. Annu Rev Virol. 2016 Sep 29;3(1):429–51.

91. Roseman AM, Berriman JA, Wynne SA, Butler PJG, Crowther RA. A structural model for maturation of the hepatitis B virus core. Proc Natl Acad Sci U S A. 2005 Nov 1;102(44):15821–6.

92. Wynne SA, Crowther RA, Leslie AGW. The crystal structure of the human hepatitis B virus capsid. Mol Cell. 1999 Jun;3(6):771–80.

93. Heger-Stevic J, Zimmermann P, Lecoq L, Böttcher B, Nassal M. Hepatitis B virus core protein phosphorylation: Identification of the SRPK1 target sites and impact of their occupancy on RNA binding and capsid structure. Siddiqui A, editor. PLoS Pathog. 2018 Dec 1;14(12):e1007488.

94. Melegari M, Wolf SK, Schneider RJ. Hepatitis B Virus DNA Replication Is Coordinated by Core Protein Serine Phosphorylation and HBx Expression. J Virol. 2005 Aug 1;79(15):9810–20.

95. Kann M, Sodeik B, Vlachou A, Gerlich WH, Helenius A. Phosphorylation-dependent Binding of Hepatitis B Virus Core Particles to the Nuclear Pore Complex. J Cell Biol. 1999;145(1):45–55.

96. Li HC, Huang EY, Su PY, Wu SY, Yang CC, Lin YS, et al. Nuclear export and import of human hepatitis B virus capsid protein and particles. Ou JJ, editor. PLoS Pathog. 2010 Oct 28;6(10):e1001162.

97. Lin Y-Y, Liu C, Chien W-H, Wu L-L, Tao Y, Wu D, et al. New Insights into the Evolutionary Rate of Hepatitis B Virus at Different Biological Scales. J Virol. 2015 Apr 1;89(7):3512–22.

98. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: Patterns and determinants. Nat Rev Genet. 2008 Apr 4;9(4):267–76.

99. Shi W, Carr MJ, Dunford L, Zhu C, Hall WW, Higgins DG. Identification of novel inter-genotypic recombinants of human hepatitis B viruses by large-scale phylogenetic analysis. Virology. 2012 May 25;427(1):51–9.

100. Cui C, Shi J, Hui L, Xi H, Zhuoma A, Quni A, et al. The dominant hepatitis B virus genotype identified in Tibet is a C/D hybrid. J Gen Virol. 2002 Nov 1;83(11):2773–7.

101. Chekaraou MA, Brichler S, Mansour W, Gal F Le, Garba A, Dény P, et al. A novel hepatitis B virus (HBV) subgenotype D (D8) strain, resulting from recombination between genotypes D and E, is circulating in Niger along with HBV/E strains. J Gen Virol. 2010 Jun 1;91(6):1609–20.

102. Rodriguez-Frias F, Buti M, Tabernero D, Homs M. Quasispecies structure, cornerstone of hepatitis B virus infection: Mass sequencing approach. World J Gastroenterol. 2013 Nov 7;19(41):6995–7023.

103. Noguchi C, Ishino H, Tsuge M, Fujimoto Y, Imamura M, Takahashi S, et al. G to A hypermutation of hepatitis B virus. Hepatology. 2005 Mar;41(3):626–33.

104. Cullen BR. Role and Mechanism of Action of the APOBEC3 Family of Antiretroviral Resistance Factors. J Virol. 2006 Feb 1;80(3):1067–76.

105. Carman W, Thomas H, Domingo E. Viral genetic variation: hepatitis B virus as a clinical example. Lancet. 1993 Feb 6;341(8841):349–53.

106. Domingo E, Baranowski E, Nuñez JI, Ruiz-Jarabo CM, Sierra S, Molina N, et al. Cuasiespecies y evolución molecular de virus. OIE Rev Sci Tech. 2000 Apr;12(1):55–63.

107. Barzon L, Lavezzo E, Militello V, Toppo S, Palù G. Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci. 2011 Nov 14;12(11):7861–84.

108. Tsiatis AC, Norris-Kirby A, Rich RG, Hafez MJ, Gocke CD, Eshleman JR, et al. Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: Diagnostic and clinical implications. J Mol Diagnostics. 2010 Jul 1;12(4):425–32.

109. Chiu AP, Tschida BR, Sham TT, Lo LH, Moriarity BS, Li XX, et al. HBx-K130M/V131I promotes liver cancer in transgenic mice via AKT/FOXO1 signaling pathway and arachidonic acid metabolism. Mol Cancer Res. 2019 Jul 1;17(7):1582–93.

110. Ge Z, Tian T, Meng L, Song C, Yu C, Xu X, et al. HBV mutations in EnhII/BCP/PC region contribute to the prognosis of hepatocellular carcinoma. Cancer Med. 2019 Jun 1;8(6):3086–93.

111. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun 1;493:227–37.

112. Liu F, Chen L, Yu DM, Deng L, Chen R, Jiang Y, et al. Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: Correlation with antiviral efficacy. Gut. 2011 Sep 1;60(9):1269–77.

113. Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol Mol Biol Rev. 2012 Jun 1;76(2):159–216.

114. Chen L, Zhang Q, Yu D min, Wan M bin, Zhang X xin. Early changes of hepatitis B virus quasispecies during lamivudine treatment and the correlation with antiviral efficacy. J Hepatol. 2009 May;50(5):895–905.

115. Cheng Y, Guindon S, Rodrigo A, Wee LY, Inoue M, Thompson AJV, et al. Cumulative viral evolutionary changes in chronic hepatitis B virus infection precedes hepatitis B e antigen seroconversion. Gut. 2013 Sep;62(9):1347–55.

116. Lim SG, Cheng Y, Guindon S, Seet BL, Lee LY, Hu P, et al. Viral Quasi-Species Evolution During Hepatitis Be Antigen Seroconversion. Gastroenterology. 2007 Sep;133(3):951–8.

117. Domingo E, Gomez J. Quasispecies and its impact on viral hepatitis. Virus Res. 2007 Aug;127(2):131–50.

118. Homs M, Caballero A, Gregori J, Tabernero D, Quer J. Clinical Application of Estimating Hepatitis B Virus Quasispecies Complexity by Massive Sequencing: Correlation between Natural Evolution and On-Treatment Evolution. PLoS One. 2014;9(11):112306.

119. Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N. Application of next-generation sequencing technologies in virology. J Gen Virol. 2012 Sep 1;93(PART 9):1853–68.

120. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 Jun 1;17(6):333–51.

121. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008 Oct 9;26(10):1135–45.

122. Lakdawalla A, Fisher J, Ronaghi M, Fan JB. Cancer genome sequencing. In: Gelmann EP, Sawyers CL, Rauscher III FJ, editors. Molecular Oncology: Causes of Cancer and Targets for Treatment. Cambridge: Cambridge University Press; 2015. p. 1–9.

123. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 1;13(5):e1005457.

124. Illumina. Illumina Sequencing Overview [Internet]. Available from: https://www.well.ox.ac.uk/ogc/wp-content/uploads/2017/09/Illumina_Sequencing_Overview_15045845_D.pdf

125. Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, et al. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273–83.

126. R Development Core Team. R: a language and environment for statistical computing [Internet]. Available from: https://www.r-project.org/.

127. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

128. Pages H, Aboyoun P, Gentleman R, DebRoy S. Memory efficient string containers, string matching algorithms, and other utilities, for fast manipulation of large biological sequences or sets of sequences. [Internet]. 2018. Available from: https://bioconductor.org/packages/release/bioc/html/Biostrings.html

129. Magoč T, Magoč M, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957–63.

130. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998 Mar 1;8(3):186–94.

131. Cuadras CM. DISTANCE ANALYSIS IN DISCRIMINATION AND CLASSIFICATION USING BOTH CONTINUOUS AND CATEGORICAL VARIABLES. In: Statistical Data Analysis and Inference. Elsevier; 1989. p. 459–73.

132. Cuadras CM. A DISTANCE BASED APPROACH TO DISCR1MINANT ANALYSIS AND ITS PROPERTIES. In: Mathematics Preprint Series. Universitat de Barcelona; 1991.

133. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Jun;16(2):111–20.

134. Zheng CL, Fu YM, Xu ZX, Zou Y, Deng K. Hepatitis B virus core protein dimer-dimer interface is critical for viral replication. Mol Med Rep. 2019 Jan 1;19(1):262–70.

135. Lavanchy D, Kane M. Global Epidemiology of Hepatitis B Virus Infection. In: Human Diseases. 2016. p. 187–203.

136. Nayagam S, Thursz M, Sicuri E, Conteh L, Wiktor S, Low-Beer D, et al. Requirements for global elimination of hepatitis B: a modelling study. Lancet Infect Dis. 2016 Dec 1;16(12):1399–408.

137. Papatheodoridis G V., Chan HLY, Hansen BE, Janssen HLA, Lampertico P. Risk of hepatocellular carcinoma in chronic hepatitis B: Assessment and modification with current antiviral therapy. J Hepatol. 2015 Apr 1;62(4):956–67.

138. Zhang Y, Wu Y, Deng M, Xu D, Li X, Xu Z, et al. CD8 T-Cell Response-Associated Evolution of Hepatitis B Virus Core Protein and Disease Progress. jvi.asm.org 1 J Virol. 2018;92:2120–37.

139. Flisiak R, Jaroszewicz J, Łucejko M. siRNA drug development against hepatitis B virus infection. Vol. 18, Expert Opinion on Biological Therapy. Taylor and Francis Ltd; 2018. p. 609–17.

140. Durantel D. New treatments to reach functional cure: Virological approaches. Best Pract Res Clin Gastroenterol. 2017 Jun 1;31(3):329–36.

141. Kaur H, Bruno JG, Kumar A, Sharma TK. Aptamers in the therapeutics and diagnostics pipelines. Theranostics. 2018;8(15):4016–32.

142. Zhang Z, Zhang J, Pei X, Zhang Q, Lu B, Zhang X, et al. An aptamer targets HBV core protein and suppresses HBV replication in HepG2.2.15 cells. Int J Mol Med. 2014 Nov 1;34(5):1423–9.

143. Orabi A, Bieringer M, Geerlof A, Bruss V. An Aptamer against the Matrix Binding Domain on the Hepatitis B Virus Capsid Impairs Virion Formation. J Virol. 2015 Sep 15;89(18):9281–7.

144. Pleshakova TO, Kaysheva AL, Shumov ID, Ziborov VS, Bayzyanova JM, Konev VA, et al. Detection of hepatitis C virus core protein in serum using aptamer-functionalized AFM chips. Micromachines. 2019 Feb 15;10(2):129.

