ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

The Codon Usage Bias and Mutational Landscape of Polymerase Overlapping Pre-Surface and Surface Genes of HBV

***Mohammed Zia[1], Khalid Omer Abualnaja[1,2], Esam I Azhar[3], Taha A. Kumosani[1,2,4], Elie K. Barbour[1,5], Mai M. El-Daly[3], Sherif A. El-Kafrawy, Hisham O. Akbar[6], Hind B. Fallatah[6], Mohammed I. Dgdgi[7], Ghazi A. Jamjoom[3], Jalaluddin A. Jalal[1].**

1. Department of Biochemistry, King Abdulaziz University Jeddah, Saudi Arabia
2. Bioactive Natural Products Research Group, King Abdulaziz University, Jeddah, Saudi Arabia.
3. Special Infectious Agent Unit, King Fahd Medical Research Centre, King Abdul Aziz University
4. Experimental Biochemistry Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia.
5. Department of Animal and Veterinary Sciences, Faculty of Agricultural and Food Sciences (FAFS), American University of Beirut (AUB).
6. Unit of Gastroenterology and Hepatology, Department of Internal Medicine, King Abdulaziz University, Jeddah, Saudi Arabia.
7. Gastroenterology Department, King Fahad Central Hospital, Jizan, Saudi Arabia.

Manuscript Info

Manuscript History:

Received: 14 October 2015
Final Accepted: 22 November 2015
Published Online: December 2015

Key words:

Codon usage bias, HBV, pre surface, surface, mutational landscape, Saudi Arabia

****Corresponding Author***

Mohammed Zia

Abstract

The Hepatitis B Virus (HBV) is a major health concern worldwide. Its genotypes show geographic distribution. Owing to its unique genomic organisation and nature of its replication cycle, this virus acquired the ability to persist inside the host for a long period of time, evolving under several selection pressures within a limited space of relaxed sites. This study explores the differences in codon usage, positional variability and the differential analysis of mutational landscape, of the pre-surface (PS) and surface (S) genes that overlap polymerase gene of HBV. It also explores the differences in variability of B/T cell epitope coding and non-coding regions in PS and S regions. Samples were collected from chronically infected patients in Saudi Arabia, who had not received anti-viral treatment. The DNA molecules of the HBV extracted from serum samples were amplified and then sequenced for PS and S region. The sequences were genotyped, and the variant frequency of synonymous and non-synonymous mutations in PS and S genes and their overlapping P gene regions were determined. The positional Shannon entropy were determined and compared within and between PS and S regions of HBV. The findings reveal higher variability in the PS region compared to that of the S region. Most of the synonymous mutations in the PS and S gene occur at position S3P1 causing a non-synonymous mutation in the overlapping Pol gene. The non-synonymous mutations in the S region are comparatively well tolerated by the P gene compared to the non-synonymous mutations occurring in the PS gene. A higher transversion was observed at the S1p2 codon site of both PS and S genes. The mutations in the PS and S genes and their overlapping Pol gene show positional and compositional pattern of distribution. There was a codon usage-bias observed between the PS and S regions. In the PS region codons with G or C at the third position in codon were favoured against A or T in the S region. There is probably no immune selective pressure to evolve the PS region, while the S region seems to evolve under that pressure.

.....

Introduction

Hepatitis B virus (HBV) belongs to the family Hepadnaviridae (Drexler et al., 2013). It is classified into eight genotypes A-H (Sunbul, 2014), based on nucleotide divergence exceeding 8% in the entire genome and into 34 sub-genotypes based on greater than 4% of nucleotide divergence within the genotype of HBV (Seeger and Mason, 2000; Zhong et al., 2012). The HBV genotypes differ in geographic distribution, HBeAg seroconversion rates, clinical outcome, prognosis, and response to antiviral treatment (Zhang and Cao, 2011). According to WHO estimates in 2014, 240 million persons were chronically infected with HBV. Chronic hepatitis B virus (CHB) infection is a major risk factor for chronic liver diseases, including cirrhosis and hepatocellular carcinoma (HCC) (Yu et al., 2010). Up to 80% of all HCC cases are due to HBV infection(Okamoto et al., 1988).

