Journal of Proteomics & Bioinformatics
ISSN: 0974-276X

Lian et al., J Proteomics Bioinform 2013, S6
DOI: 10.4172/jpb.S6-005

Research Article Open Access

Detecting Hepatitis B Viral Amino Acid Sequence Mutations in Occult Hepatitis B Infections *via* Bayesian Partition Model

Zhichao Lian[2], Qi Ning Tian[2], Yang Liu[2], Valeria Cento[3], Romina Salpini[3], Carlo Federico Perno[3], Valentina Svicher[3], Gang Chen[2], Cong Li[1] and Jing Zhang[1,2]*

[1]*Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT06511, USA*
[2]*Department of Statistics, Yale University, New Haven, CT06511, USA*
[3]*Department of Experimental Medicine and Surgery, University of Rome Tor Vergata, Rome, Italy*

Abstract

Background: With advancements in technology, a number of Hepatitis B virus (HBV) infections, where viral DNA is present in the liver or plasma, without the concomitant detection of HBsAg in plasma have been reported, and have been termed occult Hepatitis B infections (OBI). Unfortunately, the etiology and pathogenesis of OBI remain elusive to date, and the genetic characteristics of HBV sequence that lead to the development of OBI are still poorly understood.

Methods: 358 genotype-C (330 chronically infected patients and 28 occult infected patients) and 107 genotype-D (83 chronically infected patients and 24 occult infected patients) HBV Reverse Transcriptase (RT) amino acid sequences were collected. In addition to greedy search, a novel statistical approach, Bayesian Variable Partition Model is applied to pinpoint those positions, where amino acid mutations collaboratively discriminate OBI samples from chronically infected samples, in genotype-C and genotype-D, respectively.

Results: Several discriminate and correlated positions were found in genotype-C (high-order position combinations listed in tables) and genotype-D (positions 126+138, 129+131 and 138+139) respectively. By comparing amino acid distributions in these positions between genotype-C and genotype-D, six position combinations were reported to have obvious different amino acid distributions in these two HBV genotypes.

Conclusions: This paper furthers the understanding of the correlation between HBV sequence mutations and the differences of OBI in two HBV genotypes, by studying mutations in HBV RT amino acid sequences. Different from other traditional methods, the Bayesian-based method is able to analyze high-order combinations of positions.

Keywords: Hepatitis B virus; Baysian method; Occult HBV

Introduction

Each year, Hepatitis B infects 10 to 30 million people worldwide and claims an estimated 600,000 lives [1,2]. While most patients are able to clear the infection within 6 months, roughly 5%-10% of infected adults develop a chronic form of the infection [1-3]. The rates of developing a chronic Hepatitis B infection are significantly more elevated in younger populations, sitting at 30%-50% for patients between the ages of one and four, and a staggering 90% for infants [2]. One third of those who contract the chronic form of the disease will exhibit clinical symptoms, which can cause irreversible liver damage and lead to cirrhosis, as well as hepatocellular carcinoma [3]. The remaining two thirds of chronic Hepatitis B patients, though asymptomatic, retain the Hepatitis B virus in the body, and can remain highly infectious [3]. Overall, 15% to 25% of patients suffering from chronic Hepatitis B infection die from complications resulting from the disease [3].

The Hepatitis B virus (HBV) comprises an icosahedral protein capsid surrounding the viral DNA, with a lipoprotein viral envelope [4-7]. HBV DNA is divided into four open reading frames (ORF): ORF S, which encodes HBsAg, ORF C, which encodes HBcAg and HBeAg, ORF X, which encodes an X protein, whose precise function is currently unclear, and ORF P, which encodes DNA polymerase [5,7]. The nucleocapsid consists of two types of highly immunogenic proteins: the core antigen, named HBcAg, and a truncated variant termed the E antigen (HBeAg), while the viral envelope holds the less immunogenic surface antigen abbreviated HBsAg [5-7]. In most cases of HBV infections, HBsAg is among the first serological markers to become detectable in the blood, with an average onset of four weeks post-exposure; as such, immunological assays for the antigen, as well as for its antibody, hold an importance place in the diagnosis of Hepatitis B [8,9]. Indeed, a Hepatitis B blood panel tests for three markers: HBsAg, antibodies against HBsAg (anti-HBs), and antibodies against HBcAg (anti-HBc) [10].

With advancements in technology, particularly in the sensitivity of polymerase chain reactions (PCR), a number of HBV infections that exhibited viral DNA, in the absence of detectable levels of HBsAg have been reported [11,12]. These cases have been termed occult Hepatitis B infections (OBI), defined as HBV infections, where viral DNA is present in the liver or plasma, without the concomitant detection of HBsAg in plasma [13-16]. Though studies have shown that the levels of HBV DNA in OBI patients are often low [14], OBI remains a significant problem and an active research area in Hepatitis virology, as the infection can still be transmitted *via* blood transfusions or liver transplantations [14,16]. While the exact clinical implications of OBI are still under investigation, OBI has been proposed to contribute to

***Corresponding author:** Jing Zhang, Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT06511, USA, E-mail: jing.zhang.jz349@yale.edu

Received April 29, 2013; **Accepted** June 07, 2013; **Published** June 12, 2013

Citation: Lian Z, Tian QN, Liu Y, Cento V, Salpini R, et al. (2013) Detecting Hepatitis B Viral Amino Acid Sequence Mutations in Occult Hepatitis B Infections *via* Bayesian Partition Model. J Proteomics Bioinform S6: 005. doi:10.4172/jpb.S6-005

the development of cirrhosis and hepatocellular carcinoma, much like chronic HBV infections [14-16]. Indeed, Fang et al. [17] demonstrated in 2009 that, up to 70.4% of hepatocellular carcinoma patients testing negative for HBsAg are infected with OBI, and in 2011, Shi et al. [15] reported that OBI patients have as high as a 2.44-fold increased risk of developing hepatocellular carcinoma. Furthermore, OBI has been shown to be capable of being reactivated in patients undergoing immunosuppressive therapy, which can later lead to potentially life-threatening clinical situations [14,18]. As such, a better understanding of OBI holds great potential in giving rise to a significant advancement in the research, and development of medications and therapies against these major liver disorders. Unfortunately, the etiology and pathogenesis of OBI remain elusive to date, and the genetic characteristics of HBV DNA that lead to the development of OBI are still poorly understood [16].

There are eight HBV genotypes (A-H), based on the variation of the complete nucleotide sequence of the HBV genome. Among these eight genotypes, genotype-C patients were mainly distributed in Asia, while genotype-D patients dominated in Southern Europe and Middle East [19]. In this paper, we mainly focus on the analysis of genotype-C and genotype-D data. A greedy search algorithm is applied to find the positions where amino acid mutations can discriminate genotype-C and genotype-D OBI samples from chronically infected samples. Moreover, this paper applies a novel statistical approach, Bayesian Variable Partition Model [20], to pinpoint those positions where the distributions of amino acids in OBI samples are not only different from those in chronically infected samples, but highly correlated with each other. Different from other traditional methods, the Bayesian-based method is able to analyze high-order combinations of positions. This study is conducted on 358 HBV reverse transcriptase (RT) amino acid sequences from 330 chronic genotype-C HBV patients [21-30], and 28 genotype-C HBV OBI patients [31], as well as 83 chronic genotype-D and 24 genotype-D OBI nucleotide sequences [16]. We believe the findings highlighted in this paper shed light on the future understanding of genotype-C and genotype-D OBI.

Materials and Methods

Data

330 chronic genotype-C HBV and 28 genotype-C OBI RT amino acid sequences (344 aa long) were downloaded from Stanford University HBVrtDB [32], which is available at http://hivdb.stanford.edu/HBV/releaseNotes/.All these data were published [21-31]. The RT sequences of genotype-C cover the last 8 amino acids of the pre-S2, and the entire HBsAg (226 amino acid).

In addition, 83 chronic genotype-D HBV and 24 genotype-D OBI nucleotide sequences were obtained from plasma samples of HBV-infected patients, followed in different centers in Central Italy. All patients with OBI fulfilled the criteria reported in the Taormina Statement [33]. The methodology for RT/HBsAg sequencing is reported in Svicher et al. [16]. Different from the genotype-C data (344 aa long), only parts of amino acid sequences were obtained corresponding 126th to 171th amino acid (46 aa long) in the reference sequence of genotype-D [16,34]. The RT sequences of genotype-D cover the amino acid region 118-163 of the HBsAg.

Stored plasma samples derived from patients with occult HBV infection were retrospectively retrieved and included in the analysis. Ethic approval was deemed unnecessary, because, under Italian law, biomedical research is subjected to previous approval by ethics committees, only in the hypothesis of clinical trials on medicinal products for clinical use (art. 6 and art. 9, leg. decree 211/2003). The research also was conducted on DNA samples and data previously anonymized, according to the requirements set by Italian Data Protection Code (leg. decree 196/2003). Before their first genotypic test, patients sign a consent to approve future analysis of the virus detected in their blood withdrawal.