145. Percze K, Szakács Z, Scholz É, András J, Szeitner Z, Kieboom CHV Den, et al. Aptamers for respiratory syncytial virus detection. Sci Rep. 2017 Feb 21;7(1):42794.

146. Liu K, Luckenbaugh L, Ning X, Xi J, Hu J. Multiple roles of core protein linker in hepatitis B virus replication. PLoS Pathog. 2018 May 1;14(5):e1007085.

147. Ludgate L, Liu K, Luckenbaugh L, Streck N, Eng S, Voitenleitner C, et al. Cell-Free Hepatitis B Virus Capsid Assembly Dependent on the Core Protein C-Terminal Domain and Regulated by Phosphorylation. J Virol. 2016 Jun 15;90(12):5830–44.

148. Thursz M. Basis of HBV persistence and new treatment options. Hepatol Int. 2014 Sep 27;8(2):486–91.

149. Rajoriya N, Combet C, Zoulim F, Janssen HLA. How viral genetic variants and genotypes influence disease and treatment outcome of chronic hepatitis B. Time for an individualised approach? J Hepatol. 2017 Dec 1;67(6):1281–97.

150. Al-Qahtani AA, Al-Anazi MR, Nazir N, Abdo AA, Sanai FM, Al-Hamoudi WK, et al. The Correlation Between Hepatitis B Virus Precore/Core Mutations and the Progression of Severe Liver Disease. Front Cell Infect Microbiol. 2018 Oct 22;8:355.

151. Jia J, Li H, Wang H, Chen S, Wang M, Feng H, et al. Hepatitis B virus core antigen mutations predict post-operative prognosis of patients with primary hepatocellular carcinoma. J Gen Virol. 2017 Jun 1;98(6):1399–409.

152. Morgan AA, Rubenstein E. Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome. PLoS One. 2013;8(1):53785.

153. Guerrero-Murillo M, Gregori J. Characterizing viral quasispecies [Internet]. 2020. Available from: https://bioconductor.org/packages/release/bioc/vignettes/QSutils/inst/doc/QSutils-Diversity.html

154. Burns JS, Manda G. Metabolic pathways of the Warburg effect in health and disease: Perspectives of choice, chain or chance. Int J Mol Sci. 2017 Dec 19;18(12).

155. He X, Li J, Wu J, Zhang M, Gao P. Associations between activation-induced cytidine deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like cytidine deaminase expression, hepatitis B virus (HBV) replication and HBV-associated liver disease (Review). Mol Med Rep. 2015 Sep 1;12(5):6405–14.

156. Zhang Y, Qian H, Xu J, Gao W. ADAR, the carcinogenesis mechanisms of ADAR and related clinical applications. Ann Transl Med. 2019 Nov;7(22):686–686.

ANEXOS

11. ANEXOS

11.1 Anexo 1: Publicación

Publicación de los resultados del primer estudio como artículo original.

Referencia bibliográfica: Yll M, Cortese MF, Guerrero-Murillo M, Orriols G, Gregori J, Casillas R, González C, Sopena S, Godoy C, Vila M, Tabernero D, Quer J, Rando A, Lopez Martinez R, Esteban R, Riveiro-Barciela M, Buti M, Rodríguez-Frías F. Conservation and variability of hepatitis B core at different chronic hepatitis stages. *World J Gastroenterol* 2020; 26(20): 2584-2598

URL: https://www.wjgnet.com/1007-9327/full/v26/i20/2584.htm

DOI: https://dx.doi.org/10.3748/wjg.v26.i20.2584

Manuscrito completo:

World Journal of Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

World J Gastroenterol 2020 May 28; 26(20): 2584-2598

DOI: 10.3748/wjg.v26.i20.2584

ISSN 1007-9327 (print) ISSN 2219-2840 (online)

ORIGINAL ARTICLE

Basic Study

Conservation and variability of hepatitis B core at different chronic hepatitis stages

Marçal Yll, Maria Francesca Cortese, Mercedes Guerrero-Murillo, Gerard Orriols, Josep Gregori, Rosario Casillas, Carolina González, Sara Sopena, Cristina Godoy, Marta Vila, David Tabernero, Josep Quer, Ariadna Rando, Rosa Lopez-Martinez, Rafael Esteban, Mar Riveiro-Barciela, Maria Buti, Francisco Rodríguez-Frías

ORCID number: Marçal Yll (0000-0002-7030-3360); Maria Francesca Cortese (0000-0002-4318-532X); Mercedes Guerrero-Murillo (0000-0002-5556-2460); Gerard Orriols (0000-0002-1138-5909); Josep Gregori (0000-0002-4253-8015); Rosario Casillas (0000-0002-6758-6734); Carolina González (0000-0002-0169-5874); Sara Sopena (0000-0002-3309-5486); Cristina Godoy (0000-0001-5037-1916); Marta Vila (0000-0001-9303-5189); David Tabernero (0000-0002-1146-4084); Josep Quer (0000-0003-0014-084X); Ariadna Rando (0000-0003-4555-7286); Rosa López-Martínez (0000-0002-8450-6986); Rafael Esteban (0000-0001-5280-392X); Mar Riveiro-Barciela (0000-0001-9309-2052); Maria Buti (0000-0002-0732-3078); Francisco Rodríguez-Frías (0000-0001-9058-4641).

Author contributions: Rodríguez- Frías F designed the research; Cortese MF coordinated the research; Yll M and Cortese MF equally contributed to design the experiments; Yll M, Orriols G, Godoy C, Sopena S, Casillas R, González C, Vila M and Rando A performed the experiments; Yll M, Cortese MF, Gregori J and Guerrero-Murillo M analyzed data acquired during the experiments and interpreted the results; Yll M and Cortese MF drafted the manuscript; Cortese MF, Tabernero D, Lopez-Martinez R, Riveiro-

Marçal Yll, Maria Francesca Cortese, Gerard Orriols, Rosario Casillas, Carolina González, Sara Sopena, Cristina Godoy, Marta Vila, David Tabernero, Ariadna Rando, Rosa Lopez-Martinez, Francisco Rodríguez-Frías, Liver Pathology Unit, Departments of Biochemistry and Microbiology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona 08035, Spain

Marçal Yll, Maria Francesca Cortese, Mercedes Guerrero-Murillo, Josep Gregori, Rosario Casillas, Sara Sopena, Marta Vila, Josep Quer, Francisco Rodríguez-Frías, Liver Unit, Liver Disease Laboratory-Viral Hepatitis, Vall d'Hebron Institut Recerca-Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona 08035, Spain

Mercedes Guerrero-Murillo, Department of Microbiology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona 08035, Spain

Josep Gregori, Cristina Godoy, David Tabernero, Josep Quer, Rafael Esteban, Mar Riveiro- Barciela, Maria Buti, Francisco Rodríguez-Frías, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas, Instituto de Salud Carlos III, Madrid 28029, Spain

Rafael Esteban, Mar Riveiro-Barciela, Maria Buti, Liver Unit, Department of Internal Medicine, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona 08035, Spain

Corresponding author: Maria Francesca Cortese, PhD, Research Scientist, Liver Pathology Unit, Departments of Biochemistry and Microbiology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Passeig Vall d'Hebron 119-129, Barcelona 08035, Spain. maria.cortese@vhir.org

Abstract

BACKGROUND

Since it is currently not possible to eradicate hepatitis B virus (HBV) infection with existing treatments, research continues to uncover new therapeutic strategies. HBV core protein, encoded by the HBV core gene (*HBC*), intervenes in both structural and functional processes, and is a key protein in the HBV life cycle. For this reason, both the protein and the gene could be valuable targets for new therapeutic and diagnostic strategies. Moreover, alterations in the protein sequence could serve as potential markers of disease progression.

Barciela M, Buti M, Quer J, Esteban R and Rodríguez-Frías F critically reviewed the manuscript.

Supported by the Instituto de Salud Carlos III, Spain, the European Regional Development Fund, No. PI18/01436.

Institutional review board statement: The study was reviewed and approved by the Clinical Research Ethics Committee of Hospital Universitari Vall d'Hebron.

Conflict-of-interest statement: Josep Gregori is an employee of Roche Diagnostics, SL.

Data sharing statement: Next- generation sequencing data were submitted to the GenBank SRA database (BioProject accession number PRJNA625435).

ARRIVE guidelines statement: The authors have read the ARRIVE guidelines, and the manuscript was prepared and revised according to the ARRIVE guidelines.

Manuscript source: Invited manuscript

Received: February 28, 2020
Peer-review started: February 28, 2020
First decision: April 9, 2020
Revised: May 8, 2020
Accepted: May 19, 2020
Article in press: May 19, 2020
Published online: May 28, 2020

P-Reviewer: Said ZNA
S-Editor: Zhang H
L-Editor: A
E-Editor: Ma YJ

AIM

To detect, by next-generation sequencing, *HBC* hyper-conserved regions that could potentially be prognostic factors and targets for new therapies.

METHODS

Thirty-eight of 45 patients with chronic HBV initially selected were included and grouped according to liver disease stage [chronic hepatitis B infection without liver damage (CHB, $n = 16$), liver cirrhosis (LC, $n = 5$), and hepatocellular carcinoma (HCC, $n = 17$)]. HBV DNA was extracted from patients' plasma. A region between nucleotide (nt) 1863 and 2483, which includes *HBC*, was amplified and analyzed by next-generation sequencing (Illumina MiSeq platform). Sequences were genotyped by distance-based discriminant analysis. General and inter-group nt and amino acid (aa) conservation was determined by sliding window analysis. The presence of nt insertion and deletions and/or aa substitutions in the different groups was determined by aligning the sequences with genotype-specific consensus sequences.

RESULTS

Three nt (nt 1900-1929, 2249-2284, 2364-2398) and 2 aa (aa 117-120, 159-167) hyper-conserved regions were shared by all the clinical groups. All groups showed a similar pattern of conservation, except for five nt regions (nt 1946-1992, 2060-2095, 2145-2175, 2230-2250, 2270-2293) and one aa region (aa 140-160), where CHB and LC, respectively, were less conserved ($P < 0.05$). Some group-specific conserved regions were also observed at both nt (2306-2334 in CHB and 1935-1976 and 2402-2435 in LC) and aa (between aa 98-103 in CHB and 28-30 and 51-54 in LC) levels. No differences in insertion and deletions frequencies were observed. An aa substitution (P79Q) was observed in the HCC group with a median (interquartile range) frequency of 15.82 (0-78.88) *vs* 0 (0-0) in the other groups ($P < 0.05$ *vs* CHB group).