The HBV genome is partially double stranded, circular and is approximately 3.2 kb long. Two third of the HBV genome is overlapping, containing four open reading frames, translating into S (surface antigen), P (polymerase), C (core) and X (regulatory) proteins (Xie et al., 2015). The S gene is completely overlapped by the Polymerase gene. The codon in open reading frame (ORF) of surface gene has one nucleotide frame-shift downstream with respect to the codon of the P gene-ORF (Cento et al., 2013). The pre-surface (PS) region is overlapped by Spacer region of the HBV polymerase (Pol). Moreover, the small surface (S) region overlaps the reverse transcriptase domain (RT) of Pol. This genomic arrangement affects the magnitude and dynamics of the mutation of the ORFs present in S and P genes. A synonymous (syn) substitution in one ORF can potentially result in a non-synonymous (non-syn) substitution in the ORF of the overlapping gene (Norder et al., 1994). Other source of mutation in HBV is due to its reverse transcriptase which lacks proof reading (Horvat, 2011). HBV replicates through RNA-intermediated reverse transcription (Buti et al., 2005). Despite the evolutionary constraints acting on HBV, errors in its replication occurs at a much higher rate than those detected in other DNA viruses (Gao and Hu, 2007; Ingman et al., 2006; Norder et al., 1994). Mutations in HBV may occur non-randomly in the overlapping regions and may be established under selection pressure (Chen et al., 2013; Pollicino et al., 2014). The three nucleotide sites in a codon differ in their mutability due to the degeneracy of genetic code (Krakauer, 2000). It is concluded from Wobble hypothesis that the third site in a codon is most tolerant for substitution mutations (Sanger, 1981). Degenerate sites also occur at positions one and two of the codon as well, as seen in the case of leucine and serine amino acids. However, the degenerate site may be under constraint, due to non-degenerate site of the overlapping gene. Thus the frequency of the mutations in overlapping gene is expectedly lower than that of the non-overlapping genes (Torres et al., 2013b). Much of the scope of mutability in overlapping genes is restricted to regions that are not functionally important, or existing under relaxed stoichiometric constraint (Chen et al., 2013) due to usage of amino acid codons with high degenerate sites, such as those that are coded by six, four, three and two synonymous codons in the overlapping gene. There are four different combinations leading to possibility of mutation in overlapping S and P genes of HBV. A non-syn (N_S) mutation in gene PS/S, causing a non-syn (N_P) mutation in gene P-(N_SN_P); a non-syn mutation in gene PS/S (N_S) causing a syn (S_P) mutation in gene P-(N_SS_P); a syn mutation in gene PS/S (S_S) causing a non-syn (N_P) mutation in gene P-(S_SN_P) and a syn mutation in gene PS/S (S_S) causing a syn (S_P) mutation in gene P-(S_SS_P). The N_SN_P mutations may cause a purifying or negative selection, whereas the N_SS_P and S_SN_P mutations are the result of tolerance in the overlapping gene, which may determine a positive selection. The S_SS_P mutation may cause a genetic drift of the sequences.

The aims of this work were i) To analyse the differences in the codon usage of the PS and S and their overlapping spacer and RT domains respectively. ii) To determine the within region (PS/S) and between region (PS and S) variability within codon sites. iii) To determine the mutational landscape in PS and S regions and within their B or T cell epitope coding regions, elucidating the patterns of mutation occurrence in mutation groups N_SN_P, N_SS_P, S_SN_P and S_SS_P and in three codon clusters namely, the High (H: 6 syn codons), Medium (M: 3 and 4 syn codons) and Low (L: 1 and 2 syn codons) at codon sites 1,2 and 3.

