




Viral quasispecies quantitative analysis: a novel approach for appraising the immune tolerant phase of chronic hepatitis B virus infection

Mingjie Wang ^{a,b,*}, Li Chen ^{b,*}, MinHui Dong^{c,*}, Jing Li^a, Beidi Zhu^c, Zhitao Yang^d, Qiming Gong^e, Yue Han^a, Demin Yu^a, Donghua Zhang^a, Fabien Zoulim^{f,†}, Jiming Zhang^{c,†} and Xinxin Zhang ^{a,†}

^aDepartment of Infectious Diseases, Research Laboratory of Clinical Virology, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, People's Republic of China; ^bDepartment of Gastroenterology, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, People's Republic of China; ^cDepartment of Infectious Diseases, Huashan Hospital and Key Laboratory of Medical Molecular Virology (MOH & MOE), Shanghai Medical College, Fudan University, Shanghai, People's Republic of China; ^dDepartment of Emergency, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, People's Republic of China; ^eDepartment of Infectious Diseases, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Shanghai, People's Republic of China; ^fINSERM U1052, Cancer Research Centre of Lyon (CRCL), Lyon, France

ABSTRACT

Few non-invasive models were established for precisely identifying the immune tolerant (IT) phase from chronic hepatitis B (CHB). This study aimed to develop a novel approach that combined next-generation sequencing (NGS) and machine learning algorithms using our recently published viral quasispecies (QS) analysis package. 290 HBeAg positive patients from whom liver biopsies were taken were enrolled and divided into a training group ($n = 148$) and a validation group ($n = 142$). HBV DNA was extracted and QS sequences were obtained by NGS. Hierarchical clustering analysis (HCA) and principal component analysis (PCA) based on viral operational taxonomic units (OTUs) were performed to explore the correlations among QS and clinical phenotypes. Three machine learning algorithms, including K-nearest neighbour, support vector machine, and random forest algorithm, were used to construct diagnostic models for IT phase classification. Based on histopathology, 90 IT patients and 200 CHB patients were diagnosed. HBsAg titres for IT patients were higher than those of CHB patients ($p < 0.001$). HCA and PCA analysis grouped IT and CHB patients into two distinct clusters. The relative abundance of viral OTUs differed mainly within the BCP/precore/core region and was significantly correlated with liver inflammation and fibrosis. For the IT phase classification, all machine-learning models showed higher AUC values compared to models based on HBsAg, APRI, and FIB-4. The relative abundance of viral OTUs reflects the severity of liver inflammation and fibrosis. The novel QS quantitative analysis approach could be used to diagnose IT patients more precisely and reduce the need for liver biopsy.

ARTICLE HISTORY Received 11 November 2020; Revised 6 April 2021; Accepted 14 April 2021



KEYWORDS Chronic hepatitis B; quasispecies; clinical pathology; natural history; machine learning; decision support techniques

Introduction

The natural history of chronic hepatitis B virus (HBV) infection can be divided into five phases, including the hepatitis B e antigen (HBeAg)-positive chronic infection, previously known as the immune tolerant (IT) phase (Phase I); the HBeAg-positive chronic hepatitis B (CHB) (Phase II); the HBeAg-negative chronic infection (Phase III); the HBeAg-negative CHB (Phase IV); and the hepatitis B surface antigen (HBsAg)-negative (Phase V) [1,2]. However, in a significant number of patients, it is difficult to precisely classify a specific patient into one of the above phases in the clinic, even after a complete assessment of clinical and virological profiles, including HBeAg, HBV DNA, and alanine aminotransferase (ALT) levels


[1,3,4]. Eventually, an invasive liver biopsy is required for some patients to determine the infection phase and the severity of the liver disease.

In HBV infection, due to the high variability of the HBV genome, a mass of complex and dynamically distributed variants, termed quasispecies (QS), are generated during replication and contain a remarkable amount of genomic diversity [5,6]. The QS property confers virus adaptability to the changing environment by shifting fitness under host immune or antiviral pressure. Any newly generated mutation with a selective advantage under multiple pressures posed by the innate and adaptive immune responses will take over other mutations and become the dominant QS, following the Darwinian evolutionary process [7,8].

CONTACT Xinxin Zhang  zhangx@shsmu.edu.cn  Department of Infectious Diseases, Research Laboratory of Clinical Virology, Ruijin Hospital, Shanghai Jiaotong University, School of Medicine, Ruijin 2nd road 197, Shanghai 200025, People's Republic of China

*These authors contributed equally to this work.

†These authors are co-corresponding authors.

 Supplemental data for this article can be accessed <https://doi.org/10.1080/22221751.2021.1919033>.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group, on behalf of Shanghai Shangyixun Cultural Communication Co., Ltd
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lim et al. [9] analyzed the long-term evolution of HBV during HBeAg seroconversion and found that more positive selection sites were identified within the precore/core region in HBeAg seroconverters than in non-seroconverters. Our previous study showed that more positive selection sites were detected within the surface region in patients with sustained response than in patients who experienced viral relapse after nucleotide analogue withdrawal. Most of the positive selection sites in patients with sustained response were located in HLA-I and HLA-II epitopes [10]. An enhanced host immune response and concurrent selection of mutations were associated with HBeAg seroconversion or off-treatment sustained response. A recent study showed the high genetic divergence of HBV haplotype across different phases of HBV natural history following deep sequencing of the whole HBV genome [11]. Taken together, these studies suggest that deep analysis of viral QS characteristics may constitute a means by which to study the interaction between the host immune response and viral replication.

Next-generation sequencing (NGS) enables characterization of viral variants with higher sensitivity than is possible with standard population sequencing and can detect variants at frequencies as low as 1% in the QS pool [12,13]. Indeed, NGS platforms have been implemented in many clinical and research laboratories, as the cost of these platforms is progressively decreasing [14]. NGS analysis of HBV sequences has shown that patients with detectable basal core promoter (BCP) and/or precore variants and high viral diversity achieved a lower probability of HBsAg loss during long-term tenofovir therapy [15]. NGS study also showed that a high proportion of BCP mutation was associated with the risk of cirrhosis development in HBV carriers [16]. Therefore, NGS technology provides an excellent opportunity to determine the high-risk population and to choose the optimal candidates for antiviral therapy.

Recently, we developed an automatic quasispecies analysis package (QAP) software to quantitatively analyse the massive viral quasispecies data generated from next-generation sequencing [17]. In this study, the aim was to apply the novel non-invasive approach, based on machine learning-assisted viral QS quantitative analysis, to precisely identify the IT phase in HBeAg-positive patients. Such a novel approach would help physicians to identify patients who really need antiviral therapy, reducing the clinical need for liver biopsy.