CONCLUSION

The differentially conserved *HBC* and HBV core protein regions and the P79Q substitution could be involved in disease progression. The hyper-conserved regions detected could be targets for future therapeutic and diagnostic strategies.

Key words: Hepatitis B virus; Hepatitis B core gene; Next-generation sequencing; Genetic conservation; Amino acid substitution; Gene therapy; Small interfering RNA

Core tip: New tools for hepatitis B virus infection treatment and follow-up are needed. Hepatitis B virus core protein has a key role in viral replication and persistence. Analysis of viral quasispecies by next-generation sequencing can identify conserved regions in viral genes or proteins that may serve as targets for new therapeutic and diagnostic strategies. Moreover, it may help identify prognostic markers of liver disease progression. Here, we detected hyper-conserved nucleotide and amino acid regions regardless of the clinical stage. Moreover, we observed several group-specific conserved and variable regions and an amino acid substitution that could be indicative of different disease progression.

Citation: Yll M, Cortese MF, Guerrero-Murillo M, Orriols G, Gregori J, Casillas R, González C, Sopena S, Godoy C, Vila M, Tabernero D, Quer J, Rando A, Lopez-Martinez R, Esteban R, Riveiro-Barciela M, Buti M, Rodriguez-Frias F. Conservation and variability of hepatitis B core at different chronic hepatitis stages. *World J Gastroenterol* 2020; 26(20): 2584-2598
URL: https://www.wjgnet.com/1007-9327/full/v26/i20/2584.htm
DOI: https://dx.doi.org/10.3748/wjg.v26.i20.2584

INTRODUCTION

Hepatitis B virus (HBV) is a small virus with a specific tropism for the liver. It belongs

to the *Hepadnaviridae* family. Despite the existence of effective preventive vaccines, an estimated 257 million people worldwide live with chronic HBV infection and more than 880000 people die every year of HBV-related complications such as liver cirrhosis (LC) and hepatocellular carcinoma (HCC)[1].

HBV is an enveloped virus equipped with 3.2 kb of partially double-stranded circular DNA produced by the reverse transcription of an RNA intermediate known as pregenomic RNA[2]. This ribonucleic intermediate is produced from a viral DNA molecule that interacts with cellular (histone and non-histone) and viral proteins, forming a "mini-chromosome" known as covalently closed circular DNA (cccDNA) that remains in hepatocyte nuclei for the rest of the cell's life[3]. Although current antiviral therapy can control viral replication, it is not capable of interfering with the formation or persistence of cccDNA, rendering HBV infection eradication impossible. This mini-chromosome could even be a source of HBV reactivation after clinical resolution and HBsAg seroclearance[4]. Due to persistent infection, up to 1% of Caucasian patients with noncirrhotic chronic HBV infection have been found to develop HCC[5].

Gene therapy has emerged as one of the most promising strategies for blocking disease progression, and results from studies investigating the potential of small interfering RNA (siRNA) systems as adjuvant therapy are encouraging[6]. SiRNA is a double-stranded noncoding RNA [with an optimal length of 21 nucleotides (nt)] that interacts with target messenger RNA, promoting its degradation and silencing of the gene[7].

HBV reverse transcriptase lacks 3' to 5' proofreading activity, which leads to viral genome variability comparable to that observed in an RNA virus[8]. This genetic variability is further increased by inter- and intra-genotype recombination events[9]. In short, HBV circulates as a complex mixture of closely related genetic variants (haplotypes) known as quasispecies[10].

The HBV core protein (HBc) [encoded by the HBV core gene (*HBC*) from the PreCore/Core open reading frame (ORF)] is essential for viral replication. It is a structural 21-kDa protein that self-assembles to create dimers that assemble in hexamers forming the icosahedral viral capsid[11,12]. It has 183 amino acids (aa) (185 for genotype A) with a N-terminal domain and a C-terminal domain (CTD) connected through a linker region. The N-terminal domain ranges from aa position 1 to 149 (including the linker region aa 140 to 149) and constitutes the α helix-rich assembly domain[13]. The CTD is shorter (aa 150 to 183, or 185 for genotype A) and constitutes the functional domain[14]. The CTD allows HBc to intervene in a multitude of processes such as subcellular traffic, viral genome release, capsid assembly and transport, RNA metabolism, and viral pregenomic RNA reverse transcription[14]. Considering just how essential this protein is for viral replication, it could be an optimal target for gene therapy. Moreover, mutations in HBc may have different roles in liver disease progression, positioning them as potentially useful prognostic genetic markers.

Next-generation sequencing (NGS) is a highly sensitive technique for studying viral quasispecies; it is capable of detecting highly conserved regions of the HBV genome, regardless of genome or clinical stage[15]. Moreover, it supports the identification and quantitative determination[16] of specific variants that could be used as markers to predict prognosis and treatment response in patients with HBV infection.

The aim of this study was to apply NGS to analyse HBc conservation and variability at the nt and aa levels in patients with different stages of chronic HBV infection in order to identify hyper-conserved regions of the *HBC* gene that could be a target for gene therapy and to determine possible prognostic factors of disease progression

MATERIALS AND METHODS

Patients and samples

The study was reviewed and approved by the Clinical Research Ethics Committee of Hospital Universitari Vall d'Hebron (PR(AG)146/2020). No animals were used.

Forty-five patients with chronic HBV infection were recruited from members of the general population seen at the outpatient clinic at Vall d'Hebron University Hospital in Barcelona, Spain. They tested negative for hepatitis D virus, hepatitis C virus, and human immunodeficiency virus, and had a viral load > 3 log IU/mL, which is the limit of polymerase chain reaction (PCR) amplification sensitivity. HBV serological markers such as the surface antigen (HBsAg), the e antigen (HBeAg), and anti-HBe antibodies were tested using commercial chemiluminescent assays on a COBAS 8000 analyzer (Roche Diagnostics, Rotkreuz, Switzerland). HBV DNA was quantified by

real-time PCR with a detection limit of 10 IU/mL (COBAS 6800, Roche Diagnostics). Patients were divided into 3 clinical groups according to liver disease stage determined by biopsy or diagnostic imaging in line with the EASL guidelines[16]: Chronic HBV infection without liver damage (CHB group), chronic HBV infection with liver cirrhosis (LC group), and chronic HBV infection with hepatocellular carcinoma (HCC group).

HBC gene amplification and NGS

HBV DNA was extracted from 200 μL of serum using the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. The region of interest was amplified through a 3-step nested PCR protocol (Figure 1). The first step (PCR1) covered a large region between nt 1774-2930 that includes the *HBC* gene (nt 1901-2464 for genotype A and 1901-2458 for other genotypes). As the Illumina MiSeq platform (Illumina, San Diego, CA, United States) allows read lengths of up to 600 bp, the following amplification steps were performed by dividing *HBC* into 2 amplicons (amplicon 1 = nt 1863-2317 and amplicon 2 = nt 2205-2483), which overlapped in a 112 nt-long portion (PCR2). The M13-tail, added in step 2, was used for the last step (PCR3), which introduced a 10 nt-long sample-specific multiplex identifier. All the PCR steps were performed using high-fidelity Pfu Ultra II DNA polymerase (Stratagene, Agilent Technologies, Santa Clara, CA, United States). The primers and protocols are reported in Table 1.

The final PCR products were purified with Agencourt AMPure XP magnetic beads (Beckman Coulter, Beverly, LA, United States) and their quality verified using the Agilent 2200 TapeStation System and D1000 ScreenTape kit (Agilent Technologies, Waldbronn, Germany).

Purified amplicons were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific-Life Technologies, Austin, TX, United States) and pooled to guarantee that the 2 amplicons for each patient were adequately represented in the analysis (2.5x for amplicon 1 and 1x for amplicon 2, due to their different lengths). The amplicon pool was sequenced by NGS on the Illumina MiSeq platform.

The reads obtained underwent an in-house bioinformatics filtering procedure based on R scripts[17], as previously described by our group[18]. For each amplicon, a group of unique sequences (haplotypes) forming the viral quasispecies was obtained. All sequences that did not match in the overlapping 112-nt region between amplicon 1 and 2 were discarded.

The bioinformatics methods used in this study were reviewed by Mercedes Guerrero-Murillo from the Microbiology Department at Vall d'Hebron Hospital (Barcelona, Spain) and by Dr. Josep Gregori from the Liver Disease Viral Hepatitis Laboratory at Vall d'Hebron Hospital (Barcelona, Spain), CIBERehd research group, and Roche Diagnostics SL.

Genotyping of the haplotypes

The amplicons from each patient were aligned with the same region of the respective amplicons extracted from 106 full-length HBV genome sequences representative of genotypes A to J obtained from the NCBI GenBank (Supplementary Table 1). Genotyping was conducted by applying distance-based discriminant analysis (DB rule)[19,20], which considers the inter- and intra-class variability of all genotypes. Genetic distances were computed according to the Kimura-80 model[21].

Conservation and mutation analysis

Sequence conservation at nt and aa levels was determined by calculating the information content (IC) of each position in a multiple alignment of all haplotypes detected with a frequency > 0.25.

This analysis calculates the mean IC for windows of 25 nt (or 10 aa), starting from the first position in the multiple alignment and moving forward in steps of 1[22]. The hyper-conserved regions were detected by aligning all haplotypes, regardless of clinical stage. Differences in sequence conservation between the groups were determined by comparing IC values.

To identify specific nt insertions and deletions (indels) and aa substitutions that could discriminate between the groups, haplotype sequences were aligned with their genotype-specific consensus sequence. Consensus was obtained by aligning the sequences of the subgenotypes of interest extracted from the 106 full-length HBV genome sequences. Polymorphisms were identified by aligning haplotype sequences with a population consensus sequence and discarded.