Materials and methods

Datasets

The dataset used in this study is comprised of sequenced pre-surface (around 489 base pair length) and surface genes (around 393 base pair length) of HBV from 81 infected subjects that were admitted to the hepatology out-patient clinic and endoscopy unit at King Abdul Aziz University. The subjects had no viral co-infections with HCV or HIV. Sequences with INDELS and or internal stop codons were excluded from the analysis (Hu and Ng, 2012; Hu and Ng, 2013; Kramvis et al., 2008). The HBV genotypes represented in the dataset were those of A, B, C, D and E. The sequences used in this study can be retrieved from the NCBI, BioProject ID PRJNA294996.

Genotyping and Variant detection

Genotypes of sequences were determined by performing multiple sequence alignment with HBV genotype reference sequence, obtained from online NCBI Biosystems Database (Taxonomic Repository Database Reference - db_xref="taxon:10407) (Geer et al., 2010) and HBVdB, using MUSCLE program (Edgar, 2004).The highest alignment similarity scores to that of the reference sequence determined the genotype of a consensus sequence (Rozanov et al., 2004). The "Annotate and predict" tool for variant detection in Geneious software (Kearse et al., 2012) was used to determine the variants with respect to reference sequences (Nielsen et al., 2012).Descriptive analysis of variants was performed, determining the total frequencies of syn and non-syn mutations in PS and S and the corresponding syn and non-syn mutations in the overlapping spacer and RT domains respectively. Transition/ transversion bias in the PS, S, spacer and RT domains was determined using MEGE v6 software.

Comparison of variability within codon sites in PS, S and the overlapping P gene.

Sequences in the PS and S regions were aligned by MUSCLE program in Geneious software version 8 (Kearse et al., 2012). Strings of nucleotides at positions 1, 2 and 3 of each codon in PS and S coding sequences were extracted by stripping the other two nucleotides in each codon using Geneious software (Kearse et al., 2012). This resulted in three strings or sequences, in other words, sequences containing only position 1 or 2 or 3. The position in these strings corresponds to the codon number in the original PS and S sequences. These strings or sequences were imported from Bioedit v.7 (Hall, 1999) program as FASTA files. The positional entropy (Hx), a measure of variability (Hall, 1999), was quantified for each of these strings, and grouped into B/T cell epitope coding regions and non-epitope coding regions. The positional Hx of each codon in PS and S region, at positions 1, 2 and 3 were compared and plotted. Friedman's test was used for comparisons among site entropies within same region while the Mann Whitney U test was used for comparing the overall site to site entropies of PS and S regions and between B/T epitope coding and non-coding regions. These statistical analyses were performed by GraphPad version 6 program.

Compositional analysis of amino acids and mutation frequency at different positions of their codons in PS, S and the overlapping P gene

The sequences were translated in the ORF of S and P gene and the amino acid composition of the overlapping genes PS, spacer and S, RT were determined using BioEdit version 7 software (Hall, 1999). The degeneracy of codons for an amino acid were clustered into three classes, based on the number of synonymous codons used for each amino acid namely, High (H : 6 syn codons), Medium (M: 3 and 4 syn codons) and Low (L: 1 and 2 syn codons).

The detected SNVs in PS and S regions were grouped into four categories namely, the N_SN_P, N_SS_P, S_SN_P, and S_SS_P, where 'S' and 'P' in the subscripts denote surface/pre-surfaceof polymerase regions respectively. The mutation frequencies at each codon position under each amino acid class (H, M, or L) were determined in PS and S regions. For every mutation in PS and S at a codon position (1 or 2 or 3) in each class (H,M, or L), there are possibilities of changing the triplet code of an amino acid in the overlapping P gene mutations, which may be of the class H_P or M_P or L_P, where 'P' in the subscript denotes an amino acid group in the polymerase ORF of overlapping PS and S region; thus, the frequencies of mutations under each class (H, M, or L) in the PS and S region were further sub-classed into H_P, M_P and L_P.