Greedy algorithm

The flow chart of the greedy algorithm is shown in Figure 1. In the algorithm, we first only add single positions to the subset to classify all the samples, as much as possible. If all the samples cannot be classified correctly, selected discriminant combinations of two positions will be added into the subset. In our case, it is enough to use single positions and the combinations of two positions to classify all the samples. Therefore, we do not consider the combinations of three positions. After all the samples are correctly classified, we prune the subset to make it as small as possible, by testing the redundancy of each single position.

Bayesian variable partition model

Given two data matrices $A=[A_1, \ldots, A_m]$ (of dimension $n_A \times m$) and $B=[B_1, \ldots, B_m]$ (of dimension $n_B \times m$), where A and B denote occult HBV sequences and chronic HBV sequences (control group), respectively (each row is a sequence, each column is a position of HBV amino acid sequence). Here n_A or n_B denotes the number of sequences in occult HBV or control group, and m denotes the number of positions. For the distribution of the positions from these two groups, we have the following four hypotheses:

H1: The identity of the **independent** amino acid positions, where occult and chronic HBV sequences share the same probability mass function.

H2: The identity of the **independent** amino acid positions, where occult and chronic HBV sequences have different probability mass functions.

H3: The identity of the **dependent** amino acid positions, where occult and chronic HBV sequences share the same probability mass function.

H4: The identity of the **dependent** amino acid positions, where occult and chronic HBV sequences have different probability mass functions.

We are especially interested in positions in H2 and H4. Given that position i is from H2 hypothesis, assume that there are c_i possible values (amino acid) at position i. Suppose for every sequence in group A, we have p_1 for the first value, p_2 for the second value, …, p_{ci} for the last value, and $\sum_{j=1}^{c_i} p_j = 1$. Then, the likelihood for group A's data at position i is

$$P(\mathrm{A}_i \mid p_1, \ldots, p_{ci}, H2) = \prod_{j=1}^{c_i} p_j^{n_j}$$

Where n_j denotes the number of sequences who take the j-th value in A_i. For every sequence in B, we have p'_1 for the first value, p'_2 for

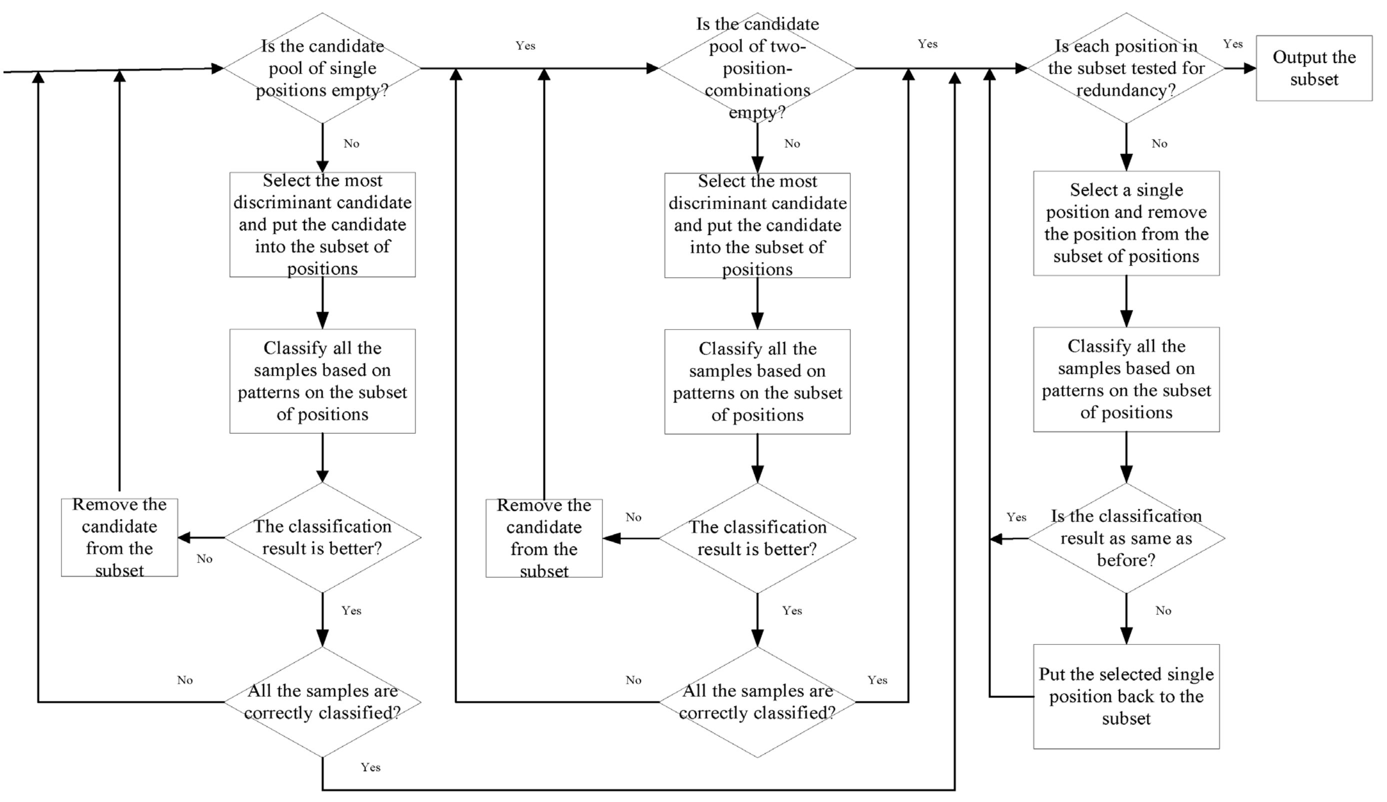
Is the candidate pool of single positions empty?
Yes
Is the candidate pool of two-position-combinations empty?
Yes
Is each position in the subset tested for redundancy?
Yes
Output the subset
No
Select the most discriminant candidate and put the candidate into the subset of positions
No
Select the most discriminant candidate and put the candidate into the subset of positions
No
Select a single position and remove the position from the subset of positions
Classify all the samples based on patterns on the subset of positions
Classify all the samples based on patterns on the subset of positions
Classify all the samples based on patterns on the subset of positions
Remove the candidate from the subset
No
The classification result is better?
Remove the candidate from the subset
No
The classification result is better?
Yes
Is the classification result as same as before?
Yes
Yes
No
No
All the samples are correctly classified?
No
All the samples are correctly classified?
Yes
Put the selected single position back to the subset
Yes

Figure 1: Flow chart of greedy algorithm.

the second value,…, p'_{ci} for the last value, and $\sum_{j=1}^{c_i} p'_j = 1$. Then, the likelihood for group B's data at the position *i* is

$$P(\mathrm{B}_i \mid p'_1,\dots,p'_{c_i}, H2) = \prod_{j=1}^{c_i} (p'_j)^{n'_j}$$

Where n'_j denotes the number of individuals who take the *j*-th value in B_i.

H2 means $p_j \neq p'_j$.However, we do not know these p_j and p'_j. So, we assume they are random and use Dirichlet prior on them.

$$p \sim Dirichlet(\alpha_1,\dots,\alpha_{c_i}) : P(p_1,\dots,p_{ci} \mid H2, \alpha_1,\dots,\alpha_{c_i}) = \frac{1}{B(\alpha)}\prod_{j=1}^{c_i} p_j^{\alpha_j - 1},$$

where $B(\alpha) = \dfrac{\prod_{j=1}^{c_i}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{c_i}\alpha_j)}$, $\alpha = (\alpha_1,\dots,\alpha_{c_i})$, and

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt;$$

$$p' \sim Dirichlet(\alpha'_1,\dots,\alpha'_{c_i}) : P(p'_1,\dots,p'_{ci} \mid H2, \alpha'_1,\dots,\alpha'_{c_i}) = \frac{1}{B(\alpha')}\prod_{j=1}^{c_i} (p'_j)^{\alpha'_j - 1},$$

where $B(\alpha') = \dfrac{\prod_{j=1}^{c_i}\Gamma(\alpha'_j)}{\Gamma(\sum_{j=1}^{c_i}\alpha'_j)}$, $\alpha' = (\alpha'_1,\dots,\alpha'_{c_i})$, and

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

So,

$$P(A_i, p_1,\dots,p_{c_i} \mid H2) = \prod_{j=1}^{c_i} p_j^{n_j} \times$$

$$Dirichlet(\alpha_1,\dots,\alpha_{c_i}) = \frac{1}{B(\alpha)}\prod_{j=1}^{c_i} p_j^{n_j+\alpha_j-1};$$

$$P(B_i, p'_1,\dots,p'_{c_i} \mid H2) = \prod_{j=1}^{c_i} (p'_j)^{n'_j} \times$$

$$Dirichlet(\alpha'_1,\dots,\alpha'_{c_i}) = \frac{1}{B(\alpha')}\prod_{j=1}^{c_i} (p'_j)^{n'_j+\alpha'_j-1},$$