Statistical analysis

Sequence conservation differences between the groups in the sliding windows were analysed using the Wilcoxon–Mann–Whitney test. Frequencies of aa changes detected

Table 1 Primer design and polymerase chain reaction protocols for each amplified region

	PCR	Primer	Primer sequence (5'-> 3')	Amplified region	Protocol
1st step	PCR1	Forward	TAGGAGGCTGTAGGCATA	1774-2930	95 °C 5 min; (95°C 20 s, 49 °C 20 s, 72 °C 15 s) × 35 cycles; 72 °C 3 min
		Reverse	GGAAAGAATCCCAGAGG		
2nd step	PCR2 A.1	Forward A.1	**GTTGTAAAACGACGGCCAGT**TTCAAGCCTCCAAGCTGT	1863-2317	95 °C 2 min; (95 °C 20 s, 58 °C 20 s, 72 °C 15 s) × 35 cycles; 72 °C 3 min
		Reverse A.1	**CACAGGAAACAGCTATGACC**GATAGGGGCATTTGGTGGTCT		
	PCR2 A.2	Forward1 A.2	**GTTGTAAAACGACGGCCAGT**GGTTTCATATTTCTTGCC	2205-2483	95 °C 2 min; (95 °C 20 s, 50 °C 20 s, 72 °C 15 s) × 35 cycles; 72 °C 3 min
		Forward2 A.2	**GTTGTAAAACGACGGCCAGT**GGTTTCACATTTCCTGTC		
		Forward3 A.2	**GTTGTAAAACGACGGCCAGT**GGTTTCACATTTCTTGTC		
		Forward4 A.2	**GTTGTAAAACGACGGCCAGT**GGTTTCACATTTCCTGCC		
		Reverse A.2	**CACAGGAAACAGCTATGACC**TCCCACCTTATGAGTCCAAG		
3rd step	PCR3	Forward (specific per sample)	**GTTGTAAAACGACGGCCAGT**+specific 10 nt MID		95 °C 2 min; (95 °C 20 s, 60 °C 20 s, 72 °C 15 s) × 20 cycles; 72 °C 3 min
		Reverse (specific per sample)	**CACAGGAAACAGCTATGACC**+specific 10 nt MID		

Bold nucleotides indicate the M13 sequence. Forward primers in PCR2-A2 were multiplexed at the same concentration to cover all HBV genotypes. The protocols of amplification are reported. A.1: Amplicon 1; A.2: Amplicon 2; PCR: Polymerase chain reaction; MID: Multiplex identifier.

were compared with the Kruskal-Wallis test and described as median and interquartile range (IQR). All analyses were performed in R version 3.2.3. $P < 0.05$ was considered significant.

RESULTS

Patients characteristics and NGS results

Of the 45 patients with chronic hepatitis initially included in the study, 38 passed the sequencing quality filters and had correctly overlapping amplicons 1 and 2. After application of the quality filters, a median (IQR) of 133156.5 (85961.25-605212) and 66571 (25958.5-2301225) sequences per patient were obtained respectively for amplicon 1 and amplicon 2. NGS data were submitted to the GenBank SRA database (BioProject accession number PRJNA625435; BioSample accession numbers are reported in Supplementary Table 2). In the clinical groups, there were 16 patients with CHB, 5 with LC, and 17 with HCC. The clinical and viral characteristics (including genotypes) are reported in Table 2.

Sequence conservation at the nt level

Sequence conservation was studied by applying a sliding window analysis to the entire *HBC* sequence overlapping the 2 amplicons at the common 112 nt-long portion. No differences in IC were observed on analyzing the sequences by haplotype considering or not their relative frequency (Figure 2A). Considering the IC of all the nt-sequence haplotypes obtained (regardless of clinical group), we identified 3 hyper-conserved regions (nt 1900-1929, 2249-2284, and 2364-2398, Figure 2B). Most of the nt positions within these regions yielded the maximum IC value of 2 bits (100% conservation).

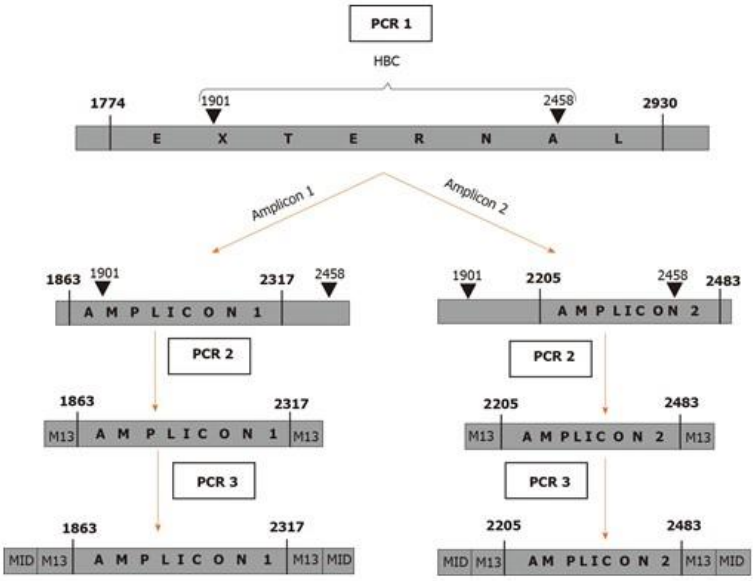


Figure 1 Schematic summary of the 3 amplification steps. In the first amplification step (PCR1), a large region was amplified. In the following step (PCR2), the region was divided into 2 amplicons that overlapped in a 112 nucleotide-long portion. In the third step (PCR 3) a sample identifier (MID) was added. PCR: Polymerase chain reaction; MID: Multiplex identifier.

On comparing the IC of each clinical group by haplotype, the HCC and LC groups showed similar conservation patterns; CHB was notably associated with the lowest level of conservation, mainly evident in 5 regions: nt 1946-1992, 2060-2095, 2145-2175, 2230-2250, and 2270-2293 ($P < 0.05$, Figure 3A). Three group-specific conserved regions were detected: 1 in the CHB group (nt 2306-2334) and 2 in the LC group (nt 1935-1976 and 2402-2435; Figure 3B). Most of the nt positions within these regions yielded the maximum IC value of 2 bits (100% conservation).

Sequence conservation at the aa level

The aa sequences of the haplotypes were translated from their respective nt sequences using the *HBC* reading frame.

Sliding window analysis of the aa haplotypes of the 38 patients by haplotype and haplotype frequency (Figure 4A) showed that the HBc protein was highly conserved throughout its sequence except for the central region (between aa 50 and 100), where conservation was slightly decreased. Two common hyper-conserved regions were detected: 1 between aa 117-120 and 1 between aa 159-167 (Figure 4B). All the aa in these regions had a conservation of around 100% (4.32 bits).

On analyzing aa conservation by haplotype in relation to clinical stage, the 3 groups showed a similar pattern, except for a region between aa 140 and 160, which was less conserved in the LC group compared with the CHB and HCC groups ($P < 0.05$, Figure 5A). Again, 3 group-specific conserved aa regions were detected: 1 in the CHB group (aa 98-103) and 2 in the LC group (aa 28-30 and 51-54, Figure 5B). All the aa in these regions had a conservation of around 100% (4.32 bits).

nt indels and aa changes

nt indels and aa changes were identified by aligning the patients' haplotypes with their genotype-specific consensus sequence.

In the CHB group, 8/16 patients had indels in *HBC*, *vs* 2/17 in the HCC group and 1/5 in the LC group. The indels consisted of the insertion or deletion of one nt at positions 1951 or 2085 (a thymine in 1951 and a guanine in 2085; Table 3). In all cases, a truncated HBc protein was produced. However, due to the limited number of patients, no statistical differences were observed on comparing the frequencies between the groups.

On analysing the presence of aa changes, we identified the aa substitution P79Q

Table 2 Main clinical and viral characteristics of hepatitis B virus-infected patients enrolled in the study

Median [IQR]	CHB (*n* = 16)	HCC (*n* = 17)	LC (*n* = 5)	*P* value
Age	38.5 [33.5-46.5]	67 [58-69]	56 [48-66]	0.002
Viral load (log IU/mL)	6.8 [5.7-8.0]	5.5 [4.7-6.7]	5.7 [4.8-6.2]	NS
ALT	56.5 [41.25-180.5]	70 [47-212]	46 [43-79]	NS
AST	56 [34.75-124]	120.5 [59-163.5]	66.45 [48.675-84.225]	NS
Platelets (10^9/L)	183 [161.5-226]	136 [98.5-255]	81.5 [61.25-101.75]	NS
Proportion				
Gender (male)	11/16	15/17	3/5	
HBeAg (positive)	8/16	3/17	0/5	
Genotype, % (*n*)				
A	18.8 (3)	5.9 (1)	20.0 (1)	
C	37.5 (6)	5.9 (1)	20.0 (1)	
D	25.0 (4)	64.7 (11)	40.0 (2)	
D/A	(0)	5.9 (1)	0.0 (0)	
D/E	6.3 (1)	11.8 (2)	0.0 (0)	
E	6.3 (1)	0.0 (0)	20.0 (1)	
F	6.3 (1)	5.9 (1)	0.0 (0)	

D/E and D/A indicate mixtures of the 2 genotypes. The frequency of each genotype within the clinical groups is reported as percentage (%) and number of patients (*n*). CHB: Chronic hepatitis B infection without liver damage; HCC: Hepatocellular carcinoma; LC: Liver cirrhosis; ALT: Alanine aminotransferase (normal value < 40 IU/mL); AST: Aspartate aminotransferase (normal value < 40 IU/mL); IQR: Interquartile range; NS: No-statistical *P* value.

(proline to glutamine) in the HCC group with a median (IQR) frequency of 15.82 (0-78.9) *vs* (0-0) in the CHB group ($P < 0.05$) and 0 (0-0) in the LC group (Figure 6).

DISCUSSION

The HBc protein, encoded by the *HBC* gene, is a key element in viral replication and disease progression and is involved in both structural and functional processes. Studying gene and protein sequences in patients with different clinical stages of HBV infection could provide important information on the pathogenic role of this protein. Moreover, the identification of hyper-conserved regions at both nt and aa levels could help develop new therapeutic approaches, including gene therapy. In this study, we used NGS to analyse *HBC* quasispecies in a group of patients with chronic HBV infection stratified by liver disease stage.

First, we studied quasispecies conservation to search for hyper-conserved nt and aa regions regardless of clinical stage or viral genotype. Current treatment based on nucleos(t)ide inhibitors does not affect cccDNA levels or transcriptional activity and therefore cannot eliminate HBV infection. This viral mini-chromosome supports the continuous expression of viral antigens that possibly contribute to disease progression, even in the presence of drug-induced viral suppression[22].

New therapeutic approaches are thus required to control HBV expression, and the targeted delivery of siRNA is one of the most promising approaches under investigation[23]. Several siRNAs are currently being tested against X and S ORFs. A study conducted in chimpanzees showed that multiple injections of ARB-1467 (a mixture of 3 interfering RNAs targeting both X and S ORFs[24]) led to a 90% reduction in HBsAg levels and a 50% reduction in cccDNA within 28 d of treatment[25]. None of the molecules currently available, however, target *HBC*, which considering its role in viral replication could be a valuable target for siRNA-based therapies.