Comparative analysis of codon- usage in PS, S region and overlapping P gene

The aligned sequences of PS, S and their corresponding Pol regions were analysed for codon usage bias, using MEGA v.6 program (Tamura et al., 2013). Correlation analysis between the relative synonymous codon usage frequencies (RSCU) in PS and S and their corresponding polymerase region were performed by Spearman's test present in statistical computing package of SPSS v 22. A lack of correlation implied a presence of dissimilar codon usage. Amino acids coded by one codon (AUG: methionine and UGG: tryptophan) and protein synthesis termination-codons (UAA, UAG and UGA) were not considered in the analysis. The nucleotide base at the third position of most favoured codon(s) of each amino acid in each region, was determined and categorised into two categories namely, those nucleotide bases using two hydrogen bonding- A or U (T) and those using three hydrogen bonds for base pairing- the G or C.

Results

Descriptive of genotypes and variants in PS and S region

The majority of the subjects were infected by genotype D (85%) followed by genotype C (6%), genotype A (4%), genotype E (4%) and genotype B (1%) (Fig. 1).

The variant analysis with respect to the reference sequence of each genotype, revealed 248/523and 106/396 variable sites in the PS and S regions, respectively. The total frequency (f) of mutations observed in the PS region were 679 (PS-S1P2 f=194, PS-S2P3 f=76 and PS-S3P1 f=409) of which 413 (S1P2 f=17, S2P3 f=0 and S3P1 f=396) were syn and 266 (S1P2 f=177, S2P3 f=76 and S3P1 f=13) were non-syn mutations (Table 4).Among the non-syn mutations in PS, 193 caused a non-syn mutation also in the spacer region. The remaining 73 non-syn mutations in PS did not result in amino acid changes and were tolerated by the overlapping codons in the P gene. Among the synonymous mutations in PS, 28 variants were syn also in the overlapping spacer region. The remaining 385 syn mutations in PS caused a non-syn mutation in spacer region (Table 4).

The total mutation frequency (f) in the S region was 285 (S-S1P2 f=35, S-S2P3 f=86 and S-S3P1 f=164), less than half the number observed in the PS region (Table 4). Among the total observed mutations in S region, 124were non-syn mutations (S1P2 f=33, S2P3 f=86 and S3P1 f=5) and 161 were syn mutations (S1P2 f=2, S2P3 f=0 and S3P1 f=159). Among the non-syn mutations, 41 mutations caused a change in both S and RT, and the remaining 83 non-syn mutations of S gene caused no alterations in P gene and were tolerated by overlapping codons in ORF of the P gene. Among syn mutations, 128 affected only the RT region, and these mutations were tolerated by overlapping codons in the ORF of the S gene. There were 33 mutations that did not affect either the S and neither the P gene (Table 4).

The transition/transversion bias (R) in PS at codon sites 1, 2 and 3 were inferred from the Neighbour-joining phylogenetic tree drawn using maximum likelihood statistics by Kimura-2 parameter model. The estimated transition/transversion bias (R) in PS at codon sites 1, 2 and 3 were 0.96, 2.22 and 2.97, respectively. However, the estimated transition/transversion bias (R) in the S region at positions 1, 2 and 3 of codon were 0.56, 1.31 and 3.21, respectively.