Integrating p and p', respectively, we have:

$$P(A_i \mid H2) = \int_p P(A_i, p_1,\dots,p_{c_i} \mid H2)dp = \prod_{j=1}^{c_i}\frac{\Gamma(n_j+\alpha_j)}{\Gamma(\alpha_j)}\frac{\Gamma(\sum_{j=1}^{c_i}\alpha_j)}{\Gamma(\sum_{j=1}^{c_i}(n_j+\alpha_j))}$$

$$P(B_i \mid H2) = \int_{p'} P(B_i, p'_1,\dots,p'_{c_i} \mid H2)dp' = \prod_{j=1}^{c_i}\frac{\Gamma(n'_j+\alpha'_j)}{\Gamma(\alpha'_j)}\frac{\Gamma(\sum_{j=1}^{c_i}\alpha'_j)}{\Gamma(\sum_{j=1}^{c_i}(n'_j+\alpha'_j))}$$

Then

$$P(A_i, B_i \mid H2) = P(A_i \mid H2) \times P(B_i \mid H2).$$

H1 means $p_j = p'_j$.So for H1 hypothesis, similarly, we have

$$P(A_i,B_i\mid H1)=\int_p P(A_i,B_i,p_1,\ldots,p_{c_i}\mid H1)dp$$

$$=\int_p \frac{1}{B(\alpha)}\prod_{j=1}^{c_i} p_j^{n_j+n_j'+\alpha_j-1}dp \quad .$$

$$=\prod_{j=1}^{c_i}\frac{\Gamma(n_j+n_j'+\alpha_j)}{\Gamma(\alpha_j)}\frac{\Gamma(\sum_{j=1}^{c_i}\alpha_j)}{\Gamma(\sum_{j=1}^{c_i}(n_j+n_j'+\alpha_j))}$$

For H4 hypothesis, assume that there are c possible value combinations of the dependent positions. Suppose for every sequence in group A, we have p_1 for the first combination, p_2 for the second combination,…, p_c for the last combination, and $\sum_{j=1}^{c}p_j=1$. For everyone in group B, we have p_1' for the first combination, p_2' for the second combination,…, p_c' for the last combination, and $\sum_{j=1}^{c}p_j'=1$. Then, we can obtain:

$$P(\text{dependent positions in A|H4})=\prod_{j=1}^{c}\frac{\Gamma(n_j+\alpha_j)}{\Gamma(\alpha_j)}\frac{\Gamma(\sum_{j=1}^{c}\alpha_j)}{\Gamma(\sum_{j=1}^{c}(n_j+\alpha_j))}$$

$$P(\text{dependent positions in B|H4})=\prod_{j=1}^{c}\frac{\Gamma(n_j'+\alpha_j')}{\Gamma(\alpha_j')}\frac{\Gamma(\sum_{j=1}^{c}\alpha_j')}{\Gamma(\sum_{j=1}^{c}(n_j'+\alpha_j'))}$$

Where n_j (n_j') is the number of the *j*th combination in A (B), and

$$P(\text{dependent positions in A,B}\mid H4)=$$
$$P(\text{dependent positions in A})\times P(\text{dependent positions in B})$$

For H3 hypothesis, similarly, we have

$$P(\text{dependent positions in A,B}\mid H3)=\prod_{j=1}^{c}\frac{\Gamma(n_j+n_j'+\alpha_j)}{\Gamma(\alpha_j)}\frac{\Gamma(\sum_{j=1}^{c}\alpha_j)}{\Gamma(\sum_{j=1}^{c}(n_j+n_j'+\alpha_j))}$$

We define an indicator vector $I=[I_1, \ldots, I_m]$ to indicate the hypothesis of different positions belong to, where $I_i=1$ means position *i* from H1, $I_i=2$ means position *i* from H2, $I_i=3$ means position *i* from H3, and $I_i=4$ means position *i* from H4. Currently, we are interested in the posterior distribution of *I,* given the data matrices A and B, i.e. $P(I\mid A,B)$. According to the Bayes' theorem, we have:

$$P(I\mid A,B)=\frac{P(I)P(A,B\mid I)}{\sum_{\text{all possible }I}P(I)P(A,B\mid I)}$$

Therefore, $P(I\mid A,B)\propto P(I)P(A,B\mid I)$. Based on the four hypotheses, we have

$$P(A,B\mid I)=\prod_{i:I_i=1,2}P(A_i,B_i\mid I_i)\times$$
$$P(\text{dependent positions from H3})\times P(\text{dependent positions from H4})$$

Markov Chain Monte Carlo (MCMC) was used to sample from the posterior probability $P(I\mid A,B)$ *via* the Metropolis- Hastings algorithm. 100 MCMC chains were run. For each HBV sequence, positions with the posterior probability *P(I=4|A,B)>0.95* ("H4 positions") were identified, and all unique combinations of positions were recorded, while duplicates were discarded. The amino acids at each H4 position were then extracted from each HBV sequence for every set of H4 positions, and then were analyzed to determine the frequency of each combination of amino acids, and to identify any commonly-occurring set of dependent mutations in OBI HBV DNA polymerase.

Results

Results on genotype-C Data

As our first trial, we searched all the single positions which discriminated part of genotype-C OBI samples from all the chronic genotype-C samples (control group), as shown in Table 1. We found that some single mutations directly discriminated part of OBI samples from the control samples. For example, S159P or S230P successfully separated two among 28 OBI samples (7.14%) from the controls, respectively (i.e. none of the chronic samples has S159P or S230P). It was clear that the 72th position (L72X) was the most discriminant single position, which correctly classified 4 genotype-C OBI samples. Note that the position 72 and 230 were reported in a highly conserved motif in genotype C [35]. Similarly, we searched the combinations of two and three positions, which discriminated part of genotype-C OBI samples from all the controls, as shown in Table 2 and 3, respectively. Although some single positions did not have the capability to discriminate the OBIs from the controls by itself, it generated a powerful discriminant combination when it was combined with other positions. For example, the single position 318 did not separate any OBI sample from the

Position	Classified OBI sample	Position	Classified OBI sample	Position	Classified OBI sample	Position	Classified OBI sample
D2G	19	T54X	3	Y124C	9	I233M	1
P20L/S	4/16	W58X	2	***Y148H***	8	I269S	25
F28L	27	P59X	4	S159P	12 19	Q271K	24
L29S	27	L66X/P	5/15	***I162T***	5	C287R	13
N36X	27	L72X	3 4 5 17	F166L	7	Y305F	12
T38S	26	S75P	10	K168R	3	Y327C	8
R41G	1	W79R	8	V173A	22	L331P	1
V44X	27	L80V	15	F178L	3	L336T	28
S50X/L	4/26	S85P	15	L217P	26		
R51X/G	4/26	F88X	2	T225A	17		
G52X	6	N121D	25	S230P	11 23		

Table 1: Discriminant single positions in genotype-C.

Positions	Classified OBI sample	Positions	Classified OBI sample	Positions	Classified OBI sample
2/72	3 4 5 17 19	66/72	3 4 17 5 15	72/217	3 4 5 17 26
20/72	3 5 17 4 16	72/75	3 4 5 17 10	72/230	3 4 5 17 11 23
28/72	3 4 5 17 27	72/79	3 4 5 17 8	72/233	1 3 4 5 17
29/72	3 4 5 17 27	72/80	3 4 5 17 15	72/269	3 4 5 17 25
36/72	3 4 5 17 27	72/85	3 4 5 17 15	72/271	3 4 5 17 24
38/72	3 4 5 17 26	72/88	2 3 4 5 17	72/287	3 4 5 17 13
41/72	1 3 4 5 17	72/121	3 4 5 17 25	72/305	3 4 5 17 12
44/72	3 4 5 17 27	72/124	3 4 5 17 9	72/327	3 4 5 17 8
50/72	3 5 17 4 26	72/148	3 4 5 17 8	72/331	1 3 4 5 17
51/72	3 5 17 4 26	72/159	3 4 5 17 12 19	72/336	3 4 5 17 28
52/72	3 4 5 17 6	72/166	3 4 5 17 7	316/336	16 17 18 19 28
58/72	2 3 4 5 17	72/173	3 4 5 17 22		

Table 2: Some discriminant combinations of two positions in genotype-C.

Positions	Classified OBI sample	Positions	Classified OBI sample
72/159/230	3 4 5 17 11 23 12 19	72/316/336	3 4 5 16 18 19 17 28

Table 3: Some discriminant combinations of three positions in genotype-C.