In this study, we analysed quasispecies conservation of the entire *HBC* gene in patients infected by different HBV genotypes and with different clinical stages of disease in order to identify hyper-conserved regions that might be useful for pangenotypic and panclinical RNA silencing strategies. On analyzing nt conservation for the group of 38 patients, we detected 3 shared hyper-conserved regions, namely the start codon of *HBC* expression (nt 1900-1929), a portion with 2 CD8 epitopes (HLA-A24 and A3303) (nt 2249-2284)[26], and an arginine-rich portion of the CTD (nt 2364-2398). All 3 sequences could be valuables targets for a new gene silencing

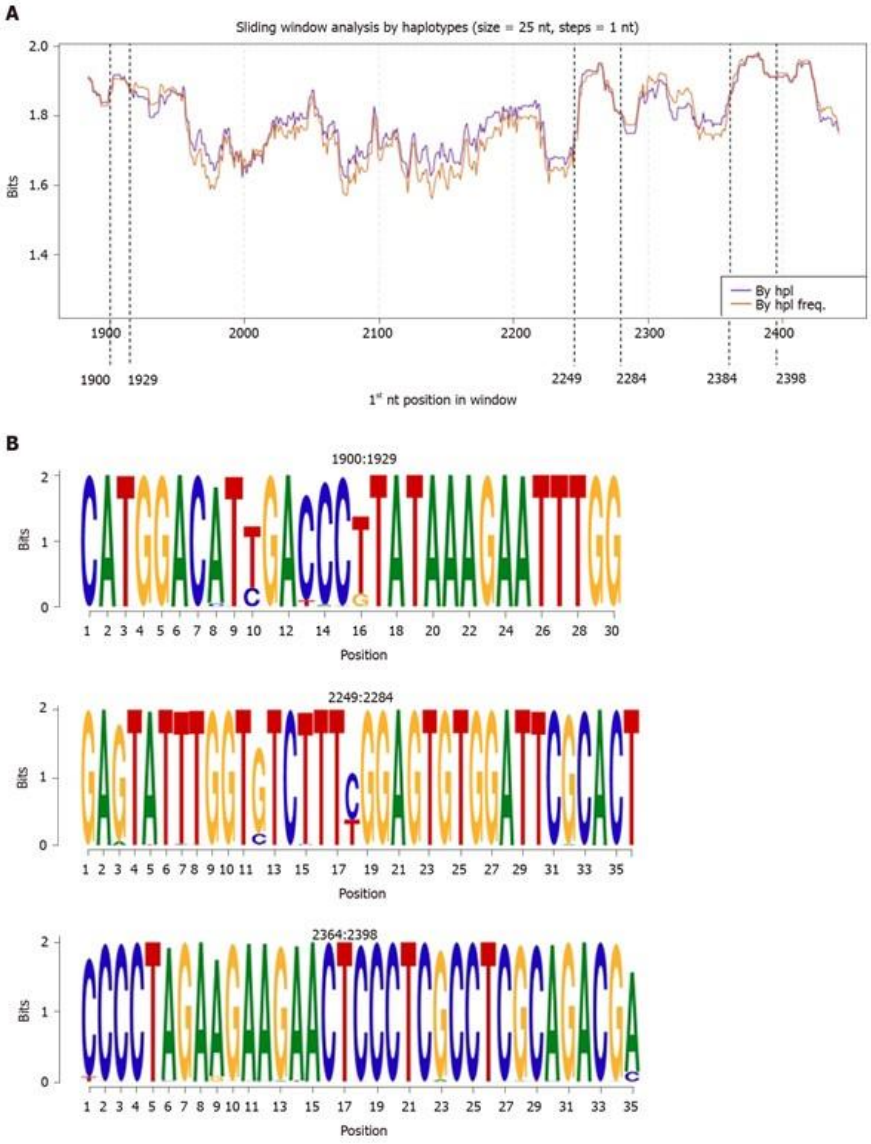


Figure 2 Information content analysis at nucleotide level. A: Sliding window analysis of Hepatitis B core gene performed by aligning the quasispecies haplotypes for all 38 patients with and without considering their relative frequency. Each point on the graph represents the mean information content (in bits) of the 25-nucleotides windows, with forward displacement of 1 nucleotide step between windows. The purple line shows the analysis by haplotype (By hpl), which is the mean information content obtained from the multiple alignments of all quasispecies haplotypes. The orange line represents the analysis by haplotype frequency (By hpl freq), which is the mean information content from the multiple alignments of all the patients' quasispecies haplotypes considering their relative frequency. The dashed lines indicate the 3 common hyper-conserved regions observed, with reporting of their positions. B: Representation of detected hyper-conserved regions as sequence logos (with reporting of nucleotide positions). The relative sizes of the letters in each stack indicate their relative frequencies at each position within the multiple alignments of nucleotide haplotypes. The total height of each stack of letters depicts the information content of each nucleotide position, measured in bits (Y-axis): from minimum (0) to maximum conservation (2). By hpl: Analysis by haplotype; By hpl freq: Analysis by haplotype frequency; nt: Nucleotide.

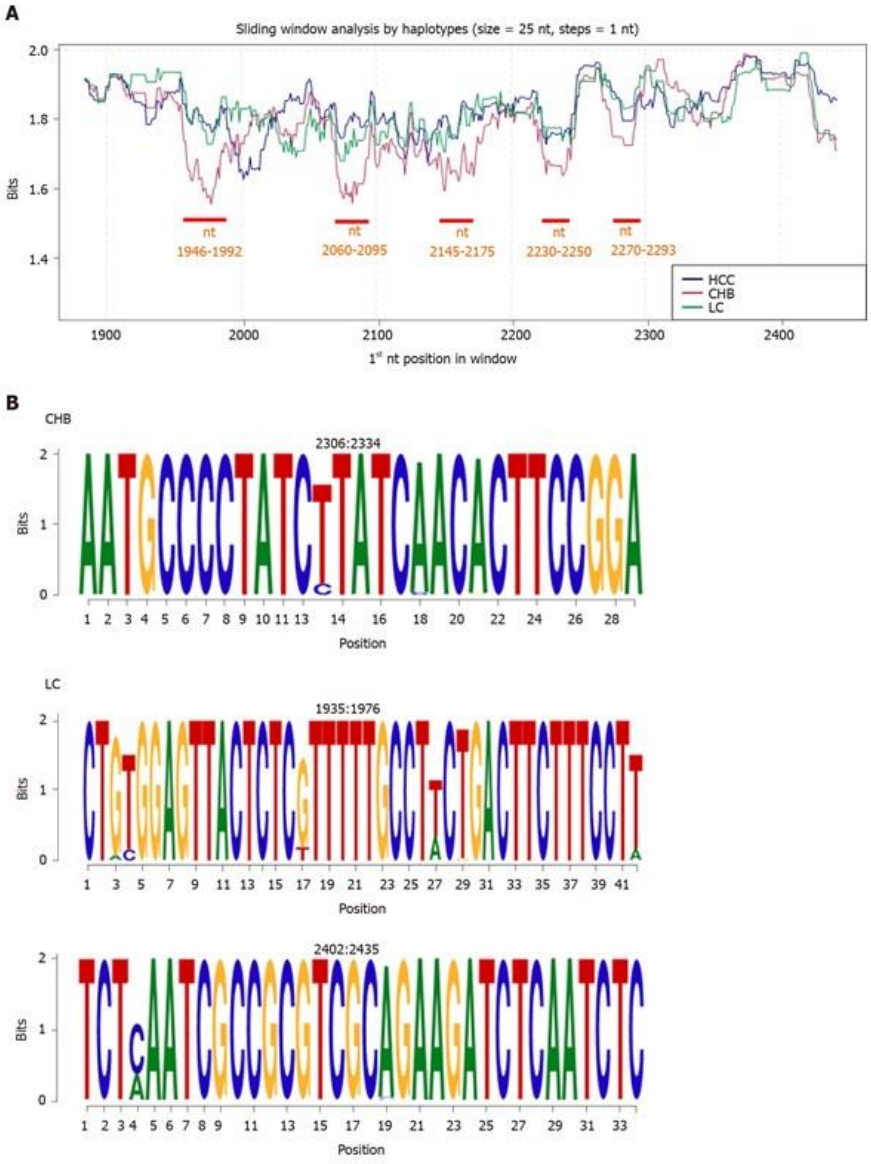


Figure 3 Information content analysis at nucleotide level by clinical stage group. A: By-haplotype sliding window analysis of the Hepatitis B core gene according to different clinical groups (HCC in blue, CHB in red, and LC in green). The portions and positions where CHB showed lower levels of conservation than the others (*P* < 0.05) are shown in red. B: Representation of the information content of CHB- and LC-specific conserved nucleotide regions as sequence logos. Positions are reported at the top of each logo. CHB: Chronic hepatitis B infection without liver damage; HCC: Hepatocellular carcinoma; LC: Liver cirrhosis; nt: Nucleotide; *P*: *P* value.

strategy.

At the aa level, we observed 2 common hyper-conserved regions (aa 117-120 and 159-167), which fell into the second and third hyper-conserved nt portions (nt 2249-2284 and 2364-2398 respectively). The CTD plays a key role in HBc function. It contains the 4 arginine-rich domains (RRR aa 150-152, RRR aa 157-159, RRRR aa 164-