Materials and methods

Patients

HBeAg-positive CHB patients who underwent liver biopsy were enrolled retrospectively from 2008 to 2017 at Ruijin Hospital, Shanghai Jiaotong University

School of Medicine as a training group, whereas a validation group was enrolled from Huashan Hospital, Fudan University. Patients were diagnosed based on the criteria recommended by the American Association for the Study of Liver Disease (AASLD) [2]. All patients were HBsAg and HBeAg positive for >6 months. Exclusion criteria included: (1) HBV DNA levels <10⁴ IU/ml; (2) hepatitis C virus (HCV) or hepatitis D virus (HDV) co-infection; (3) previous history of antiviral therapy; (4) had received immunosuppressive therapy within the preceding 6 months; and (5) insufficient serum sample available for NGS analysis. Two hundred and ninety patients were enrolled in this study. Of them, 148 patients were in the training group and 142 patients were in the validation group. The flowchart for patient inclusion is shown in Figure 1A. Written informed consent according to the Declaration of Helsinki was obtained from each patient. This study was approved by the Ethics Committee of the Ruijin Hospital, Shanghai Jiaotong University School of Medicine (2016–17), and the Huashan Hospital, Shanghai Medical College, Fudan University (2016–124).

Clinical and laboratory tests

The serum samples were collected at the time of liver biopsy, aliquoted, and stored at –80°C. Baseline demographic variables and clinical profiles for each patient were recorded. HBV serological biomarkers (HBsAg, HBs antibody, HBeAg, HBe antibody, and hepatitis B core antibody) were measured using automated chemiluminescent microparticle immunoassays (CMIA) (Abbott, Chicago, IL, USA). HBV DNA levels were measured by real-time polymerase chain reaction (PCR) (Pj Co. Ltd., Shenzhen, China or Roche, Mannheim, Germany). Serum ALT and aspartate aminotransferase (AST) levels (upper limit of normal: 60 IU/L) were assessed with an automatic biochemical analyzer (Beckman Coulter, Brea, CA, USA or Abbott, Chicago, IL, USA).

Histopathological evaluation

Liver samples were obtained by percutaneous liver biopsy using 16-G Menghini needles, fixed in formalin, and embedded in paraffin. Hematoxylin-eosin and reticular fibre staining or Masson's staining were undertaken on each section. Section slides with less than three portal tracts were regarded as poor biopsy specimens and were excluded. Liver inflammation grading and fibrosis staging were based on a modified Scheuer scoring system by two experienced clinical pathologists [18]. The definition of the IT phase was: patients with HBsAg and HBeAg positivity, very high levels of HBV DNA (typically >1 million IU/mL), and normal or minimally elevated ALT levels

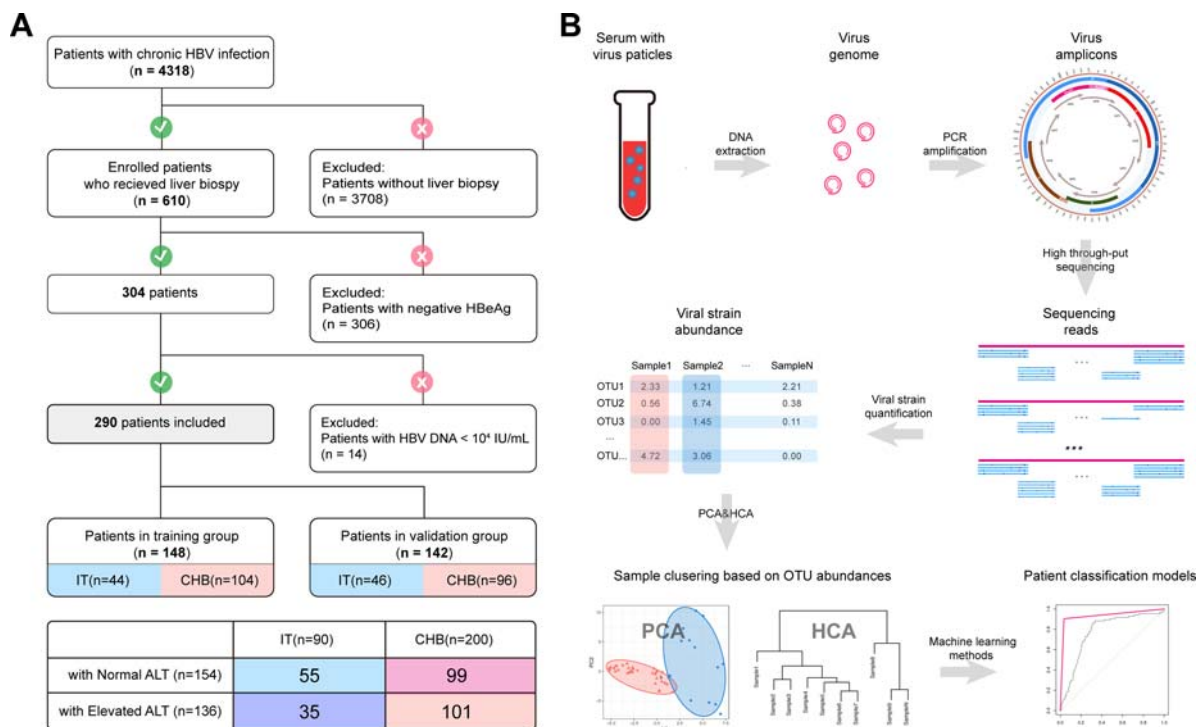


Figure 1. (A) Flowchart of patient enrolment in the study. (B) A schematic diagram of the experiment and data analysis workflow. Briefly, HBV genome DNA was extracted from serum and amplified by 9 pairs of primers, then detected by NGS which generated a pooled sequencing reads of different viral strains. QS were then quantified based on the abundances of viral OTUs, and clustered using HCA and PCA. Finally, classification models were constructed using machine learning algorithms based on sample clusters.

(less than two times of upper limit of normal) with no or mild liver inflammation and fibrosis (G0-1/S0-1) by liver histopathology. The definition of CHB phase was: patients with HBsAg and HBeAg positivity, high levels of HBV DNA ($>20,000$ IU/ml), elevated ALT levels with moderate to severe liver inflammation or fibrosis (G2-4/S2-4) by liver histopathology, according to the AASLD guideline [2].

HBV DNA extraction, amplification, and NGS process

HBV DNA was extracted from 200 μ l of serum using the QIAamp UltraSens Virus Kit (Qiagen, Hilden, Germany). The whole HBV genome was amplified for 50 randomly selected samples from the training group, using nine pairs of primers with nine overlapping fragments (Supplementary Table S1, and Figure S1, primers P1–P9). For other samples, only fragment P5 corresponding to the BCP/precore and core region was amplified. Each HBV fragment was amplified by PCR with the corresponding primers. A library of PCR products was established using a Nextera DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Each library was subjected to size selection to remove fragments <400 bp using AMPure XP beads (Beckman Coulter) and verified using an Agilent Bioanalyzer. Before sequencing, each library was quantified by real-time PCR using an NGS Library Quantification Kit (Takara, Mountain View, CA, USA). Sequencing

of the PCR products was performed using an Illumina Miseq platform, according to the manufacturer's PE 2×300 bp protocol (maximum read length 300 bp, maximum read pair span 600 bp). Image analysis and base calling were performed using Illumina CASAVA version 1.8.2 with default parameters.