Patterns at (28/38/41/52/55/66/ 72/75/79/121/124/159/ 166/173/230/256/ 271/287/316/336)	Classified OBI sample	Patterns at (28/38/41/52/55/66/ 72/75/79/121/124/159/ 166/173/230/256/ 271/287/316/336)	Classified OBI sample
FT**G**GHLLSWN**N**SFVS**C**QCQL	1	FTRG**R**LLSWNYSFVS**C**QCQL	14
FTRG**R**LLSWN**N**SFVS**C**QCQL	2	FTRGH**P**LSWNYSFVSSQCQL	15
FTRGHL**X**SWN**N**SFVSSQCQL	3	FTRGHLLSWNYSFVSSQC**HM**	16 18
FTRGHL**X**SWNYSFVSSQCQL	4	FTRGHL**X**SWNYSFVSSQC**HM**	17
FTRGH**XX**SWNYSFVSSQCQL	5	FTRGHLLSWNY**P**FVSSQC**HM**	19
FTR**X**HLLSWNYSFVSSQCQL	6	FTRGHLLSWNYSF**L**SSQCQ**M**	20
FTRGHLLSWNYS**L**VSSQCQL	7	FTRGHLLSWNYSF**L**SSQCQL	21
FTRGHLLS**R**NYSFVSSQCQL	8	FTRGHLLSWNYSF**A**SSQCQL	22
FTRGHLLSWN**C**SFVSSQCQL	9	FTRGHLLSWN**N**SFVSS**K**CQL	24
FTRGHLL**P**WNYSFVSSQCQL	10	FTRGHLLSW**D**YSFVSSQCQL	25
FTRGHLLSWNYSFV**P**SQCQL	11 23	F**S**RGHLLSWNYSFVSSQCQL	26
FTRGHLLSWNY**P**FVSSQCQL	12	**L**TRGHLLSWNYSFVSSQCQL	27
FTRG**R**LLSWNYSFVSSQ**R**QL	13	FTRGHLLSWNYSFVSSQCQ**T**	28

Table 4: A subset of discriminant positions and corresponding patterns.

controls. However, when it was combined with position 336, the amino acids 318H+336M at these two positions successfully discriminated additional 4 OBI samples (14.29%) from the controls, besides one OBI sample originally separated by L336T.

Based on the above results, we applied a greedy algorithm to search a subset of all the positions to classify all the samples into the controls and the OBIs correctly. The details of the algorithm can be referred in the section of Methods. The position subset and corresponding amino acids of OBI samples at these positions are listed in Table 4. The blue amino acid in each position is a minor pattern in controls (although the pattern is rare in controls, but still exists). The red amino acid at each position is an important mutation which never exists in the control samples. Although the subset of positions discriminated all the OBI samples from the controls successfully, the subset may not be optimal due to greediness of the search. From Table 4, it is interesting that V173L+S230+S256+Q271+C287+Q316+L336 is a discriminant combination, although V173L is not a discriminant mutation (which also exists in controls), and but if no mutation in S230+S256+Q271+C287+Q316+L336, then V173L only exists in OBI.

The RT mutation V173L has been reported to be associated with drug resistance in patients receiving antiviral treatments, such as adefovir and lamivudine [36]. As a result, this mutation has also been associated with HBV vaccine escape [37]. Some positions in the subset were reported in the highly reserved motif such as 28/38/41/72/75/79/166/173/230/256 [35]. Further experimental investigations on these reserved positions may provide more insights of genotype-C OBI.

Without considering computational burden, we can find more discriminant combinations containing four or more positions. However, it is extremely computational intensive to analyze high-order combinations in this way, because the number of possible combinations is too large (2^{344-1} combinations for genotype C HBV sequence containing 344 amino acids). In observing the complexity and the limitation of current computing power, we applied Bayesian Variable Partition Model [20], to analyze the high-order combinations containing more positions in the genotype-C OBI and control sequences. The details of Bayesian Variable Partition Model can be referred to the section of Method.

Before applying Bayesian Variable Partition Model, prior probabilities of four hypothesizes (*P(H1), P(H2), P(H3) and P(H4)*) must be determined. To investigate the effect of prior probabilities on the number of H4 positions (interactively associated mutations), and the number of unique combinations of H4 positions, six sets of prior probabilities, where *P(H4)=P(H2) and P(H1)=P(H3)*, were chosen, and 25 Markov chains were run for each set. The prior probabilities *P(H4)*=0.25, 0.01, 10^{-4}, 10^{-6}, 10^{-8}, and 10^{-10} were used. The results shown in Figure S1 demonstrate that the magnitude of *P(H4)* does not affect significantly, either the number of H4 positions in each chain or the number of combinations of H4 positions within a given number of runs. Note that we observe several different H4 results from totally 100 runs, which results from the MH chains being "stuck" in one of many local modes. Since different local modes may imply different OBI mechanism, we analyze all the local modes in the results.

As such, the prior probabilities *P(H4)=P(H2)* were chosen to be 10^{-5} while *P(H1)=P(H3)* were set as 0.49999. Among totally 100 runs, 44 unique combinations of H4 positions were found with a mean length of 51.3 positions, in a range of 40 to 67 positions (i.e. the orders of 44 high-way interactions range from 40 to 67) and a standard deviation of 6.6. Each unique combination of H4 positions is given a Combo ID, which is shown in Figure S2, with observed frequencies. Furthermore, we observed the distributions of amino acids at these positions in each H4 Combo, as shown in Table 5. The amino acid patterns listed in Table 5 have frequency difference between controls and OBI samples greater than 10%, and the two-sided p-value smaller than 0.05. The amino acid in red was a mutation. It is interesting to observe that the HBV samples with the mutation H9Y+V278I in Combo 11, 19 and 22, or N337H in Combo 11 and 22, or L267Q in Combo 32, only existed in Chronic HBV group when amino acids at other positions in each Combo do not mutate.

Results on genotype-D Data

In this section, we analyzed genotype-D data using the same methods. Different from Genotype-C data, only parts of amino acid sequences were obtained corresponding 126th to 171th amino acid (46 aa long), in the reference sequence of genotype-D. Similar to genotype-C, we firstly obtained discriminant single positions and discriminant combinations of two positions and three positions, which separated part of OBI samples from the controls, as shown in Tables 6-8, respectively. From the tables, we saw some interesting results. For example, 3 out of 24 samples (12, 13 and 21) were separated from the controls, when the mutation N131K existed. Similarly, 3 OBI samples (6, 11 and 16) were discriminated from the controls due to the mutation R153Q. After that, a greedy search was carried out to find a subset of positions, which discriminated all the OBI samples from the controls. The subset and amino acids on these positions are listed in the Table 9. Similar to Table 4, we also found some interesting mutation positions combined with other positions discriminated part of OBI samples, such as H126Q, M129L+P130Q, and N139K, although these mutation positions did not have the discriminating capability by itself .

Different from genotype-C, four OBI samples (No. 8, 14, 17 and 24) were not separated from the controls. The reason was that their amino acid sequences were the same as those in the controls. However, we observed that some of them (8 and 17) had different nucleotides sequences from the controls, as shown in Table 10. In the table, the red triplet codon in OBI and that in control were translated into the same amino acid in red. The italic nucleotide of the codon in OBI sample was the mutation, while the italic nucleotide in the control was the original nucleotide.

Furthermore, we applied the Bayesian Variable Partition model to identify HBV genetic determinants' association with genotype-D OBI, while chronic genotype-D data were used as control. We generated 100 Markov chains with the prior probabilities *P(H2)=P(H4)=*10^{-3} and *P(H1)=P(H3).* It turned out that there were 16 different combinations of positions satisfying H4 hypothesis. As in Table 5, only part of amino acid patterns are listed in Table 11 whose frequency differences between chronic and OBI samples are greater than 10%, and the two-sided p-value are small than 0.05. It was clear that the pattern R138+N139 was distributed unevenly in OBI samples and Chronic samples. Furthermore, the mutation H126R resulted in that the frequency of amino acids H126R+138R in controls was obviously greater than that in OBI samples, while the mutation at the same position, but different amino acid H126Q contributed to the much higher frequency of H126Q+138R in OBIs, compared to that in controls. For positions 129 and 131, the reference pattern M129+N131 was distributed unevenly in OBI samples and Chronic samples. Moreover, all the genotype-D HBV samples with wildtype M129 and the mutation N131K were missed by the HBsAg detection.

Result comparison between genotype-C and genotype-D Data

First, we compared the amino acid reference sequence of genotype-C (344aa long) with that of genotype-D (46 aa long, from 126 to 171), and found that the part reference sequence of genotype-C from position 126 to 171 was the same as that of genotype-D, except at position 131 and 149. In position 131, the reference amino acid of genotype-C was D, while that in genotype-D was N [34]. In position 149, the reference amino acid in genotype-C was K while that in genotype-D was Q [34].