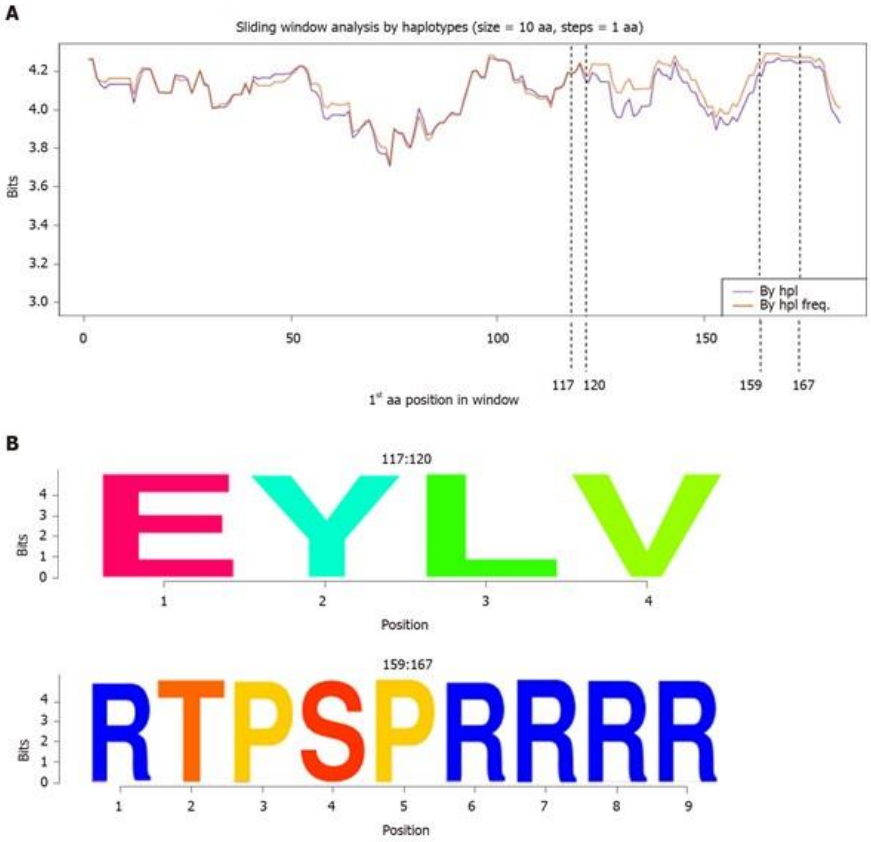


Figure 4 Information content analysis at amino acid level. A: Sliding window analysis of the Hepatitis B core protein sequence for all 38 patients with and without consideration of relative frequency. Each point on the graph is the result of the mean information content (in bits) of the 10-amino acid in size windows, with forward displacement between them of 1 amino acid step. The purple line represents the information content of all the quasispecies haplotypes (By hpl) whereas the orange line indicates the information content considering haplotype frequency (By hpl freq). The dashed lines show the 2 common amino acid hyper-conserved regions observed, with reporting of their positions. B: Representation of amino acid hyper-conserved regions detected as sequence logos (with reporting of amino acid positions). The relative sizes of the letters in each stack indicate their relative frequencies at each position within the multiple alignments of amino acid haplotypes. The total height of each stack depicts the information content of each amino acid position, measured in bits (Y-axis); range: 0 bits (0% conservation) to 4.32 bits (100% conservation). By hpl: Analysis by haplotype; By hpl freq: Analysis by haplotype frequency; aa: Amino acid.

167, and RRRR aa 172-175) that guarantee adequate protein subcellular localization acting as nuclear or cytoplasmic localization signals[27]. The second hyper-conserved aa region (aa 159-167) included one of these arginine-rich domains.

The high degree of sequence conservation observed in HBc may be indicative of its importance in protein function, positioning it as a possible target for diagnostic and therapeutic strategies. Recent studies have defined HBV core-related antigen (HBcrAg, which consists of HBc, HBeAg, and HBV p22 protein) as a promising serological viral marker, particularly for patients with low viral loads, such as treated patients[28] and patients with chronic HBeAg-negative infection[29]. This potential marker, however, has some limitations related to its high limits of detection (2 log IU/mL) and quantification (3.7 log IU/mL). The hyper-conserved regions observed in our study could be used as targets to improve HBc detection technology.

Aptamers are emerging as a promising diagnostic and therapeutic option for different diseases[30]. These molecules consist of single-strand DNA or RNA with high affinity and specificity and no toxicity or immunogenicity[31]. *In vitro* testing of an aptamer generated using the matrix domain of HBV (located in the large surface protein L and related to the nucleocapsid envelope) resulted in a 50% decrease in

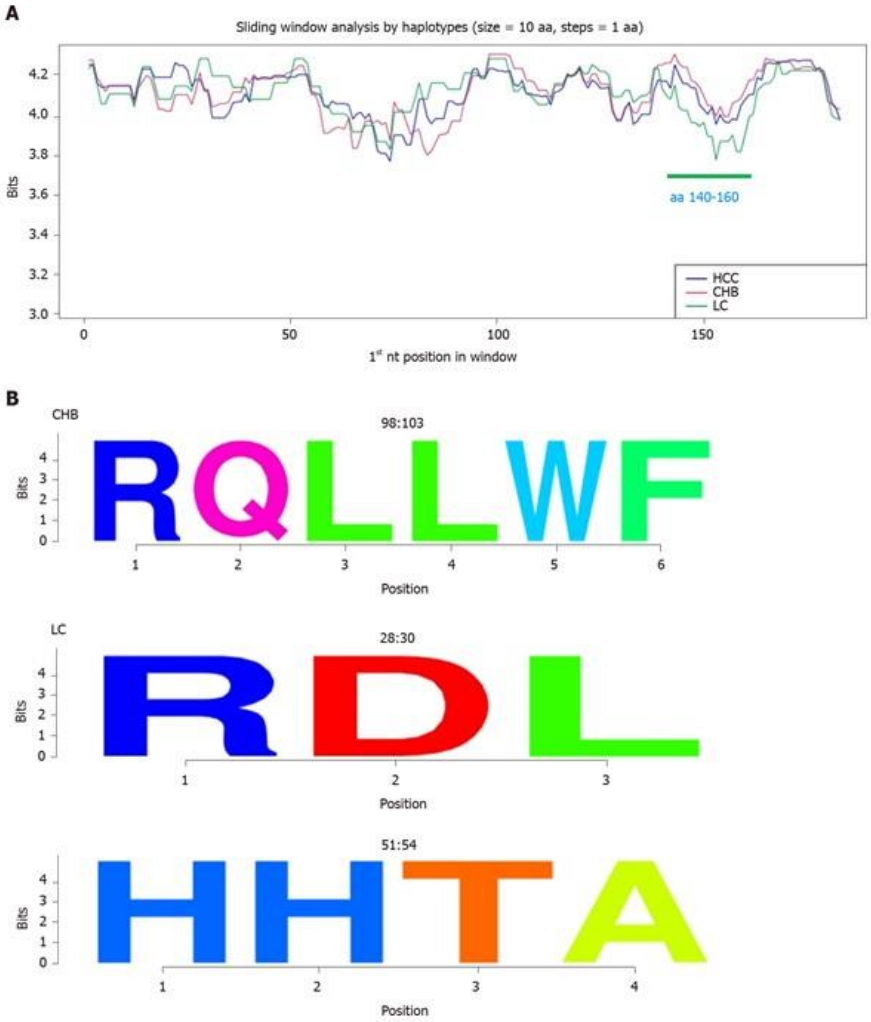


Figure 5 Information content analysis at amino acid level by clinical group. A: Sliding window analysis of the Hepatitis B core protein by haplotype between the different clinical groups (HCC in blue, CHB in red, and LC in green). The green horizontal line corresponds to the region where LC group is less conserved compared to the CHB and HCC groups ($P < 0.05$). B: Representation of CHB- and LC-specific conserved amino acid regions as sequence logos. Positions are reported at the top of each logo. CHB: Chronic hepatitis B infection without liver damage; HCC: Hepatocellular carcinoma; LC: Liver cirrhosis; aa: Amino acid; *P*: *P* value.

HBV titre in treated cell supernatants[44]. In another study, an aptamer targeting *HBC* resulted in a reduction in extracellular HBV DNA by interfering with nucleocapsid assembly[44]. Again, the hyper-conserved regions detected in our study could be novel targets for aptamer-based strategies that might work independently of clinical stage or HBV genotype. They could be also used to elaborate a new HBV detection system, as has been done with hepatitis C virus[43] and syncytial viruses[44].

On analyzing nt and aa conservation in relation to clinical stage of HBV infection, all 3 groups showed similar patterns at the aa level, although the HBV quasispecies in the LC group was slightly less conserved (mainly between aa 140-160). At the nt level, conservation was lower in the CHB group than in the other 2 groups, largely in the 5 regions between nt 1946-1992, 2060-2095, 2145-2175, 2230-2250, and 2270-2293. This finding could be consistent with the high replication rate of HBV during this clinical stage. Moreover, the first variable region (nt 1946-1992) includes three CD8 HLA

Table 3 Relative frequencies of nucleotide insertions/deletions detected

Clinical stage (*n*/total)	Patient	Relative frequency (% of mutated haplotypes)	
		1951 (1 nt: T)	2085 (1 nt: G)
CHB (8/16)	1	836 (8.7)	
	2		17.12 (40)
	3		3.19 (5)
	4	0.37 (5.9)	
	9	202 (882)	
	10		1.34 (50)
	12		1.04 (10)
	13	2.79 (22.22)	
HCC (2/17)	28		0.78 (4)
	33		242 (48)
LC (1/5)	34		17.42 (19.2)

The table shows the relative frequency of insertions/deletions, together with the percentage (%) of mutated haplotypes per patient. Only patients carrying these mutations were included in the table. CHB: Chronic hepatitis B infection without liver damage; HCC: Hepatocellular carcinoma; LC: Liver cirrhosis; T: Thymine; G: Guanine; nt Nucleotide.

epitopes (epitopes B5101, B3501, and B0702 at nt positions 1958-1982)[36], suggesting an attempt at immune evasion. Although the CHB group had the lowest levels of sequence conservation, we detected 2 group-specific conserved regions: aa 98-103 and nt 2306-2334. The nt region included the first 5 aa of the linker region, suggesting thus an important role for this region, which is involved in capsid assembly[35,36] and viral DNA synthesis[36]. In the LC group we detected 2 exclusively conserved nt regions (nt 1935-1976 and 2402-2435, which would translate respectively to aa 11-25 and 167-178) and 2 exclusively conserved aa regions (aa 28-30 and 51-54). The first related regions (nt 1935-1976 and aa 28-30) included portions of HBc (aa 14-18 and aa 23-39 respectively) that are involved in capsid assembly and envelopment and virion production[37], highlighting the importance of these functions in LC. The second LC-specific nt region (nt 2402-2435) contained an arginine-rich domain of the CTD when translated.

The identification of group-specific conserved regions suggests different evolutionary histories that may have different effects on disease progression. Further studies, however, are needed to prove the association between these regions and different clinical stages and to investigate their role in liver disease progression.

Considering the risk and severity of disease progression, identification of prognostic factors would be of great help. A number of studies have focused on detecting aa changes possibly related to different clinical stages. The mutations T1753C and A1762T/G1764A (K130M/V131I in HBx) of basal core promoter, for example, were identified as possible prognostic markers for HCC[38,39], while HBc aa mutations F24Y, E64D, E77Q, A80I/T/V, L116I, and E180A were linked to the development of cirrhosis and HCC[40]. In our study, one of the aa changes detected, P79Q, was exclusively observed in the HCC group. Mutations at this position have been found to be slightly associated with tumour relapse after resection[41]. More *in vitro* studies are required to investigate the role of the P79Q mutation in liver disease progression.