Raw data pre-processing

Raw NGS data in fastq format was quality filtered by using the software QAP [17], with the following criteria: read length ≥ 250 bp and base quality ≥ 25 to remove adaptors and filter out low-quality or short reads. Totally 16,010,732 clean reads from 17,309,648 raw reads were left after filtration and used for subsequent analysis. Next, clean reads were mapped to the reference genome (genotype B, GenBank accession D00329; genotype C, GenBank accession X04615), and read pairs were assembled to amplicon sequences based on their mapping positions. Amplicon sequences were subsequently processed to correct sequencing errors, and finally, the viral haplotypes in fastq format were generated. All QS data were analyzed using QAP software [17].

QS quantification

To define a unified quantitative unit, the concept of an operational taxonomic unit (OTU) was borrowed from bacterial metagenomics analysis and redefined

here as viral strains with high homology [17]. Briefly, variant viral strains were first extracted from viral haplotypes. Next, all viral strains were clustered, and those with high homology were regarded as the same potential viral OTU. If a potential OTU met the following criteria: abundance greater than 0.1% and prevalence in the cohort greater than 2%, it would be regarded as an OTU and used for subsequent quantification. Then, OTUs were picked based on their abundance and frequency among samples. Finally, the abundance of OTUs within each QS was quantified and an OTU table was generated, in which rows corresponded to samples and columns to OTUs.

Clustering analysis

To further explore the correlations among QS and clinical phenotypes, hierarchical clustering analysis (HCA) and principal component analysis (PCA) were performed based on QS quantification. In the present study, HCA was performed in an unsupervised manner to explore potential groups in the QS of all samples. PCA converts the abundances of OTUs into a set of values of linearly uncorrelated variables known as PCs. HCA and PCA were carried out by using QAP software [17].

Diagnostic model construction using machine learning algorithms

Based on pathological classification (IT vs. CHB), three machine learning methods were applied to construct diagnostic models based on viral OTU quantification of patients in the training group, including K-nearest neighbour (KNN), support vector machine (SVM), and the random forest (RF) algorithm. These three machine-learning methods were implemented using R package KNN, e1071, and randomForest, respectively. All R packages can be downloaded from CRAN or Bioconductor. All three models were performed with five-fold cross-validation to avoid overfitting. And then the diagnostic models were validated in the validation group. The workflow of the experiment and data analysis is shown in Figure 1B.

Statistical analysis

All analysis was performed using R version 4.0.1. Continuous data are presented as mean \pm SD or median (interquartile range) and compared using Student's *t*-test or Mann-Whitney U test as appropriate. The categorical data were expressed as proportions and analyzed using the χ^2 test. Correlations were evaluated by the Spearman, Pearson, or Kendall rank correlation coefficient as appropriate. To evaluate the performance of diagnostic models, receiver operating characteristic (ROC) curve analysis was performed using R

package pROC with a 95% confidence interval (CI). Diagnostic accuracy was expressed as the sensitivity, specificity, and area under the ROC curve (AUC). A two-sided *p*-value < 0.05 was considered statistically significant for all tests.

Results

Clinical characteristics of patients

A total of 290 patients were enrolled retrospectively. Ninety and 200 patients were diagnosed as IT and CHB patients respectively, based on the clinical, virological, and histopathological profiles. Most of the patients were male, and the median age was 32.83 and 36.75 years for IT and CHB patients, respectively. HBsAg titres [(4.60 \pm 0.66) vs. (3.90 \pm 0.77) log₁₀IU/mL] and HBV DNA load [(7.56 \pm 0.52) vs. (6.59 \pm 1.29) log₁₀IU/mL] in IT patients were higher than those in CHB patients (*p* < 0.001) (Table 1). Interestingly, thirty-five patients (38.8%, 35/90) in the IT group had slightly elevated ALT levels [76.00(64.50–93.00) IU/L]; while 99 patients (49.5%, 99/200) in the CHB group had normal ALT levels. The age, ALT, HBsAg, and HBV DNA levels were different among the subgroups (*p* < 0.05) (Table S2). The clinical characteristics of the training group and validation group were shown in Supplementary Table S3.

The relative abundance of viral OTUs between IT and CHB patients differed mainly within the BCP/precore/core region

To identify the specific regions of sequence divergence between viral strains of IT and CHB patients, 50 patients were randomly selected from the training group, and the whole HBV genome in these patients was sequenced with nine pairs of overlapping primers. PCA was performed based on viral OTU quantification of nine amplicons (Figure 2). Among the nine amplicons, sample clustering based on viral OTUs in amplicon P5 (Figure 2E) showed two distinct clusters corresponding to the IT and CHB patients. Associations between PC1 and the sample groups were calculated, and amplicon P5 had the most statistically significant association (*p* = 1.11E–10, supplementary Table S4). Thus, amplicon P5 was regarded as the most divergent region between IT and CHB patients. As amplicon P5 corresponds to the BCP/precore and core region in the HBV genome, the results indicate that most of the sequence divergence between IT and CHB patients occurred within this region. The frequency of hot-spot mutation within the BCP/precore and core regions, such as A1762T/G1764A and G1896A mutation, was higher in the CHB group than in the IT group (Table S5).

Table 1. The clinical characteristics of the study patients in the IT and CHB group.

| | IT group (n = 90) | CHB group (n = 200) | p value |
|-----------------------------------|---------------------|----------------------|---------|
| Sex (Male, %) | 80 (88.89) | 152 (76.00) | 0.02 |
| Age (years) | 32.83 ± 9.18 | 36.75 ± 10.68 | <0.01 |
| ALT (IU/L) | 47.00 (29.50–70.50) | 61.00 (40.50–110.50) | <0.01 |
| AST (IU/L) | 29.50 (23.00–39.25) | 40.50 (30.25–63.75) | <0.01 |
| PLT (10 ⁹ /L) | 202.96 ± 50.99 | 176.50 ± 54.94 | <0.01 |
| HBV DNA (log ₁₀ IU/ml) | 7.56 ± 0.52 | 6.59 ± 1.29 | <0.01 |
| HBsAg (log ₁₀ IU/ml) | 4.60 ± 0.66 | 3.90 ± 0.77 | <0.01 |
| Genotype B/C (n) | 38/52 | 78/122 | 0.75 |
| G0/G1/G2/G3/G4(n) | 30/60/0/0/0 | 4/26/112/47/11 | <0.01 |
| S0/S1/S2/S3/S4(n) | 55/35/0/0/0 | 2/48/86/31/33 | <0.01 |

PCA analysis demonstrated clusters overlapped slightly between IT and CHB patients

PCA based on viral OTU quantification within amplicon P5 was analyzed for all training group patients. The abundance of viral OTUs in IT and CHB patients was significantly different with distinct clusters in the two groups. PCA of viral OTUs revealed that IT and CHB patients were grouped into two distinct clusters. There was a small overlap between the two clusters in both HCA and PCA, indicating a potential transitional phase between IT and CHB phases (Figure 3A and B).