Furthermore, we found there were two common single discriminant positions existing in both genotype-C and genotype-D OBI samples, among all the discriminant single positions. Moreover, the mutations at these two positions in genotype-C and genotype-D OBI samples were almost the same as Y148H and I162T. Although the mutation at position 148 was shown as "CMC" codon in one genotype D OBI sample, it was possible to obtain the same mutation H as the genotype-C OBI sample, because the "CMC" could be translated into H or P corresponding to "M=A" or "M=C", respectively. No other single discriminant positions co-exist in both of genotype-C and genotype-D data.

Next, we compared the positions satisfying H4 hypothesis in genotype-C, with those in genotype-D. Because there were only 46 amino acids in genotype-D data corresponding to position 126 to 171 in genotype-C data, we focused on H4 positions in this common range. It is possible that the H4 positions in genotype-C only depends on some positions out of the range between position 126 and 171. Therefore, to

Combo ID	Amino acid combination	Chronic	Occult	2-sided p-value
8	W3+H9+N13+P20+R41+V43+V44+S50+R51+T54+W58+L66+L73+S75+W79+L80+V84+S85+F88+I91+T118+S119+F178+V207+K212+L217+S223+I224+I233+L235+S246+V253+Q262+E263+V278+I290+S317+P325+Y327+L331	32.42%	53.57%	0.0359
11	W3+G4+E8+H9+P20+L29+N36+T38+R41+V44+S50+R51+T54+H55+W58+P59+L66+L80+S85+F88+T118+S119+Q125+V142+L144+Y148+K168+F178+I233+S246+V253+S256+E263+R274+V278+C287+P325+A329+L331+N337	26.06%	53.57%	0.0037
11	W3+G4+E8+H9Y+P20+L29+N36+T38+R41+V44+S50+R51+ T54+H55+W58+P59+L66+L80+S85+F88+T118+S119+Q125+V142+L144+Y148+K168+F178+I233+S246+V253+S256+E263+R274+ V278I+C287+P325+A329+L331+N337	17.27%	0	0.0125
11	W3+G4+E8+H9+P20+L29+N36+T38+R41+V44+S50+R51+T54+H55+W58+P59+L66+L80+S85+F88+T118+S119+Q125+V142+L144+Y148+K168+F178+I233+S246+V253+S256+E263+R274+V278+C287+P325+A329+L331+N337H	15.45%	0	0.0209
12	D2+W3+G4+T7+E11+R15+P20+F28+L29+K32+N36+T38+R41+S50+R51+H55+W58+L66+L72+L73+W79+L80+V84+S85+P109+S117+S119+N121+Y124+Q125+M129+S135+S137+V142+L144+S159+I162+I169+V173+L180+M204+K212+V214+L217+T225+L229+I233+F249+V253+K270+Y305+P325+A329	49.09%	25%	0.0172
17	W3+T7+E11+R15+P20+F28+N36+T38+R41+S50+R51+H55+W58+P59+L66+L73+L80+S85+V103+G104+S119+Q125+H126+S135+S137+N139+V142+L144+Y148+L164+L199+K212+L217+S223+I224+I233+N238+K241+V253+Q262+E263+L267+K270+C272+L311+A313+T322+P325+Y327+A329	26.97%	46.43%	0.0471
18	W3+G4+E11+P20+L29+K32+N33+T38+V44+S50+R51+T54+W58+L66+L72+L73+N76+F88+V103+G104+P109+S117+N121+Y124+Q125+M129+D131+L145+Y148+R153+I162+K168+V173+F178+L180+M204+K212+V214+L217+E218+T225+I233+K241+F249+S256+T322+L331+A342	54.24%	32.14%	0.0298
19	W3+G4+E8+H9Y+E11+R15+P20+L29+K32+T38+R41+S50+R51+T54+H55+P59+L66+L72+L73+W79+L80+V84+S85+F88+T118+S119+Y124+Q125+H126+N139+V142+Y148+I162+I169+V173+F178+K212+Q215+L217+I233+S246+F249+V253+E263+K270+R274+V278I+A313+T322+P325+A329+L331	15.45%	0	0.0209
22	D2+W3+G4+H9Y+P20+F28+L29+K32+N36+T38+R41+V43+V44+S50+R51+T54+H55+W58+P59+L66+L72+T118+S119+Q125+Y148+S159+K168+I169+V173+V207+L217+T225+I233+F249+S256+R274+V278I+C287+Y305+T322+L331+N337	17.88%	0	0.0074
22	D2+W3+G4+H9+P20+F28+L29+K32+N36+T38+R41+V43+V44+S50+R51+T54+H55+W58+P59+L66+L72+T118+S119+Q125+Y148+S159+K168+I169+V173+V207+L217+T225+I233+F249+S256+R274+V278+C287+Y305+T322+L331+N337H	18.18%	0	0.0072
26	D2+W3+G4+T7+H9+N13+P20+F28+L29+K32+N36+T38+R41+V44+S50+R51+S53+W58+P59+L66+L73+L80+S85+F88+I91+S117+N121+I122+Q125+S135+S137+N139+Y141+V142+L144+S159+I162+L180+M204+K212+V214+L217+S223+I224+I233+L235+N238+S246+F249+V253+Q262+E263+C272+V278+I290+Y305+A313+S317+L331	25.45%	53.57%	0.0033

31	W3+H9+N13+P20+F28+L29+K32+R41+S50+R51+G52+T54+H55+W58+P59+L66+L72+L80+V84+S85+F88+I91+S117+N121+Q125+S135+N139+V142+L144+Y148+F151+F166+K168+L175+L180+M204+V214+L217+S219+T225+I233+L235+S246+F249+V253+R274+V278+I290+A313+S317+Q319+P325+L331+K333+N337	18.48%	35.71%	0.0446
32	D2+W3+G4+E8+P20+F28+K32+N36+T38+V43+S50+R51+T54+H55+W58+P59+L66+L72+L73+L80+S85+G104+V142+L144+Y148+S159+I162+L164+I169+V173+F178+L199+K212+T225+I233+N238+K241+F249+V253+S256+L267Q+C287+Y305+L311+T322+P325+Y327+A329+C332	14.55%	0	0.0212
34	W3+G4+P20+F28+L29+N33+N36+T38+R41+V43+V44+S50+R51+S53+P59+L66+L73+S85+V103+G104+P109+T118+S119+N121+Y124+M129+D131+Y148+K168+I169+V173+F178+L180+M204+V207+K212+L217+I233+K241+N248+F249+S256+T322+Y327+A342	60%	35.71%	0.0160
39	W3+G4+R15+P20+L29+K32+N36+T38+R41+V43+S50+R51+T54+H55+W58+P59+L66+L73+W79+L80+S85+F88+P109+S117+S119+N121+Y124+Q125+D131+S135+V142+L144+L145+Y148+I162+L164+K168+I169+V173+L180+L199+V207+K212+V214+L217+T225+I233+F249+S256+T322+Y327+A329+L331+A342	49.39%	28.57%	0.0474
41	W3+G4+T7+E11+P20+F28+N33+N36+T38+R41+S50+R51+S53+W58+P59+L66+L72+L73+W79+L80+P109+S117+T118+S119+N121+Y124+M129+D131+S135+S137+L144+L145+R153+I169+V173+L180+M204+K212+V214+L217+T225+I233+N248+K270+L331+A342	55.76%	35.71%	0.0486
42	T7+H9+N13+P20+K32+T38+S50+R51+S53+L66+L72+W79+L80+S85+F88+I91+S117+N121+D131+S137+Y141+I162+K168+F178+L180+L217+S223+I224+T225+I233+L235+S246+F249+V253+Q262+E263+C272+V278+I290+S317+L331+A342	31.21%	53.57%	0.0208
43	H9+N13+P20+F28+L29+K32+R41+V43+S50+R51+T54+W58+P59+L66+L72+W79+L80+V84+S85+F88+I91+S135+N139+V142+L144+Y148+I169+F178+L180+V207+L217+S223+I224+T225+I233+L235+S246+F249+V253+Q262+E263+C272+V278+I290+S317+T322+P325+L331	28.79%	50%	0.0304
44	W3+G4+E11+R15+P20+F28+L29+K32+N36+T38+R41+V43+V44+S50+R51+T54+H55+L66+L72+V84+S85+F88+V103+G104+P109+T118+S119+N121+Y124+Q125+M129+S135+V142+L144+L145+Y148+I162+L164+K168+I169+V173+F178+L180+L199+M204+V207+V214+L217+T225+I233+K241+F249+K270+L311+Y327+L331	48.48%	28.57%	0.0491

Table 5: Detailed interactions for H4 positions in genotype-C.

make the comparison more convincing, we cut amino acid sequences from position 126 to 171 in genotype-C, and repeated our methods on the cut sequences. The H4 positions found by Bayesian Variable Partition model and other detailed results were attached as Table S1-S5 in the Supplementary Material. There are some interesting results through the comparisons as follows.