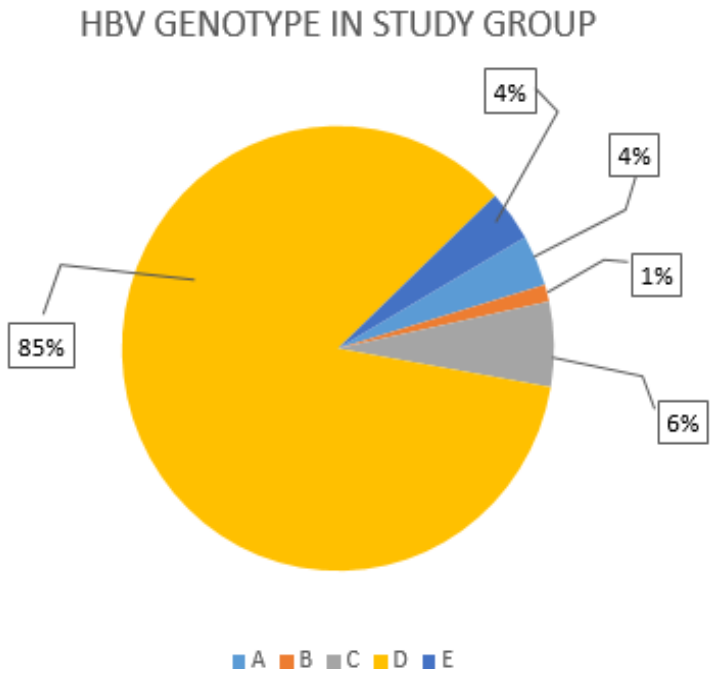

Fig. 1. The frequency of HBV genotypes (A-E) in the studied population

Comparison of variability within codon sites of PS, S and their overlapping regions in the P gene

Positional variability of sites in codon within regions

The PS region showed a significantly higher nucleotide variability (entropy) ($p<0.0001$) at positions 3 compared to that of position 1 (corresponding positions 1 and 2 in spacer region of polymerase). Similarly, there was a significantly higher nucleotide variability at positions 3 compared to that of position 2 ($p<0.0001$) (corresponding positions 1 and 3 in spacer region of polymerase). However there was no significant difference in the nucleotide variability between positions 1 and 2 of codons (corresponding positions 2 and 3 in spacer region of polymerase) (Tables 1&2).

The S region showed a significantly high nucleotide variability at position 2 compared to position 1 of the codon ($p<0.05$). There was no significant difference in variability of position 1 compared to that at position 3 and between positions 2 compared to that at position 3 of the S region.

Positional variability of sites in codon between regions

When comparing counterpart codon positions in the PS and S region, there was a significantly higher variability observed in PS than S ($p<0.001$) as determined by the rank sum differences. (Table 1&2).

Table 1. Descriptive of positional entropies (Nucleotide variability)

Δ in codon	ps1p2	ps2p3	ps3p1	s1p2	s2p3	s3p1
Number of values	161	161	161	132	132	132
25% Percentile	0.0605	0.0605	0.1056	0	0	0
Median	0.1056	0.1056	0.205	0	0.06161	0
75% Percentile	0.1803	0.1871	0.324	0.06161	0.1075	0.1231
Maximum	0.9178	0.819	1.145	0.6329	0.6646	1.069
Mean	0.1373	0.1465	0.2396	0.05978	0.08742	0.1181
Std. Deviation	0.1414	0.1524	0.2039	0.1133	0.121	0.2119
Std. Error of Mean	0.01072	0.01155	0.01546	0.009858	0.01053	0.01845
Lower 95% CI	0.1162	0.1237	0.2091	0.04028	0.06658	0.08158
Upper 95% CI	0.1585	0.1693	0.2701	0.07928	0.1083	0.1546

Table 2. Comparative analysis of codon site variations within and across PS and S regions

Friedman test- positional variations within PS and S regions				
HBV Region/	comparison	Rank sum diff.	Test statistics	P Value[a]
PS	ps1p2 vs. ps2p3	-13	45.35	> 0.9999
	ps1p2 vs. ps3p1	-99.5	45.35	< 0.0001**
	ps2p3 vs. ps3p1	-86.5	45.35	< 0.0001**
S	s1p2 vs. s2p3	-39	11.88	0.0491*
	s1p2 vs. s3p1	-33	11.88	0.1268
	s2p3 vs. s3p1	6	11.88	> 0.9999
Mann Whitney U test- variations at codon sites between PS and S region				
Δ in Codon	comparison	Rank sum diff.	Test statistics	P Value[a]
1	PS vs. S	15791	6812	< 0.0001**
2	PS vs. S	11812	8802	0.0004**
3	PS vs. S	17268	6074	< 0.0001**
a-significant at α<0.05, *-P value <0.05, **-P value <0.001				