The relative abundance of viral OTUs was significantly correlated with liver inflammation and fibrosis

To evaluate the associations among commonly used clinical markers and viral QS quantification, the first two PCs were extracted and their correlations with clinical markers were evaluated using correlation coefficients. The first two PCs were more correlated with liver inflammation and fibrosis than with HBsAg, HBV DNA, and PLT levels (Table 2).

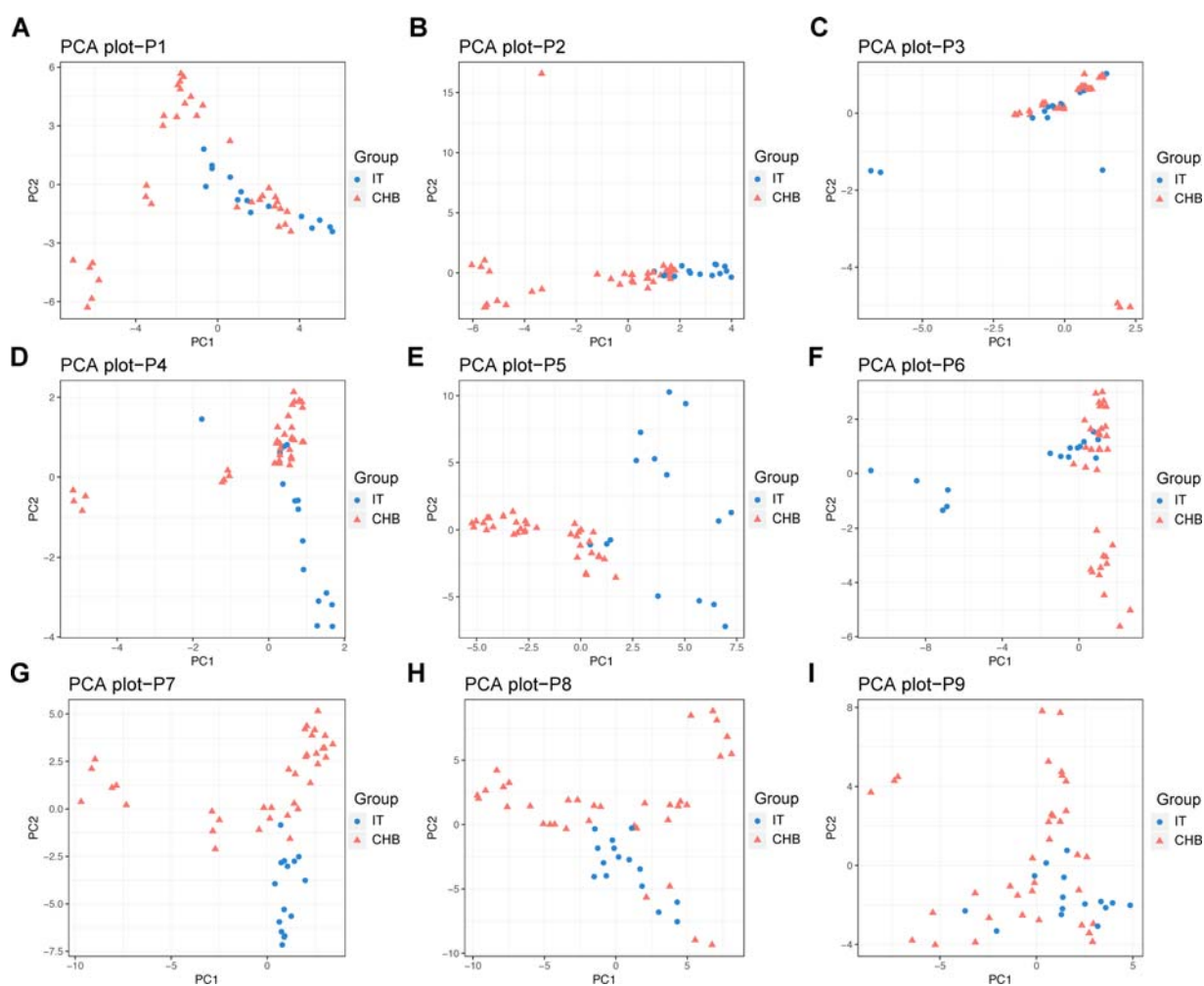


Figure 2. Scatter plots of PCA results of 9 amplicons in the training group. (A~I) corresponds to amplicon P1 to P9 (amplified by primers P1 to P9, respectively). Each dot in the plot represents a sample, of which dots in the red represent CHB patients and dots in the blue represent IT patients. The x-axis and y-axis represent the top 2 principal components (PC), PC1 and PC2, respectively.

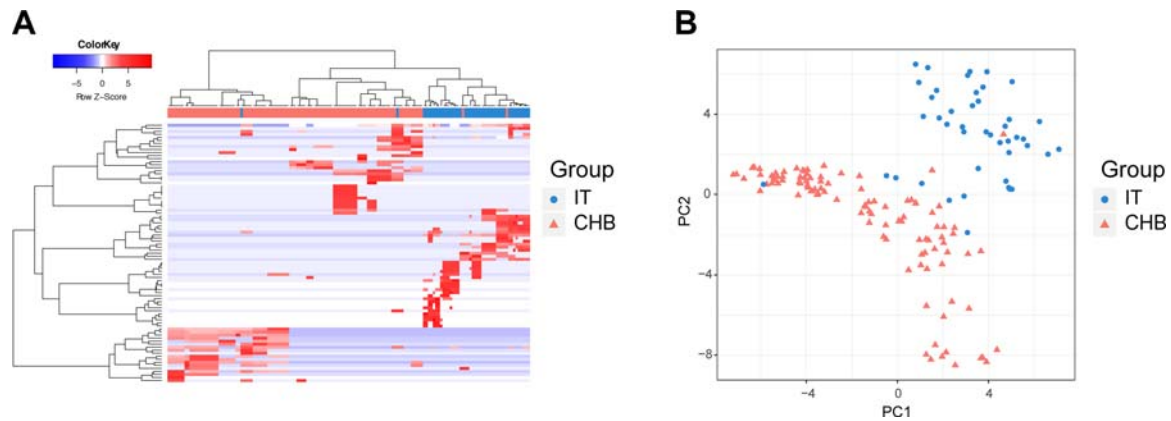


Figure 3. HCA and PCA based on viral OTUs of amplicon P5 in the training group. (A) Hierarchical clustering heatmap of viral OTUs. A column corresponds to viral OTUs within a patient, and a row corresponds to the relative abundance of a representative OTU in all patients. The colours corresponding to the scales bars and traits are shown on the left. (B) Principal component analysis of viral OTUs in 148 patients in the training group. PC1 and PC2 were used as x-axis and y-axis in two dimensions, respectively. Each dot represents one sample, and the colours indicate different groups.