For positions 129 and 131, as shown in Table 12, the reference amino acids at position 131 in genotype-C and genotype-D are different as mentioned before. In genotype-C, the HBV samples with any mutation at either position 129 or position 131 were successfully detected as HBsAg positive. In genotype-D, the HBV samples with no mutation at position 129, and any mutation at position 131 were missed by HBsAg detection with a high probability.

For positions 135 and 139, as shown in Table 13, there is no mutation at position 139 in genotype-C. The genotype-C HBV samples with the wildtype amino acid S at position 135, and any mutation at position 139 were detected as HBsAg positive successfully, while the genotype-D HBV samples with the wildtype amino acid S at position 135 and any mutation at position 139 were missed by HBsAg detection, with a high probability. On the other hand, in genotype-D HBV samples with no mutation at position 139, the mutation S135F/TYC(TYC may translated into F or S) leaded to HBsAg detection false

Position	Classified OBI sample	Position	Classified OBI sample	Position	Classified OBI sample
H126(-AC)	5	D134(GAM)	4	Y148(CMC)	22
G127 (GRG)	3 19	R138(MRR/ARG/K)	3/4/5	T150I	21
T128(RCT/A--)	2/ 3	N139(RAM/H/AMC)	4/5/20	R153Q	6 11 16
M129(-TG/AKG)	3/16	L140F	16	I162(37)T	13
P130(S)	16	Y141F	15		
N131(I/AMC/K/S)	2/3/12 13 21/16	V142(GWR/RCA/T)	3/4/9		

Table 6: Discriminant single positions in genotype-D.

Positions	Classified OBI sample	Positions	Classified OBI sample	Positions	Classified OBI sample
126(1)/131(6)	2 3 5 10 12 13 16 21	126(1)/153(28)	5 6 10 11 13 16 18 23	131(6)/138(13)	2 3 4 5 12 13 16 21
131(6)/139(14)	2 3 4 5 10 12 13 16 20 21	131(6)/142(17)	2 3 4 9 12 13 16 21	131(6)/153(28)	2 3 6 11 12 13 16 21

Table 7: Some discriminant combinations of two positions in genotype-D.

Positions	Classified OBI sample	Positions	Classified OBI sample	Positions	Classified OBI sample
126(1)/131(6)/134(9)	2 3 4 5 10 12 13 16 18 21 23	126(1)/131(6)/153(28)	2 3 5 6 10 11 12 13 16 18 21 23	126(1)/139(14)/153(28)	3 4 5 6 10 11 13 16 18 20 23
126(1)/142(17)/153(28)	3 4 5 6 9 10 11 13 16 18 23	127(2)/131(6)/139(14)	2 3 4 5 10 12 13 16 19 20 21	131(6)/134(9)/139(14)	1 2 3 4 5 10 12 13 16 20 21
131(6)/135(10)/142(17)	2 3 4 5 9 11 12 13 16 18 21	131(6)/139(14)/141(16)	2 3 4 5 10 12 13 15 16 20 21	131(6)/139(14)/142(17)	2 3 4 5 9 10 12 13 16 20 21
131(6)/139(14)/ 148(23)	2 3 4 5 10 12 13 16 20 21 22	131(6)/139(14)/151(26)	2 3 4 5 9 10 12 13 16 20 21	131(6)/139(14)/ 153(28)	2 3 4 5 6 10 11 12 13 16 20 21

Table 8: Some discriminant combinations of three positions in genotype-D.

Patterns at (126/127/129/130/131/134/139/141/142/148/151/153)	Classified OBI sample	Patterns at (126/127/129/130/131/134/139/141/142/148/151/153)	Classified OBI sample
H G M P N D K Y V Y F R	1	H G M P K D N Y V Y F R	12 21
H G M P I D T Y E Y F R	2	Q G M P K D N Y V Y Y R	13
CAM GRG -TG P AMC D N Y GWR Y F R	3	H G M Q N D N F V Y F R	15
H G ATR P N GAM RAM Y RCA Y F R	4	H G AKG S S D N Y E Y F Q	16
-AC G M P N D H Y E Y F R	5	Q G M P N D N Y V Y F R	18,23
H G M P N D N Y V Y F Q	6	H GRG M P N D N Y V Y F R	19
H G L Q N D N Y V Y F R	7	H G M P N D AMC Y V Y F R	20
Y G M P N D N Y T Y Y R	9	H G M P N D N Y V TMC F R	22
Q G M P H D T Y V Y F R	10		
H G M P N D N Y I Y F Q	11		

Table 9: A subset of discriminant positions and corresponding patterns in genotype-D.

negative with a high probability, while the genotype-D HBV samples with other mutations at the same position 135 were still detected as HBsAg positive successfully with a high probability.

For positions 126 and 135, as shown in Table 14, there are 23 different mutation combinations in genotype-D, while there are only 3 mutation combinations in genotype-C, and there is no mutation in the OBI samples of genotype-C especially. In genotype-D, the HBV samples have mutations at position 135 when H126R observed, and these HBV samples were detected as HBsAg positive successfully, no matter what mutation present in position 135. While H126Q present in genotype-D, the HBV samples were missed by HBsAg detection with a high probability, no matter what is present at position 135.

For positions 138 and 139, as shown in Table 15, there are fewer mutations in position 138 in both genotype-C and genotype D. When no mutation is present in position 138, the HBV with any mutation at position 139 in genotype-C were detected as HBsAg positive successfully, while the cases in genotype-D were different. When there is no mutation at position 138 and the mutation N139T/K/AMC observed, the HBV samples were missed by HBsAg detection with a high probability, while the HBV samples with other mutations at position 139 were still detected as HBsAg positive.

For positions 140 and 143, as shown in Table 16, there is no mutation in either position 140 or position 143 in both genotype-C chronic and OBI samples. When no mutation existed in position 140 in genotype-D, the HBV samples with mutation at position 143 from S mutated into L/P were detected as HBsAg positive successfully, while the HBV samples with S143I were missed by HBsAg detection.

OBI sample No.	OBI sample nucleotide sequence	Chronic sample No.	Chronic sample nucleotide sequence	Amino acid sequence
8	CACGGGATCATGCCGAACCT GCACGACTC***A***TGCTC***C***AGGA ACCTCTATGAATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCCTGGGC TTTCG***G***AAAATTCCTATG	49	CACGGGATCATGCCGAACCT GCACGACTC***C***TGCTC***A***AGGA ACCTCTATGAATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCCTGGGC TTTCG***C***AAAATTCCTATG	HGIMPNLH DSCSRNLY ESLLLLYQT FGRKLHLY SHPIILGFR KIPM
14	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA ACCTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCCTGGGC TTTCGGAAAATTCCTATG	3 8 14 26 66 73 82	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA ACCTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCCTGGGC TTTCGGAAAATTCCTATG	HGTMPNLH DSCSRNLY VSLLLLYQT FGRKLHLY SHPIILGFR KIPM
17	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA A***T***CTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCATGGGC TTTCGGAAAATTCCTATG	7 55	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA A***C***CTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCATGGGC TTTCGGAAAATTCCTATG	HGTMPNLH DSCSRNLY VSLLLLYQT FGRKLHLY SHPIIMGFR KIPM
24	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA ACCTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCATGGGC TTTCGGAAAATTCCTATG	7 55	CACGGGACCATGCCGAACCT GCACGACTCCTGCTCAAGGA ACCTCTATGTATCCCTCCTG TTGCTGTACCAAACCTTCGG ACGGAAATTGCACCTGTATT CCCATCCCATCATCATGGGC TTTCGGAAAATTCCTATG	HGTMPNLH DSCSRNLY VSLLLLYQT FGRKLHLY SHPIIMGFR KIPM

Table 10: Nucleotide sequences and amino acid sequences of unclassified OBI samples in genotype-D.