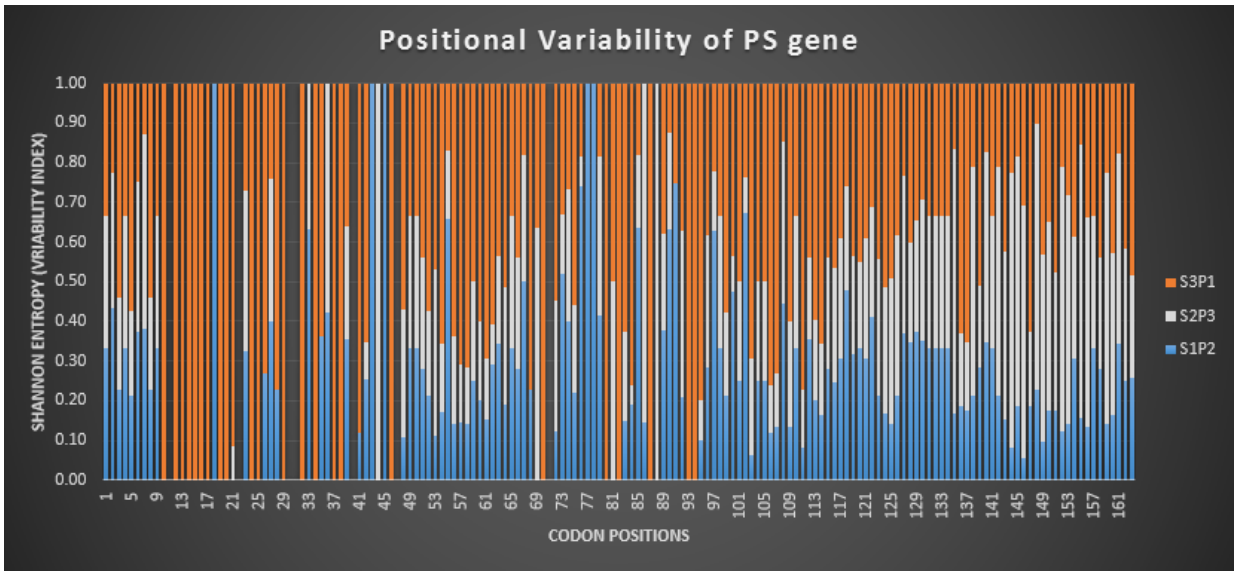

Figure2. Positional variability within codon site along the length of PS gene in HBV