Machine-learning models had superior diagnostic performance

Firstly, the diagnostic models were constructed to precisely identify the IT or CHB phase of HBV infection based on viral QS quantification using machine learning methods.

The relative abundance of viral OTUs within QS was quantified according to haplotype counts in the viral spectra. RF, SVM, and KNN algorithms were used to construct predictive models without involving any clinical phenotypic parameters. The models were then validated in the validation group. The sensitivity, specificity, and classification accuracy of each model were compared with quantitative HBsAg, ALT level, and widely used liver fibrosis models APRI and FIB-4 [19,20]. All models showed significantly higher specificity, sensitivity, accuracy, and AUC values than HBsAg, ALT, or APRI and FIB-4 in the classification of clinical phenotypes, both in the training and in the validation group (Table 3 and Figure 4).

Secondly, the diagnostic models were constructed to evaluate the severity of liver histopathology in patients with normal and elevated ALT levels, respectively. Three machine learning models showed significantly higher specificity, sensitivity, accuracy, and AUC values than HBsAg and APRI and FIB-4 in identifying

the IT or CHB patients, either in the patients with normal ALT or elevated ALT levels (Tables S6, S7).

Discussion

We developed a novel non-invasive approach to precisely identify the IT phase in HBV patients using machine learning-assisted analysis of NGS-generated viral QS sequences in a very well-defined patient cohort based on a set of clinical, biological, virologic, and histopathological findings. PCA of viral OTUs revealed that IT and CHB patients were grouped into two distinct clusters. PCA analysis also clustered different levels of liver inflammation and fibrosis, either in the normal ALT subgroup or in the elevated ALT subgroup. Predictive models that analyze the relative abundance of viral OTUs, based on machine learning algorithms, can accurately identify the HBV infection phase with significant accuracy.

The complicated interaction between the host immune response and HBV results in different clinical outcomes [21]. It is commonly believed that immunological events play important roles in the shift from the IT phase to the immune clearance phase. The innate and adaptive immune responses place multiple selective constraints on viral replication. Under such circumstances, viruses tend to increase mutations and gain replicative fitness, maintaining persistent infection [7]. A consequence of viral QS adaptation to the changing environment is rapid generation of phenotypic diversity and escape mutations [6]. The BCP/pre-core region, which does not overlap with other ORFs, exhibits wide sequence variation and immunogenicity [22]. Deep sequencing of mother-to-child transmission CHB patients has shown that the BCP and precore sequences are highly conserved during the IT phase in contrast to frequent mutation during the immune clearance phase [23]. The lesser mutation

Table 2. Statistical significance of associations between PCs and clinical profiles.

| Phenotype | PC1 | | PC2 | |
|-----------------------------------|----------------|-------------|----------------|-------------|
| | <i>p</i> value | Correlation | <i>p</i> value | Correlation |
| Gender (Male/Female) | 3.40E-02 | 0.07 | 2.74E-01 | 0.02 |
| Age (years) | 1.06E-01 | 0.10 | 6.32E-01 | 0.03 |
| ALT (IU/L) | 5.67E-01 | 0.03 | 2.97E-04 | 0.21 |
| AST (IU/L) | 2.20E-02 | 0.14 | 1.38E-05 | 0.26 |
| PLT(10 ⁹ /L) | 4.86E-02 | -0.12 | 1.94E-03 | -0.19 |
| HBV DNA (log ₁₀ IU/mL) | 4.25E-07 | -0.30 | 1.96E-01 | -0.08 |
| HBsAg (log ₁₀ IU/mL) | 1.81E-08 | -0.33 | 1.53E-05 | -0.26 |
| Inflammation grade (G) | 2.98E-10 | 0.42 | 8.19E-06 | 0.38 |
| Fibrosis stage(S) | 6.47E-10 | 0.43 | 2.01E-05 | 0.34 |

Table 3. Comparison of the performance between diagnostic models and clinical parameters in identifying the IT or CHB patients.

| | | SVM | KNN | RF | HBsAg | ALT | FIB-4 | APRI |
|------------------|------|--------|--------|--------|--------|--------|--------|--------|
| Training group | SPEC | 0.9689 | 0.9689 | 0.9678 | 0.8218 | 0.6602 | 0.4353 | 0.7326 |
| | SENS | 1.0000 | 0.9598 | 0.9670 | 0.9211 | 0.5854 | 0.7857 | 0.6552 |
| | ACC | 0.9817 | 0.9657 | 0.9675 | 0.8489 | 0.6389 | 0.5221 | 0.7130 |
| | AUC | 0.9845 | 0.9652 | 0.9681 | 0.8876 | 0.6153 | 0.5576 | 0.7153 |
| Validation group | SPEC | 0.9349 | 0.9411 | 0.9446 | 0.6744 | 0.4255 | 0.6484 | 0.6813 |
| | SENS | 0.8327 | 0.8192 | 0.8141 | 0.7447 | 0.8958 | 0.6905 | 0.7143 |
| | ACC | 0.9031 | 0.9032 | 0.9040 | 0.6992 | 0.5845 | 0.6617 | 0.6917 |
| | AUC | 0.8838 | 0.8801 | 0.8793 | 0.6759 | 0.6806 | 0.7033 | 0.7276 |

SPEC, specificity; SENS, sensitivity; ACC, accuracy; AUC, area under ROC curve; SVM, support vector machine; KNN, K-nearest neighbour; RF, random forest; FIB-4, fibrosis-4 index; APRI, AST-to-platelet ratio index.

rate during the IT phase may reflect the lack of significant selective pressure on the virus. Similar to a recent study that HBV haplotype generally separated, particularly within precore/core gene, according to different phases of HBV natural history [11], the PCA analysis based on precore/core gene can identify the IT and CHB patients precisely. Recent studies have shown that HBV-specific T-cell responses may not be as exhausted as previously thought and that HBV DNA integration events may occur during the IT phase [24,25], which challenged the prevailing opinion that the immune response and virus mutation is quiescent in such patients. The present study showed a slight overlapping area between IT and CHB patients in cluster analysis and a minimally elevated ALT level in some IT patients. Moreover, a recent study

on HBV variants by NGS from treatment naïve IT patients revealed that the level of viral diversity was correlated with age and ALT level and negatively correlated with HBVDNA, HBsAg and HBeAg level, indicating immune tolerant phase transition towards immune clearance phase despite of normal ALT level [26]. These observations further supported the concept that antiviral immune responses and virus evolution may occur in certain IT patients.