One limitation of our study is that we were not able to include large numbers of patients with different stages of liver disease due to the limits of PCR detection. This was particularly evident in the LC group, which was very small. Larger samples are needed to confirm our results. Moreover, although the Illumina MiSeq platform offers long read lengths, they are not sufficient to cover the entire *HBC* gene, making it necessary to divide it into 2 partially overlapping amplicons. Nonetheless, these 2 fragments were treated as independent samples during sequencing and subsequently analysed as such.

In summary, we have identified a number of nt and aa hyper-conserved regions that could be valuable targets for new therapeutic and diagnostic strategies. The role of group-specific conserved regions in liver disease progression requires further analysis. The P79Q substitution could be a possible prognostic factor for HCC. *In vitro* studies, however, are required to determine whether this change might affect viral

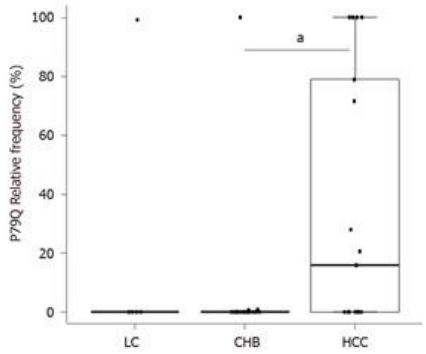


Figure 6 Relative frequency of P79Q substitution in the 3 clinical groups. Each dot represents a patient. The Bonferroni-corrected *P* value was calculated by Kruskal-Wallis test with posthoc Dunn multiple comparison test. ($^aP < 0.05$). CHB: Chronic hepatitis B infection without liver damage; HCC: Hepatocellular carcinoma; LC: Liver cirrhosis; *P*: *P* value; P79Q: Proline to glutamine in position 79.

replication and to investigate associations between cellular damage and onset of HCC.

ARTICLE HIGHLIGHTS

Research background

Despite the existence of effective preventive vaccines, an estimated 257 million people worldwide live with chronic hepatitis B virus (HBV) infection and more than 880000 people die due to the development of liver cirrhosis and/or hepatocellular carcinoma. Although infection can be controlled with existing treatment, eradication is currently impossible due to the persistence of covalently closed circular DNA in hepatocyte nuclei that acts as a template for viral expression. New therapeutic approaches are needed, and gene therapy has been proposed as one of the most promising options. HBV core protein [encoded by the HBV core gene (*HBC*)] is a structural protein with functional activity that has a key role in viral replication and disease progression. Accordingly, it could be a potential target for new therapeutic and diagnostic strategies, and its variability could be a valuable prognostic factor for disease progression.

Research motivation

As eradication of HBV infection is currently unachievable, new therapeutic strategies are necessary. Moreover, current treatments cannot interfere with the expression of viral proteins that can favor disease progression. Gene therapy based on silencing RNA is one of the most promising therapeutic approaches currently under investigation. The identification of hyper-conserved regions in key viral genes and proteins (such as *HBC*) is essential to orchestrate an effective strategy regardless of clinical stage or viral genotype.

Research objectives

This study aimed to identify, by next-generation sequencing, hyper-conserved regions in *HBC* quasispecies of patients with different clinical stages of chronic HBV infection that could be a valuable target for gene therapy. Considering the essential role of the *HBC* gene and its encoded protein HBV core protein in HBV infection, changes in gene and protein conservation in specific clinical groups could be determining factors in disease progression and hence serve as prognostic factors for clinical follow-up.

Research methods

The *HBC* gene was amplified by a 3-nested PCR protocol and later sequenced by next-generation sequencing (MiSeq, Illumina, United States) in 38 HBV-monoinfected chronic patients [16 with chronic hepatitis B infection without liver damages (CHB group), 5 with liver cirrhosis (LC group) and 17 with hepatocellular carcinoma (HCC group)]. Quasispecies sequences were genotyped by distance-based discriminant analysis, and general and intergroup nucleotide (nt) and amino acid (aa) conservation was determined by sliding window analysis. The presence of nt insertion and deletions and/or aa substitutions in the different groups was determined by aligning the sequences with a genotype-specific consensus sequence.

Research results

Three nt (nt 1900-1929, 2249-2284, 2364-2398) and two aa (aa 117-120, 159-167) hyper-conserved regions shared by all the clinical groups were identified. By comparing gene and protein conservation between the different clinical groups, a similar pattern of conservation was observed, although CHB showed five ntless conserved regions (nt1946-1992, 2060-2095, 2145-

2175, 2230-2250, 2270-2293) and LC one aa less conserved region (between aa 140 and 160). Moreover, some group-specific conserved regions were detected at both nt (nt 2306-2334 in CHB and 1935-1976 and 2402-2435 in LC) and aa (aa 98-103 in CHB and 28-30 and 51-54 in LC) levels. No differences in indel frequency were observed between the clinical groups. Contrarily, we identified an aa substitution (P79Q) that was more frequent in HCC [median (interquartile range) frequency of 15.82 (0-78.9) *vs* 0 (0-0) for the other groups; $P < 0.05$ *vs* the CHB group].

Research conclusions

We have identified a number of nt and aa regions that were highly conserved in the presence of different viral genotypes and clinical stages. These could be valuable targets for future pangenotypic and panclinical therapeutic and diagnostic strategies. The different clinically related conserved regions and the P79Q aa substitution could potentially be used as prognostic factors for disease progression.

Research perspectives

Our findings could guide the creation of a new gene therapy strategy based on RNA silencing. In-depth analysis of group-specific conserved or variable regions and their role in disease progression is needed. Further *in vitro* studies are required to determine whether the P79Q aa substitution might affect viral replication and to investigate associations between cell damage and onset of HCC.

ACKNOWLEDGEMENTS

The statistical and bioinformatics methods used in this study were reviewed by Mercedes Guerrero-Murillo from the Microbiology Department at Vall d'Hebron Hospital (Barcelona, Spain) and by Dr. Josep Gregori from the Liver Disease Viral Hepatitis Laboratory of Vall d'Hebron Hospital (Barcelona, Spain), CIBERehd research group, and Roche Diagnostics SL.

REFERENCES

1 **World Health Organization**. Global Hepatitis Report, 2017. World Heal Organ [Internet] 2017; 62. Available from: http://www.who.int/hepatitis
2 **Yuen MF**, Chen DS, Dusheiko GM, Janssen HLA, Lau DTY, Locarnini SA, Peters MG, Lai CL. Hepatitis B virus infection. *Nat Rev Dis Primers* 2018; **4**: 18035 [PMID: 29877316 DOI: 10.1038/nrdp.2018.35]
3 **Allweiss L**, Dandri M. The Role of cccDNA in HBV Maintenance. *Viruses* 2017; **9**: 156 [PMID: 28635668 DOI: 10.3390/v9060156]
4 **Bowden S**, Locarnini S, Chang TT, Chao YC, Han KH, Gish RG, de Man RA, Yu M, Llamoso C, Tang H. Covalently closed-circular hepatitis B virus DNA reduction with entecavir or lamivudine. *World J Gastroenterol* 2015; **21**: 4644-4651 [PMID: 25914474 DOI: 10.3748/wjg.v21.i15.4644]
5 **Papatheodoridis GV**, Chan HL, Hansen BE, Janssen HL, Lampertico P. Risk of hepatocellular carcinoma in chronic hepatitis B: assessment and modification with current antiviral therapy. *J Hepatol* 2015; **62**: 956-967 [PMID: 25595883 DOI: 10.1016/j.jhep.2015.01.002]
6 **Ren GL**, Huang GY, Zheng H, Fang Y, Ma HH, Xu MC, Zhang HB, Zhang WY, Zhao YG, Sun DY, Hu WK, Liu J. Changes in innate and permissive immune responses after hbv transgenic mouse vaccination and long-term-siRNA treatment [corrected]. *PLoS One* 2013; **8**: e57525 [PMID: 23472088 DOI: 10.1371/journal.pone.0057525]
7 **Gish RG**, Yuen MF, Chan HL, Given BD, Lai CL, Locarnini SA, Lau JY, Wooddell CI, Schluep T, Lewis DL. Synthetic RNAi triggers and their use in chronic hepatitis B therapies with curative intent. *Antiviral Res* 2015; **121**: 97-108 [PMID: 26129970 DOI: 10.1016/j.antiviral.2015.06.019]
8 **Lin YY**, Liu C, Chien WH, Wu LL, Tao Y, Wu D, Lu X, Hsieh CH, Chen PJ, Wang HY, Kao JH, Chen DS. New insights into the evolutionary rate of hepatitis B virus at different biological scales. *J Virol* 2015; **89**: 3512-3522 [PMID: 25589664 DOI: 10.1128/JVI.03131-14]
9 **González C**, Tabernero D, Cortese MF, Gregori J, Casillas R, Riveiro-Barciela M, Godoy C, Sopena S, Rando A, Yll M, Lopez-Martinez R, Quer J, Esteban R, Buti M, Rodriguez-Frias F. Detection of hyper-conserved regions in hepatitis B virus X gene potentially useful for gene therapy. *World J Gastroenterol* 2018; **24**: 2095-2107 [PMID: 29785078 DOI: 10.3748/wjg.v24.i19.2095]
10 **Locarnini S**, Zoulim F. Molecular genetics of HBV infection. *Antivir Ther* 2010; **15** Suppl 3: 3-14 [PMID: 21041899 DOI: 10.3851/IMP1619]
11 **Nassal M**, Rieger A, Steinau O. Topological analysis of the hepatitis B virus core particle by cysteine-cysteine cross-linking. *J Mol Biol* 1992; **225**: 1013-1025 [PMID: 1613786 DOI: 10.1016/0022-2836(92)90101-o]
12 **Diab A**, Foca A, Zoulim F, Durantel D, Andrisani O. The diverse functions of the hepatitis B core/capsid protein (HBc) in the viral life cycle: Implications for the development of HBc-targeting antivirals. *Antiviral Res* 2018; **149**: 211-220 [PMID: 29183719 DOI: 10.1016/j.antiviral.2017.11.015]
13 **Zlotnick A**, Venkatakrishnan B, Tan Z, Lewellyn E, Turner W, Francis S. Core protein: A pleiotropic keystone in the HBV lifecycle. *Antiviral Res* 2015; **121**: 82-93 [PMID: 26129969 DOI: 10.1016/j.antiviral.2015.06.020]
14 **Ning X**, Luckenbaugh L, Liu K, Bruss V, Sureau C, Hu J. Common and Distinct Capsid and Surface Protein Requirements for Secretion of Complete and Genome-Free Hepatitis B Virions. *J Virol* 2018; **92**: e00272-18 [PMID: 29743374 DOI: 10.1128/JVI.00272-18]
15 **Teng CF**, Huang HY, Li TC, Shyu WC, Wu HC, Lin CY, Su IJ, Jeng LB. A Next-Generation Sequencing-Based Platform for Quantitative Detection of Hepatitis B Virus Pre-S Mutants in Plasma of Hepatocellular Carcinoma Patients. *Sci Rep* 2018; **8**: 14816 [PMID: 30287845 DOI: 10.1038/s41598-018-33051-4]