Combo ID	Amino acid combination	Chronic	Occult	Two-sided p-value
4	R138+N139	90.36%	70.83%	0.0390
6	H126R+R138	16.87%	0	0.0362
6	H126Q+R138	1.20%	16.67%	0.0087
16	M129+N131	80.72%	62.5%	0.0976
16	M129+N131K	0	12.5%	0.0102

Table 11: Detailed interactions for H4 positions in genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
M129+D131	0.872727273	1	0.036144578	0
M129+N131	0.012121212	0	0.807228916	0.625
M129+131K	0	0	0	0.125
M129+131I	0	0	0	0.041666667
M129+131H	0	0	0	0.041666667
M129+131G	0.006060606	0	0	0
M129+131E	0.003030303	0	0	0
M129+131RAC	0	0	0.024096386	0
M129+131MAC	0	0	0.012048193	0
M129L+131D	0.063636364	0	0	0
M129L+N131	0.042424242	0	0.048192771	0.041666667
M129L+131H	0	0	0.012048193	0
M129HTG+N131	0	0	0.012048193	0
M129MTG+N131	0	0	0.012048193	0
M129RTG+N131	0	0	0.012048193	0
M129ATR+N131	0	0	0.012048193	0.041666667
M129-TG+131AMC	0	0	0	0.041666667
M129K+131D	0	0	0.012048193	0

Table 12: Distribution comparison of amino acids at positions 129 and 131 in genotype-C and genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
S135+N139	0.933333333	1	0.56626506	0.541666667
S135N+N139	0	0	0.060240964	0
S135H+N139	0	0	0.012048193	0
S135Y+N139	0	0	0.180722892	0.083333333
S135T+N139	0	0	0.012048193	0
S135C+N139	0	0	0.036144578	0
S135KAC+N139	0	0	0.012048193	0
S135F+N139	0	0	0.012048193	0.083333333
S135TYC+N139	0	0	0.012048193	0.041666667
S135+N139Q	0	0	0.024096386	0
S135+N139T	0.012121212	0	0.012048193	0.083333333
S135+N139AMC	0	0	0	0.041666667
S135 N139H	0.027272727	0	0	0
S135+N139MAC	0	0	0.012048193	0
S135+N139RAC	0	0	0.012048193	0
S135+N139D	0.009090909	0	0	0
S135+N139K	0.009090909	0	0.012048193	0.041666667
S135+N139RAM	0	0	0	0.041666667
S135Y+N139K	0.003030303	0	0	0
S135T+N139K	0.003030303	0	0	0
S135Y+N139MAC	0	0	0.012048193	0
S135Y+N139D	0	0	0.012048193	0
S135T+N139Q	0.003030303	0	0	0

Table 13: Distribution comparison of amino acids at positions 135 and 139 in genotype-C and genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
H126+ S135	0.948484848	1	0.578313253	0.625
H126R+ S135Y	0	0	0.13253012	0
H126R+ S135H	0	0	0.012048193	0
H126R+ S135N	0	0	0.012048193	0
H126R+ S135KAC	0	0	0.012048193	0
H126Q+ S135F	0	0	0	0.041666667
H126Q+S135Y	0	0	0	0.041666667
H126Q+S135	0	0	0.012048193	0.083333333
H126CRC+ S135Y	0	0	0.012048193	0
H126CAM+ S135Y	0	0	0.012048193	0
H126+ S135Y	0.003030303	0	0.048192771	0.041666667
H126Y+ S135	0.042424242	0	0.036144578	0.041666667
H126Y+S135TYC	0	0	0.012048193	0
H126Y+S135C	0	0	0.012048193	0
H126+ S135C	0	0	0.024096386	0
H126+S135N	0	0	0.048192771	0
H126+S135F	0	0	0.012048193	0.041666667
H126+S135T	0.006060606	0	0.012048193	0
H126-AC+S135C	0	0	0	0.041666667
H126CAM+S135TYC	0	0	0	0.041666667
H126YAC+S135	0	0	0.012048193	0

Table 14: Distribution comparison of amino acids at positions 126 and 135 in genotype-C and genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
R138+N139	0.924242424	1	0.903614458	0.708333333
R138+N139T	0.012121212	0	0.012048193	0.083333333
R138+N139K	0.015151515	0	0.012048193	0.041666667
R138+N139AMC	0	0	0	0.041666667
R138+N139H	0.018181818	0	0	0
R138+N139D	0.009090909	0	0.012048193	0
R138+N139Q	0.003030303	0	0.024096386	0
R138+N139MAC	0	0	0.012048193	0
R138+N139RAC	0	0	0.012048193	0
R138Q+N139	0.003030303	0	0	0
R138ARG+N139RAM	0	0	0	0.041666667
R138MRR+N139	0	0	0	0.041666667
R138ARR+N139MAC	0	0	0.012048193	0
R138K+N139H	0.009090909	0	0	0.041666667
R138K+N139N	0.006060606	0	0	0

Table 15: Distribution comparison of amino acids at positions 138 and 139 in genotype-C and genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
L140+S143	1	1	0.975903614	0.916666667
L140+S143L	0	0	0.012048193	0
L140+S143P	0	0	0.012048193	0
L140F+S143T	0	0	0	0.041666667
L140+S143I	0	0	0	0.041666667

Table 16: Distribution comparison of amino acids at positions 140 and 143 in genotype-C and genotype-D.

Amino acids	genotype-C chronics	genotype-C OBI	genotype-D chronics	genotype-D OBI
V142+L144	0.996969697	1	0.939759036	0.666666667
V142D+L144I	0.003030303	0	0	0
V142RCA+L144	0	0	0	0.041666667
V142GWR+L144	0	0	0	0.041666667
V142I+L144	0	0	0.012048193	0.041666667
V142RTA+L144	0	0	0.012048193	0
V142E+L144	0	0	0.036144578	0.166666667
V142T+L144	0	0	0	0.041666667

Table 17: Distribution comparison of amino acids at positions 142 and 144 in genotype-C and genotype-D.

For positions 142 and 144, as shown in Table 17, there is almost no mutation in either position 142 or position 144 in both genotype-C chronic and OBI samples. When no mutation existed in position 144 in genotype-D, the genotype-D HBV samples with any mutation at position 142 were missed by HBsAg detection with a high probability.

Discussion and Conclusions

In our recent study [20], we developed advanced Bayesian methods (Bayesian partition model, BVP, and Recursive Model Selection, RMS) to study epstasis among the drug-resistance associated mutations, and the structural basis of these epstatic effects. These Bayesian methods now have been widely applied in a number of HIV and HBV studies [38-43]. In this paper, the pioneering Bayesian method enabled us to detect and analyze the complex interactions of HBV OBI mutations, which directly provide means to elicit the occult sample from the chronic, and hopefully will also give us some insights on the etiology of HBV.

Previous studies have mainly focused the attention on the analysis of the HBsAg. In Yuan et al. [44], specific amino acid substitutions in the regions from amino acids 117 to 121 and amino acids 144 to 147, located in the major hydrophilic region of the S gene, was reported to be associated with carriers with OBIs. In Huang et al. [45], ten representative mutations in the regions from amino acids 119 to 145 were identified in the OBI group. In our study, some discriminant and correlated position combinations were found in the corresponding region in the RT amino acid sequences of genotype-D, such as RT amino acid position combinations 126+138, 129+131 and 138+139 (corresponding to amino acid 118+130, 121+123 and 130+131 in the S gene). We also detected substantial amino acid distribution differences between genotype-C and genotype-D in six different position combinations, which are located corresponding to the same region of S gene. All of these correlations among multiple amino acid positions can only be revealed by our pioneering Bayesian method compared to conventional analysis, focusing on the mutations in single positions. Furthermore, some interesting discriminant position combinations, but out of the region were detected, such as the mutation H9Y+V278I in Combo 11, 19 and 22, or N337H in Combo 11 and 22, or L267Q in Combo 32 when amino acids at other positions in each Combo do not mutate. These high-order position combinations correlated with the OBI, which may exist only in the RT amino acid sequences, may provide new insight into the involvement of the RT proteins in mechanisms underlying occult HBV infection.

Indeed, there remains a significant amount of further work that can be done on the project. As the numbers of unique combinations of H4 positions obtained were quite high after just 200 chains, it is likely that the MCMC runs were prone to being trapped in local modes. Hence, more advanced MCMC technologies may need to be conducted, and analyses regarding the frequencies of each unique combination may need to be performed in order to locate the global optimum.

Despite all the possibilities that may emerge, we are positive that the method, and results presented here will enlighten new and more accurate ways to decipher the myths behind HBV and other related diseases.

Acknowledgements

Jing Zhang was supported by start-up funding and Sesseel Award from Yale University. Cong Li acknowledges the support from China Scholarship Council. The computation was done with the help from the Yale University Biomedical High Performance Computing Center, which was supported by the NIH grant RR19895. Supplementary Material can be downloaded from www.stat.yale.edu/~jz349/HBV_OBI/Supplementary_Material.zip (including all the sequences data).