Further, increased viral mutations, due to enhanced immune pressure during the CHB phase, are associated with liver inflammation and fibrosis. The frequency of the I27 V mutation in the HBV core gene is higher in severe compared to moderate liver inflammation [27]. CHB patients who carry the A1762T/G1764A mutation have a higher risk of severe fibrosis

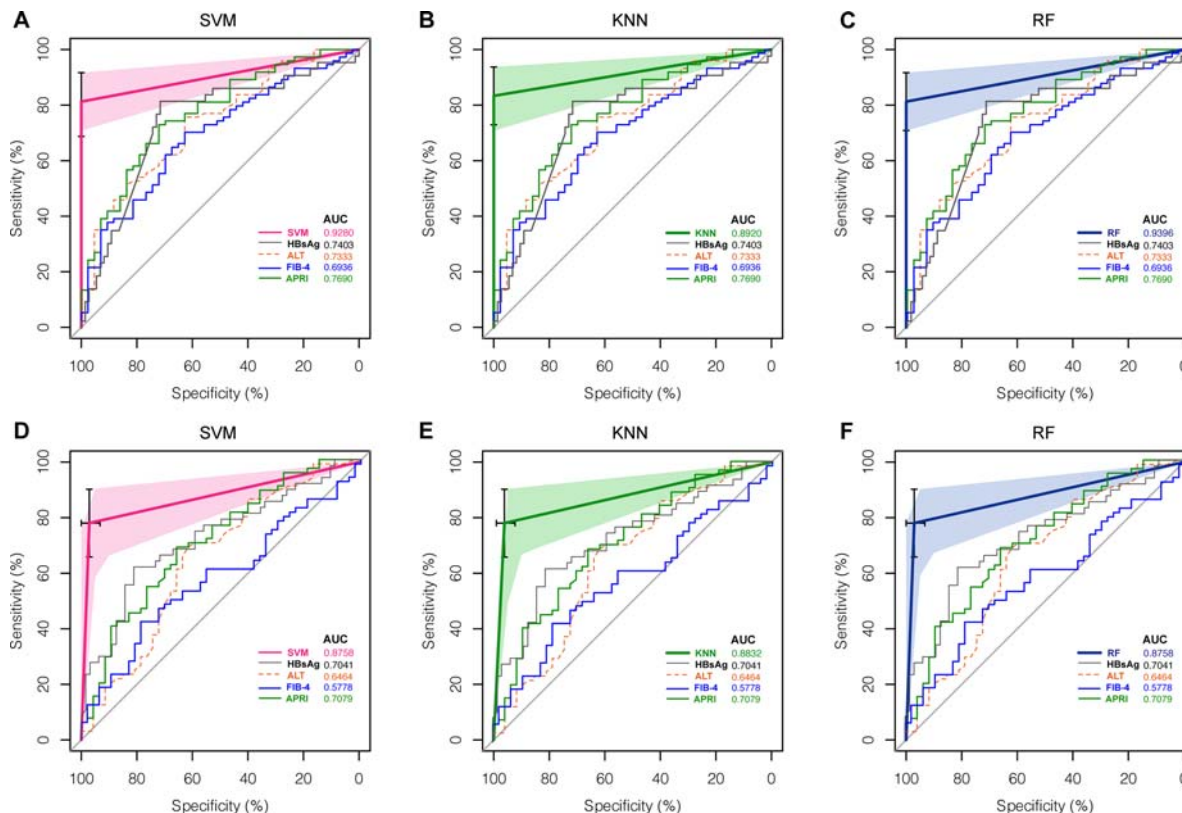


Figure 4. ROC curves of three diagnostic models using machine learning methods compared with HBsAg and ALT level in identifying IT and CHB patients. (A~C) ROC curves of three diagnostic models constructed using SVM, KNN and RF methods in the training group. (D~F) ROC curves of three diagnostic models constructed by using SVM, KNN and RF methods in the validation group. The coloured ribbon corresponds to the 95% CI of ROC curves. SVM, support vector machine; KNN, K-nearest neighbour; RF, random forest; FIB-4, fibrosis-4 index; APRI, AST-to-platelet ratio index.

[28]. NGS in a longitudinal cohort revealed that the BCP mutation is significantly associated with cirrhosis development [16]. Massively parallel pyrosequencing of HBV quaspecies showed that the frequency of viral substitution within BCP/pre-core/core region, including A1762T/G1764A/G1896A and other two novel mutations, was significantly associated with the advanced liver disease compared to chronic HBV carrier [29]. In our study, PCA analysis based on P5 amplicons, which corresponding to the BCP/pre-core/core region, significantly grouped the IT and CHB patients into two distinct clusters, and the first two PCs were significantly correlated with liver inflammation and fibrosis. Our data demonstrated for the first time that liver inflammation and fibrosis can be inferred using deep analysis of viral QS spectra.

Despite an encouraging randomized clinical trial with IT children [30], current clinical practice guidelines recommend against antiviral therapy for adults with the IT phase [1,2]. Precise identification of the IT or CHB phase is crucial for physicians when considering the initiation of antiviral therapy. In the clinic, the classification of the different phases is mainly based on clinical, biochemical, and virological profiles and non-invasive fibrosis tests. However, 37~40.2% of HBeAg positive CHB patients with persistently normal ALT levels had significant liver fibrosis or inflammation [31,32]. Lower HBsAg quantification and HBV DNA load was correlated with moderate or severe fibrosis in HBeAg positive CHB patients, with an AUC of 0.77 [33]. The non-invasive method for measuring liver stiffness using Fibroscan identified patients with \geq F2 fibrosis, with a sensitivity of 58~82% and specificity of 75~79% [34]. The widely used non-invasive panels, including FIB-4 and APRI, are more suitable for the evaluation of significant or advanced fibrosis in normal ALT patients, but with low PPV. However, previous biomarker or panels failed to precisely identify the IT phase, which was characterized by no or minimal liver necro-inflammation or fibrosis. In the present study, the predictive models using machine learning algorithms had high accuracy in identifying the IT phase patients and determining the severity of liver histopathology, compared with HBsAg quantification, ALT level, and FIB-4 and APRI. Specifically, the clustering analysis and predictive models precisely determined the subgroup of CHB patients with normal ALT levels. Such patients would have been excluded from antiviral therapy if a liver biopsy was not performed. Further, this novel approach precisely identified a subgroup of IT patients with no or mild liver inflammation and fibrosis but exhibited mildly elevated ALT levels. The ALT level can be minimally elevated in the IT patients according to the AASLD guideline [2] in contrast to the EASL guideline [1]. Actually, a substantial proportion of patients with

slightly elevated ALT levels had near-normal liver histopathology [35]. Some HBeAg positive CHB patients (52~71%) with ALT levels less than twice the upper limit of normal (60 IU/L) do have mild liver inflammation or fibrosis [36,37]. Antiviral therapy may not be urgent but close surveillance may be a better strategy for such patients. Other causes for elevated ALT level should be investigated, including but not limited to non-alcoholic fatty liver disease, alcoholic, and autoimmune liver disease. Our novel approach would provide clinicians with a means by which to precisely identify the patients who really need antiviral therapy, reducing the need for liver biopsy.