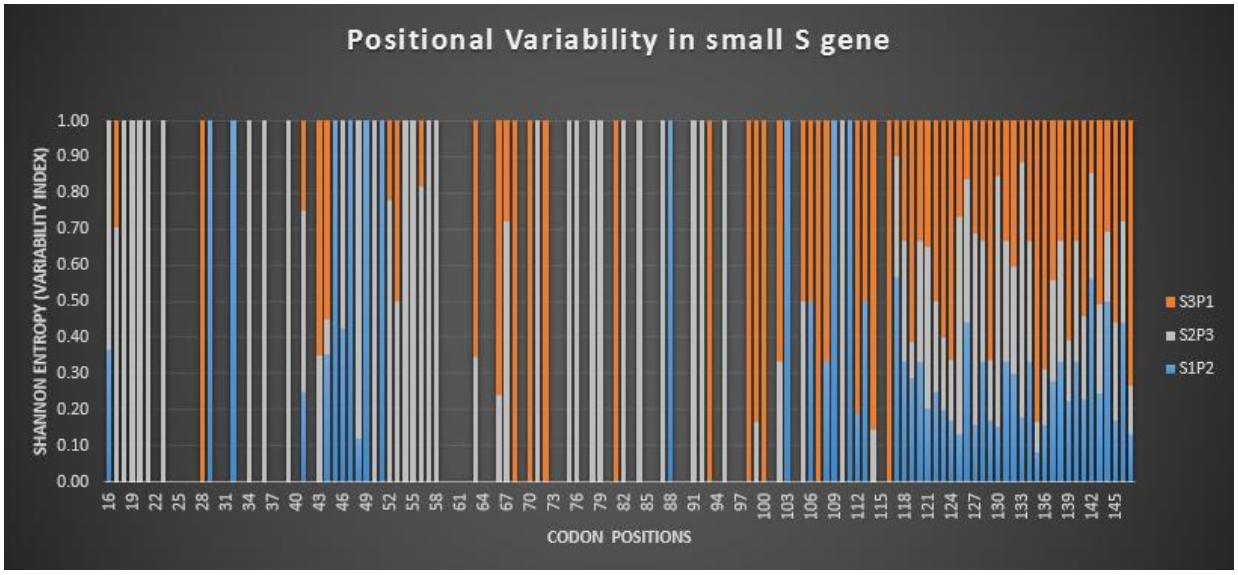

Figure3. Positional variability within codon site along the length of PS gene in HBV

Compositional analysis of amino acids and mutation frequency at different positions of their codons in PS, S and the overlapping P gene

The analysis of predicted amino acid class composition of PS, S, spacer and RT regions revealed a presence of non-significant difference between regions (Table 3). However the slight variations shown in Fig. 4 for classes L and H of amino acids are comparatively more in the ORF of the Pol compared to that of the S, while the M class was higher in composition in the ORF of S compared to that of the Pol.

Table 3. Difference in the amino acid class composition

	Chi-square between Sets of Frequencies			
	PS	SPACER	S	RT
PS	.000	6.516	5.970	5.928
SPACER	6.516	.000	7.020	5.120
S	5.970	7.020	.000	6.599
RT	5.928	5.120	6.599	.000

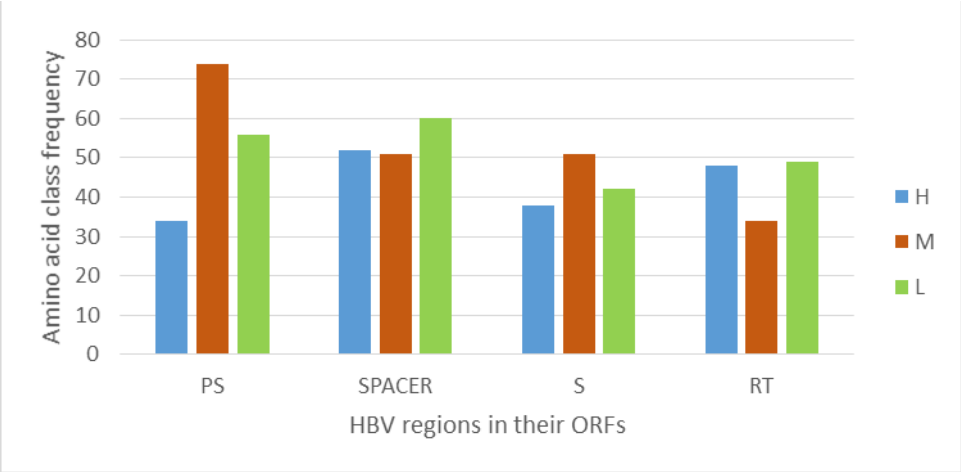

Figure 4. Regional composition of amino acid classes

Table 4 Distribution of mutation frequencies occurring in PS and S regions at each site within a codon and in each codon class