16 **European Association for the Study of the Liver**. EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. *J Hepatol* 2017; **67**: 370-398 [PMID: 28427875 DOI: 10.1016/j.jhep.2017.03.021]

17 **Team RC**. A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria. 2016; Available from: https://www.r-project.org/

18 **Ramirez C**, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. *Antiviral Res* 2013; **98**: 273-283 [PMID: 23523552 DOI: 10.1016/j.antiviral.2013.03.007]

19 **Cuadras C**. A distance approach to discriminant analysis and its properties. In: Mathematics preprint series. Barcelona: 1991

20 **Cuadras C**. Distance analysis in discrimination and classification using both continuous and categorical variables. In: Statistical AData analysis and Interference. Amsterdam, 1989: 459-473

21 **Kimura M**. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980; **16**: 111-120 [PMID: 7463489 DOI: 10.1007/bf01731581]

22 **Nassal M**. HBV cccDNA: viral persistence reservoir and key obstacle for a cure of chronic hepatitis B. *Gut* 2015; **64**: 1972-1984 [PMID: 26048673 DOI: 10.1136/gutjnl-2015-309809]

23 **Soriano V**, Barreiro P, Benitez L, Peña JM, de Mendoza C. New antivirals for the treatment of chronic hepatitis B. *Expert Opin Investig Drugs* 2017; **26**: 843-851 [PMID: 28521532 DOI: 10.1080/13543784.2017.1333105]

24 **Flisiak R**, Jaroszewicz J, Łucejko M. siRNA drug development against hepatitis B virus infection. *Expert Opin Biol Ther* 2018; **18**: 609-617 [PMID: 29718723 DOI: 10.1080/14712598.2018.1472231]

25 **Durantel D**. New treatments to reach functional cure: Virological approaches. *Best Pract Res Clin Gastroenterol* 2017; **31**: 329-336 [PMID: 28774415 DOI: 10.1016/j.bpg.2017.05.002]

26 **Zhang Y**, Wu Y, Deng M, Xu D, Li X, Xu Z, Hu J, Zhang H, Liu K, Zhao Y, Gao F, Bi S, Gao GF, Zhao J, Liu WJ, Meng S. CD8$^+$ T-Cell Response-Associated Evolution of Hepatitis B Virus Core Protein and Disease Progress. *J Virol* 2018; **92**: e02120-17 [PMID: 29950410 DOI: 10.1128/JVI.02120-17]

27 **Li HC**, Huang EY, Su PY, Wu SY, Yang CC, Lin YS, Chang WC, Shih C. Nuclear export and import of human hepatitis B virus capsid protein and particles. *PLoS Pathog* 2010; **6**: e1001162 [PMID: 21060813 DOI: 10.1371/journal.ppat.1001162]

28 **Hosaka T**, Suzuki F, Kobayashi M, Fujiyama S, Kawamura Y, Sezaki H, Akuta N, Suzuki Y, Saitoh S, Arase Y, Ikeda K, Kobayashi M, Kumada H. Impact of hepatitis B core-related antigen on the incidence of hepatocellular carcinoma in patients treated with nucleos(t)ide analogues. *Aliment Pharmacol Ther* 2019; **49**: 457-471 [PMID: 30663078 DOI: 10.1111/apt.15108]

29 **Riveiro-Barciela M**, Bes M, Rodriguez-Frias F, Tabernero D, Ruiz A, Casillas R, Vidal-González J, Homs M, Nieto L, Sauleda S, Esteban R, Buti M. Serum hepatitis B core-related antigen is more accurate than hepatitis B surface antigen to identify inactive carriers, regardless of hepatitis B virus genotype. *Clin Microbiol Infect* 2017; **23**: 860-867 [PMID: 28288829 DOI: 10.1016/j.cmi.2017.03.003]

30 **Kaur H**, Bruno JG, Kumar A, Sharma TK. Aptamers in the Therapeutics and Diagnostics Pipelines. *Theranostics* 2018; **8**: 4016-4032 [PMID: 30128033 DOI: 10.7150/thno.25958]

31 **Zhang Z**, Zhang J, Pei X, Zhang Q, Lu B, Zhang X, Liu J. An aptamer targets HBV core protein and suppresses HBV replication in HepG2.2.15 cells. *Int J Mol Med* 2014; **34**: 1423-1429 [PMID: 25174447 DOI: 10.3892/ijmm.2014.1908]

32 **Orabi A**, Bieringer M, Geerlof A, Bruss V. An Aptamer against the Matrix Binding Domain on the Hepatitis B Virus Capsid Impairs Virion Formation. *J Virol* 2015; **89**: 9281-9287 [PMID: 26136564 DOI: 10.1128/JVI.00466-15]

33 **Pleshakova TO**, Kaysheva AL, Shumov ID, Ziborov VS, Bayzyanova JM, Konev VA, Uchaikin VF, Archakov AI, Ivanov YD. Detection of Hepatitis C Virus Core Protein in Serum Using Aptamer-Functionalized AFM Chips. *Micromachines (Basel)* 2019; **10**: 129 [PMID: 30781415 DOI: 10.3390/mi10020129]

34 **Percze K**, Szakács Z, Scholz É, András J, Szeitner Z, Kieboom CH, Ferwerda G, Jonge MI, Gyurcsányi RE, Mészáros T. Aptamers for respiratory syncytial virus detection. *Sci Rep* 2017; **7**: 42794 [PMID: 28220811 DOI: 10.1038/srep42794]

35 **Ludgate L**, Liu K, Luckenbaugh L, Streck N, Eng S, Voitenleitner C, Delaney WE 4th, Hu J. Cell-Free Hepatitis B Virus Capsid Assembly Dependent on the Core Protein C-Terminal Domain and Regulated by Phosphorylation. *J Virol* 2016; **90**: 5830-5844 [PMID: 27076641 DOI: 10.1128/JVI.00394-16]

36 **Liu K**, Luckenbaugh L, Ning X, Xi J, Hu J. Multiple roles of core protein linker in hepatitis B virus replication. *PLoS Pathog* 2018; **14**: e1007085 [PMID: 29782550 DOI: 10.1371/journal.ppat.1007085]

37 **Zheng CL**, Fu YM, Xu ZX, Zou Y, Deng K. Hepatitis B virus core protein dimer/dimer interface is critical for viral replication. *Mol Med Rep* 2019; **19**: 262-270 [PMID: 30387827 DOI: 10.3892/mmr.2018.9620]

38 **Ge Z**, Tian T, Meng L, Song C, Yu C, Xu X, Liu J, Dai J, Hu Z. HBV mutations in EnhII/BCP/PC region contribute to the prognosis of hepatocellular carcinoma. *Cancer Med* 2019; **8**: 3086-3093 [PMID: 31033235 DOI: 10.1002/cam4.2169]

39 **Chiu AP**, Tschida BR, Sham TT, Lo LH, Moriarity BS, Li XX, Lo RC, Hinton DE, Rowlands DK, Chan CO, Mok DKW, Largaespada DA, Warner N, Keng VW. HBx-K130M/V131I Promotes Liver Cancer in Transgenic Mice via AKT/FOXO1 Signaling Pathway and Arachidonic Acid Metabolism. *Mol Cancer Res* 2019; **17**: 1582-1593 [PMID: 30975706 DOI: 10.1158/1541-7786.MCR-18-1127]

40 **Al-Qahtani AA**, Al-Anazi MR, Nazir N, Abdo AA, Sanai FM, Al-Hamoudi WK, Alswat KA, Al-Ashgar HI, Khan MQ, Albenmousa A, El-Shamy A, Alanazi SK, Dela Cruz D, Bohol MFF, Al-Ahdal MN. The Correlation Between Hepatitis B Virus Precore/Core Mutations and the Progression of Severe Liver Disease. *Front Cell Infect Microbiol* 2018; **8**: 355 [PMID: 30406036 DOI: 10.3389/fcimb.2018.00355]

41 **Jia J**, Li H, Wang H, Chen S, Wang M, Feng H, Gao Y, Wang Y, Fang M, Gao C. Hepatitis B virus core antigen mutations predict post-operative prognosis of patients with primary hepatocellular carcinoma. *J Gen Virol* 2017; **98**: 1399-1409 [PMID: 28640739 DOI: 10.1099/jgv.0.000790]

11.2 Anexo 2: Tabla suplementaria

Número de acceso en NCBI GenBank de las 106 secuencias completas del genoma del VHB representativas de los genotipos A-J usadas en los estudios.

Genotipo VHB (n secuencias)	Número de acceso en GenBank
A (13)	AY233278, AB241115, GQ477501, AF090841, AF090839, AB194952, AB194951, AM180623, AY934764, FJ692609, FJ692613, GQ331047, GQ331048
B (20)	AB073858, AB362933, D00329, AY596111, AP011084, GQ924653, M54923, AP011085, AB073835, AB115551, AB219427, AP011086, AB287316, DQ463787, EF473977, AP011091, AP011093, AP011094, GQ358148, GQ358152
C (24)	AB112066, AB031265, X52939, AF533983, AB033553, X75656, X75665, AB048705, AB048704, AP011099, AB241109, AP011102, AP011103, EU670263, AP011107, AP011104, AP011108, AB540583, AB554019, AB554025, AB644281, AB644284, AB644286, AB644287
D (17)	AB555496, GU456647, AB104712, AF280817, AB210820, X97848, Z35716, EU939680, AY233291, AJ132335, GQ922005, AB048701, DQ315779, AB033558, AB493846, FJ904405, AM494716
E (8)	X75664, X75657, FJ349237, AM494694, HM363569, FJ349226, DQ060828, JQ000008
F (10)	AY090459, HQ378247, HE981184, AY090455, AY311369, X69798, X75663, AB036911, AB166850, DQ823090
G (6)	EF464098, HE981172, HE981176, AP007264, GU565217, AB064311
H (5)	AP007261, AB275308, AY090460, AY090454, AB516393
I (2)	FJ023660, FJ023664
J (1)	AB486012