The traditional method used to analyze viral QS is cloning and sequencing. This is a costly, labor-intensive process that requires multiple, complex, experimental steps, with limited resolution [38]. NGS approaches enable high-throughput analysis of thousands of sequences and are a powerful tool for the characterization of genetic diversity in viral strains [38]. Deep sequencing analysis of the HBV genome can increase our understanding of HBV diversity and evolution, host immune responses, resistance to treatment, and disparities in clinical outcomes [39].

Studies on HBV QS have focused on the global description of viral spectra, including mutation frequency and QS complexity and diversity [40–42]. Due to the lack of a uniform quantification unit, it is difficult to quantitatively compare the relative abundance among different samples. In this study, we introduced the concept of OTU, which is widely used as a unit of microbial diversity [43], for global and local quantitative analysis of viral QS. Each OTU represents a cluster of nucleotide sequences that are highly similar and likely to represent one QS. The assumption is that sequences with a high degree of nucleotide identity belong to the same OTU [44]. Moreover, specific variants within QS can be recognized and extracted when the OTUs are matched with clinical and viral profiles.

Machine learning methods have been applied to a variety of problems in genomics and genetics [45]. Compared with previous diagnostic tools and traditional QS analysis, the novel machine learning-assisted QS analysis has some advantages. First, high-throughput NGS technology provides a more detailed description of the viral QS spectra. The predictive models have a high degree of classification accuracy (>90%) in classifying the IT and CHB phases, no matter which machine learning algorithm was applied. Second, this novel approach combines big data and machine learning algorithms in a high-quality computer processing platform, significantly overcoming the disadvantages of traditional QS analysis methods that are labor-intensive and time-consuming. Third, an unsupervised learning method [45] is applied so that the input OTUs can be trained without

any other clinical or viral information. The output results show the clinical phenotype automatically. Finally, the novel approach is non-invasive and practical for clinical application.

There are limitations to this approach. First, to understand and perform the procedure, a bioinformatics background is required. To overcome this necessity, we have developed software (QAP) with a visual interface that can process the NGS data automatically [17]. Second, the raw NGS data must be filtered to subtract background information to improve sequence quality before QS analysis.

In summary, we have developed a novel non-invasive diagnostic approach, based on machine learning-assisted viral QS analysis, to precisely identify IT and CHB patients and to determine the severity of liver histopathology. A difference in the relative abundance of viral OTUs reflects the severity of liver inflammation and fibrosis. The primary results demonstrate high sensitivity, specificity, and diagnostic efficiency when machine learning algorithms were applied. These results demonstrate the slight overlap between IT and CHB patients and imply an interaction between HBV and the host immune response, deepening our understanding of the natural history of HBV.

List of abbreviations

AASLD, American Association for the Study of Liver Disease; ALT, alanine aminotransferase; APRI, AST to platelet ratio index; AST, aspartate aminotransferase; AUC, area under ROC curve; BCP, basal core promoter; CHB, chronic hepatitis B; CI, confidence interval; EASL, European Association for the Study of the Liver; FIB-4, fibrosis-4 score; HBV, hepatitis B virus; HCA, hierarchical clustering analysis; HLA, human leukocyte antigen; IT, immune tolerant; KNN, K-nearest neighbour; NGS, next-generation sequencing; ORFs, overlapping reading frames; OTU, operational taxonomic unit; PCA, principal component analysis; PCs, principal components; PCR, polymerase chain reaction; QS, quasispecies; RF, random forest; ROC, receiver-operating characteristic; SVM, support vector machine.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was sponsored by grants from the National Natural Science Foundation of China [grant numbers 81371860, 81672069, 81770590], Shanghai Municipal Committee of Science and Technology [grant number 16410711900], Shanghai Shen Kang Hospital Developing Centre [grant number SHDC12016101], the Major Science and

Technology Special Project of China [grant numbers 2017ZX10202202, 2018ZX10302204-001-003].

ORCID

Mingjie Wang  <http://orcid.org/0000-0002-5680-2390>

Li Chen  <http://orcid.org/0000-0001-8492-1963>

Xinxin Zhang  <http://orcid.org/0000-0002-0598-6425>

References

- [1] European Association for the Study of the Liver, Electronic address: easloffice@easloffice.eu, 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(2):370–398.
- [2] Terrault NA, Lok ASF, McMahon BJ, et al. Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. *Hepatology*. 2018;67(4):1560–1599.
- [3] Sarin SK, Kumar M, Lau GK, et al. Asian-Pacific clinical practice guidelines on the management of hepatitis B: a 2015 update. *Hepatol Int*. 2016;10(1):1–98.
- [4] Terrault NA, Bzowej NH, Chang K-M, et al. AASLD guidelines for treatment of chronic hepatitis B. *Hepatology*. 2016;63(1):261–283.
- [5] Domingo E, Martín V, Perales C, et al. Viruses as quasispecies: biological implications. *Curr Top Microbiol Immunol*. 2006;299:51–82.
- [6] Domingo E, Gomez J. Quasispecies and its impact on viral hepatitis. *Virus Res*. 2007;127(2):131–150.
- [7] Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev*. 2012;76(2):159–216.
- [8] Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog*. 2010;6(7):e1001005.
- [9] Lim SG, Cheng Y, Guindon S, et al. Viral quasi-species evolution during hepatitis Be antigen seroconversion. *Gastroenterology*. 2007;133(3):951–958.
- [10] Chen L, Gan QR, Zhang DQ, et al. Increased intrahepatic quasispecies heterogeneity correlates with off-treatment sustained response to nucleos(t)ide analogues in e antigen-positive chronic hepatitis B patients. *Clin Microbiol Infect*. 2016;22(2):201–207.
- [11] Wagner J, Yuen L, Littlejohn M, et al. Analysis of Hepatitis B virus haplotype diversity detects striking sequence conservation across genotypes and chronic disease phase. *Hepatology*. 2020. doi:10.1002/hep.31516. PMID: 32780526.
- [12] Ayres A, Yuen L, Jackson KM, et al. Short duration of lamivudine for the prevention of hepatitis B virus transmission in pregnancy: lack of potency and selection of resistance mutations. *J Viral Hepat*. 2014;21(11):809–817.
- [13] Bayliss J, Nguyen T, Lesmana C, et al. Advances in the molecular diagnosis of hepatitis B infection: providing insight into the next generation of disease. *Semin Liver Dis*. 2013;33(2):113–121.
- [14] Barzon L, Lavezzo E, Costanzi G, et al. Next-generation sequencing technologies in diagnostic virology. *J Clin Virol*. 2013;58(2):346–350.
- [15] Bayliss J, Yuen L, Rosenberg G, et al. Deep sequencing shows that HBV basal core promoter and precore variants reduce the likelihood of HBsAg loss following