References

1. Hepatitis B Foundation (2009) Statistics. Hepatitis B Foundation, PA, USA.
2. World Health Organization (2012) Hepatitis B: Fact sheet No. 204.
3. Nettleman M (2008) Hepatitis B, eMedicineHealth.
4. Bourne CR, Katen SP, Fulz MR, Packianathan C, Zlotnick A (2009) A mutant hepatitis B virus core protein mimics inhibitors of icosahedral capsid self-assembly. Biochemistry 48: 1736-1742.
5. World Health Organization (2002) Hepatitis B: The Hepatitis B virus.
6. Miguet JP, Dhumeaux D (1993) Progress in Hepatology 93. John Libbey Eurotext, France.
7. Voevodin AF, Marx PA (2009) Simian virology. Wiley-Blackwell, USA.
8. Centers for Disease Control and Prevention (2012) Hepatitis B FAQs for Health Professionas. CDC, USA.
9. Kao JH (2008) Diagnosis of hepatitis B virus infection through serological and virological markers. Expert Rev Gastroenterol Hepatol 2: 553-562.
10. Hepatitis B Foundation. Hepatitis B Blood Tests: FAQ.
11. Allain JP (2004) Occult hepatitis B virus infection: implications in transfusion. Vox Sang 86: 83-91.
12. Conjeevaram HS, Lok AS (2001) Occult hepatitis B virus infection: a hidden menace? Hepatology 34: 204-206.
13. van Hemert FJ, Zaaijer HL, Berkhout B, Lukashov VV (2008) Occult hepatitis B infection: an evolutionary scenario. Virol J 5: 146.
14. Gutiérrez-García ML, Fernandez-Rodriguez CM, Lledo-Navarro JL, Buhigas-Garcia I (2011) Prevalence of occult hepatitis B virus infection. World J Gastroenterol 17: 1538-1542.
15. Shi Y, Wu YH, Wu W, Zhang WJ, Yang J, et al. (2012) Association between occult hepatitis B infection and the risk of hepatocellular carcinoma: a meta-analysis. Liver Int 32: 231-240.
16. Svicher V, Cento V, Bernassola M, Neumann-Fraune M, Van Hemert F, et al. (2012) Novel HBsAg markers tightly correlate with occult HBV infection and strongly affect HBsAg detection. Antiviral Res 93: 86-93.
17. Fang Y, Shang QL, Liu JY, Li D, Xu WZ, et al. (2009) Prevalence of occult hepatitis B virus infection among hepatopathy patients and healthy people in China. J Infect 58: 383-388.
18. Coppola N, Tonziello G, Pisaturo M, Messina V, Guastafierro S, et al. (2011) Reactivation of overt and occult hepatitis B infection in various immunosuppressive settings. J Med Virol 83: 1909-1916.
19. Mohanty SR, Kupfer SS, Khiani V (2006) Treatment of chronic hepatitis B. Nat Clin Pract Gastroenterol Hepatol 3: 446-458.
20. Zhang J, Hou T, Wang W, Liu JS (2010) Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. Proc Natl Acad Sci U S A 107: 1321-1326.
21. Liu BM, Li T, Xu J, Li XG, Dong JP, et al. (2010) Characterization of potential antiviral resistance mutations in hepatitis B virus reverse transcriptase sequences in treatment-naïve Chinese patients. Antiviral Res 85: 512-519.
22. Ahn SH, Yuen L, Han KH, Littlejohn M, Chang HY, et al. (2010) Molecular and clinical characteristics of hepatitis B virus in Korea. J Med Virol 82: 1126-1134.
23. Okamoto H, Tsuda F, Sakugawa H, Sastrosoewignjo RI, Imai M, et al. (1988) Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. J Gen Virol 69 : 2575-2583.
24. Odgerel Z, Nho KB, Moon JY, Kee SH, Park KS, et al. (2003) Complete genome sequence and phylogenetic analysis of hepatitis B virus (HBV) isolates from patients with chronic HBV infection in Korea. J Med Virol 71: 499-503.
25. Sakurai M, Sugauchi F, Tsai N, Suzuki S, Hasegawa I, et al. (2004) Genotype and phylogenetic characterization of hepatitis B virus among multi-ethnic cohort in Hawaii. World J Gastroenterol 10: 2218-2222.
26. Chan HL, Tsui SK, Tse CH, Ng EY, Au TC, et al. (2005) Epidemiological and virological characteristics of 2 subgroups of hepatitis B virus genotype C. J Infect Dis 191: 2022-2032.
27. Sugauchi F, Chutaputti A, Orito E, Kato H, Suzuki S, et al. (2002) Hepatitis B virus genotypes and clinical manifestation among hepatitis B carriers in Thailand. J Gastroenterol Hepatol 17: 671-676.
28. Sun X, Rokuhara A, Tanaka E, Gad A, Mutou H, et al. (2005) Nucleotide mutations associated with hepatitis B e antigen negativity. J Med Virol 76: 170-175.
29. Horiike N, Duong TN, Michitaka K, Joko K, Hiasa Y, et al. (2007) Characteristics of lamivudine-resistant hepatitis B virus (HBV) strains with and without breakthrough hepatitis in patients with chronic hepatitis B evaluated by serial HBV full-genome sequences. J Med Virol 79: 911-918.
30. Wang Z, Hou J, Zeng G, Wen S, Tanaka Y, et al. (2007) Distribution and characteristics of hepatitis B virus genotype C subgenotypes in China. J Viral Hepat 14: 426-434.
31. Fang Y, Teng X, Xu WZ, Li D, Zhao HW, et al. (2009) Molecular characterization and functional analysis of occult hepatitis B virus infection in Chinese patients infected with genotype C. J Med Virol 81: 826-835.
32. Rhee SY, Margeridon-Thermet S, Nguyen MH, Liu TF, Kagan RM, et al. (2010) Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. Antiviral Res 88: 269-275.
33. Raimondo G, Allain JP, Brunetto MR, Buendia MA, Chen DS, et al. (2008) Statements from the Taormina expert meeting on occult hepatitis B virus infection. J Hepatol 49: 652-657.
34. Standford University. HBV Site Release Notes Appendix 1 Consensus amino acid reference sequences.
35. Warner N, Locarnini S, Kuiper M, Bartholomeusz A, Ayres A, et al. (2007) The L80I substitution in the reverse transcriptase domain of the hepatitis B virus polymerase is associated with lamivudine resistance and enhanced viral

replication *in vitro*. Antimicrobial Agents Chemother 51: 2285-2292.

36. Lada O, Benhamou Y, Cahour A, Katlama C, Poynard T, et al. (2004) *In vitro* susceptibility of lamivudine-resistant hepatitis B virus to adefovir and tenofovir. Antivir Ther 9: 353-363.

37. Sheldon J, Ramos B, Garcia-Samaniego J, Rios P, Bartholomeusz A, et al. (2007) Selection of hepatitis B virus (HBV) vaccine escape mutants in HBV-infected and HBV/HIV-coinfected patients failing antiretroviral drugs with anti-HBV activity. J Acquir Immune Defic Syndr 46: 279-282.

38. Zhang J, Hou TJ, Liu Y, Chen G, Yang X, et al. (2012) Systematic Investigation on interactions for HIV drug resistance and cross-resistance among protease inhibitors. J Proteome Sci Comput Biol 1: 2.

39. Svicher V, Alteri C, Artese A, Zhang JM, Costa G, et al. (2011) Identification and structural characterization of novel genetic elements in the HIV-1 V3 loop regulating coreceptor usage. Antivir Ther 16: 1035-1045.

40. Svicher V, Chen M, Alteri C, Costa G, Dimonte S, et al. (2011) Key-genetic elements in HIV-1 gp120 V1, V2, and C4 domains tightly and differentially modulate gp120 interaction with the CCR5 and CXCR4 N-terminus and HIV-1 antigenic potential. Antiviral Therapy 16: A14-A14.

41. Svicher V, Cento V, Bernassola M, Neumann-Fraune M, Chen M, et al. (2011) Specific HBsAg genetic-determinants are associated with occult HBV-infection *in vivo* and HBsAg-detection. Antiviral Therapy 16: A85-A85.

42. Alteri C, Artese A, Zhang J, Mercurio F, Costa G, et al. (2011) Signature mutations in V3 and bridging sheet domain of HIV-1 gp120 HIV-1 are specifically associated with dual tropism and modulate the interaction with CCR5 N-terminus. Italian Conference on AIDS and Retroviruses-ICAR 2011, Florence.

43. Chen M, Svicher V, Artese A, Costa G, Alteri C, et al. (2013) Detecting and understanding genetic and structural features in HIV-1 B subtype V3 underlying HIV-1 co-receptor usage. Bioinformatics 29: 451-460.

44. Yuan Q, Ou SH, Chen CR, Ge SX, Pei B, et al. (2010) Molecular characteristics of occult hepatitis B virus from blood donors in southeast China. J Clin Microbiol 48: 357-362.

45. Huang CH, Yuan Q, Chen PJ, Zhang YL, Chen CR, et al. (2012) Influence of mutations in hepatitis B virus surface protein on viral antigenicity and phenotype in occult HBV strains from blood donors. J Hepatol 57: 720-729.

This article was originally published in a special issue, **Microarray Proteomics** handled by Editor(s). Dr. Qiangwei Xia, University of Wisconsin-Madison, USA