- tenofovir disoproxil fumarate therapy in HBeAg-positive chronic hepatitis B. *Gut*. 2017;66(11):2013–2023.
- [16] Tseng TC, Liu C-J, Yang H-C, et al. Higher proportion of viral basal core promoter mutant increases the risk of liver cirrhosis in hepatitis B carriers. *Gut*. 2015;64(2):292–302.
- [17] Wang M, Li J, Zhang X, et al. An integrated software for virus community sequencing data analysis. *BMC Genomics*. 2020;21(1):363.
- [18] Desmet VJ, Gerber M, Hoofnagle JH, et al. Classification of chronic hepatitis: diagnosis, grading and staging. *Hepatology*. 1994;19(6):1513–1520.
- [19] Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43(6):1317–1325.
- [20] Wai CT, Greenson JK, Fontana RJ, et al. A simple non-invasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. *Hepatology*. 2003;38(2):518–526.
- [21] Chisari FV, Isogawa M, Wieland SF. Pathogenesis of hepatitis B virus infection. *Pathologie Biologie*. 2010;58(4):258–266.
- [22] Vanlandschoot P, Cao T, Leroux-Roels G. The nucleocapsid of the hepatitis B virus: a remarkable immunogenic structure. *Antiviral Res*. 2003;60(2):67–74.
- [23] Sede M, Lopez-Ledesma M, Frider B, et al. Hepatitis B virus depicts a high degree of conservation during the immune-tolerant phase in familiarly transmitted chronic hepatitis B infection: deep-sequencing and phylogenetic analysis. *J Viral Hepat*. 2014;21(9):650–661.
- [24] Kennedy PTF, Sandalova E, Jo J, et al. Preserved T-cell function in children and young adults with immune-tolerant chronic hepatitis B. *Gastroenterology*. 2012;143(3):637–645.
- [25] Mason WS, Gill US, Litwin S, et al. HBV DNA integration and clonal hepatocyte expansion in chronic Hepatitis B patients considered immune tolerant. *Gastroenterology*. 2016;151(5):986–998.e4.
- [26] Yuen L, Revill PA, Rosenberg G, et al. HBV variants are common in the ‘immune-tolerant’ phase of chronic hepatitis B. *J Viral Hepat*. 2020;27(10):1061–1070.
- [27] Yang L, Ma S, Hu X, et al. Presence of valine at position 27 of the hepatitis B virus core gene is associated with severe liver inflammation in Chinese patients. *J Med Virol*. 2011;83(2):218–224.
- [28] Ducancelle A, Pivert A, Bertrais S, et al. Different pre-core/core mutations of hepatitis B interact with, limit, or favor liver fibrosis severity. *J Gastroenterol Hepatol*. 2016;31(10):1750–1756.
- [29] Li F, Zhang D, Li Y, et al. Whole genome characterization of hepatitis B virus quasispecies with massively parallel pyrosequencing. *Clin Microbiol Infect*. 2015;21(3):280–287.
- [30] Zhu S, Zhang H, Dong Y, et al. Antiviral therapy in hepatitis B virus-infected children with immune-tolerant characteristics: a pilot open-label randomized study. *Journal of Hepatology*. 2018;68(6):1123–1128.
- [31] Kumar M, Sarin SK, Hissar S, et al. Virologic and histologic features of chronic hepatitis B virus-infected asymptomatic patients with persistently normal ALT. *Gastroenterology*. 2008;134(5):1376–1384.
- [32] Lai M, Hyatt BJ, Nasser I, et al. The clinical significance of persistently normal ALT in chronic hepatitis B infection. *J Hepatol*. 2007;47(6):760–767.
- [33] Martinot-Peignoux M, Carvalho-Filho R, Lapalus M, et al. Hepatitis B surface antigen serum level is associated with fibrosis severity in treatment-naïve, e antigen-positive patients. *J Hepatol*. 2013;58(6):1089–1095.
- [34] Afdhal NH, Bacon BR, Patel K, et al. Accuracy of fibroscan, compared with histology, in analysis of liver fibrosis in patients with hepatitis B or C: a United States multicenter study. *Clin Gastroenterol Hepatol*. 2015;13(4):772–779.e1-3.
- [35] Ettel MG, Appelman HD. Approach to the liver biopsy in the patient with chronic low-level aminotransferase elevations. *Arch Pathol Lab Med*. 2018;142(10):1186–1190.
- [36] Gao S, Li X-Y, Fan Y-C, et al. A noninvasive model to predict liver histology in HBeAg-positive chronic hepatitis B with alanine aminotransferase ≤ 2 upper limit of normal. *J Gastroenterol Hepatol*. 2017;32(1):215–220.
- [37] Chen J, Xu CR, Xi M, et al. Predictors of liver histological changes and a sustained virological response to peginterferon among chronic hepatitis B e antigen-positive patients with normal or minimally elevated alanine aminotransferase levels. *J Viral Hepat*. 2017;24(7):573–579.
- [38] Rodriguez-Frias F, Buti M, Taberner D, et al. Quasispecies structure, cornerstone of hepatitis B virus infection: mass sequencing approach. *World J Gastroenterol*. 2013;19(41):6995–7023.
- [39] McNaughton AL, D’Arienzo V, Ansari MA, et al. Insights from deep sequencing of the HBV genome—unique, tiny, and misunderstood. *Gastroenterology*. 2019;156(2):384–399.
- [40] Liu F, Chen L, Yu D-M, et al. Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: correlation with antiviral efficacy. *Gut*. 2011;60(9):1269–1277.
- [41] Chen L, Zhang Q, Yu D-m, et al. Early changes of hepatitis B virus quasispecies during lamivudine treatment and the correlation with antiviral efficacy. *J Hepatol*. 2009;50(5):895–905.
- [42] Gregori J, Perales C, Rodriguez-Frias F, et al. Viral quasispecies complexity measures. *Virology*. 2016;493:227–237.
- [43] Sarangi AN, Goel A, Aggarwal R. Methods for studying gut microbiota: a primer for physicians. *J Clin Exp Hepatol*. 2019;9(1):62–73.
- [44] Porter TM, Hajibabaei M. Scaling up: a guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol*. 2018;27(2):313–338.
- [45] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321–332.