Region	**PS**									**S**								
Mutation type	N_SN_P									N_SN_P								
Δ in codon	S1P2			S2P3			S3P1			S1P2			S2P3			S3P1		
Codon class	H=56	M=55	L=60	H=0	M=7	L=2	H=0	M=1	L=12	H=1	M=18	L=13	H=2	M=3	L=1	H=0	M=0	L=3
H_P	8	23	5	0	2	1	0	0	4	1	7	0	0	0	0	0	0	0
M_P	27	13	15	0	0	0	0	1	4	0	2	11	0	0	1	0	0	3
L_P	21	19	40	0	5	1	0	0	4	0	9	2	2	3	0	0	0	0
Mutation type	N_SS_P									N_SS_P								
Δ in codon	S1P2			S2P3			S3P1			S1P2			S2P3			S3P1		
	H=0	M=0	L=6	H=23	M=27	L=7	H=0	M=0	L=0	H=0	M=0	L=1	H=30	M=28	L=22	H=0	M=0	L=2
H_P	0	0	0	12	9	0	0	0	0	0	0	1	15	5	7	0	0	2
M_P	0	0	0	8	6	0	0	0	0	0	0	0	3	12	6	0	0	0
L_P	0	0	6	13	12	7	0	0	0	0	0	0	12	11	9	0	0	0
Mutation type	S_SN_P									S_SN_P								
Δ in codon	S1P2			S2P3			S3P1			S1P2			S2P3			S3P1		
	H=17	M=0	L=0	H=0	M=0	L=0	H=70	M=159	L=139	H=2	M=0	L=0	H=0	M=0	L=0	H=31	M=81	L=14
H_P	5	0	0	0	0	0	14	59	68	0	0	0	0	0	0	0	17	12
M_P	4	0	0	0	0	0	37	61	47	2	0	0	0	0	0	0	12	1
L_P	8	0	0	0	0	0	19	39	24	0	0	0	0	0	0	31	52	1
Mutation type	S_SS_P									S_SS_P								
	S1P2			S2P3			S3P1			S1P2			S2P3			S3P1		
Δ in codon	H=0	M=0	L-0	H=0	M=0	L=0	H=1	M=20	L=7	H=0	M=0	L=0	H=0	M=0	L=0	H=6	M=8	L=19
H_P	0	0	0	0	0	0	1	18	7	0	0	0	0	0	0	6	8	19
M_P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L_P	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0

NOTE: codon class in P gene : H_P = high degeneracy, M_P = medium degeneracy and L_P = low degeneracy

Mutations causing change in amino acids of both overlapping S and P genes- N_SN_P

In both PS and S region, the S1P2 showed most non-syn mutations, affecting both proteins (PS- N_SN_P f=111, S-N_SN_P f=32). On the other hand, the mutation frequencies in positions S2P3 and S3P1 were low. There was not much difference in the mutation frequencies in amino acid codon classes of the PS (H f=56, M f=55 and L f=60) (Table 4 and Fig. 6). The affected classes in spacer region, due to non-syn mutation in PS gene and in decreasing order were the L_P (f=90), M_P (f=60) and H_P (f=43) (Table 4). Another interesting observation was detected in the PS, where most non syn mutations in L class were confined to L_P class of the Pol, whereas most non-syn mutations in the M class affected the H class and vice versa.

Regarding the S gene, the non-syn mutations, affecting both genes, occurred in the M (f=18) and L (f=13) classes of amino acids, with a lower frequency compared to that of the PS region. The affected amino acid in the RT region were mostly of L_P and M_P classes (Fig. 7). The N_SN_P was the second most frequent mutation type in the PS (PS-N_SN_P f=193) region and the third most frequent in the S region (S-N_SN_P f =41) (Table 4).

Mutations causing no change in amino acid of both overlapping genes- S_SS_P

All syn mutations affecting PS, S and their overlapping spacer and RT regions are exclusively observed at position S3P1 (PS-S_SS_P f =28 and S-S_SS_P f =33) (Figs. 5& 6). These mutations even though seen across all classes of codon in PS (H f=1, M f=20 and L f=7) and S (H f=6, M f=8 and L f=19), are almost exclusively tolerated by H_P class of overlapping codon of P gene and rarely seen in M_P and L_P classes (PS-S_SS_P f =26 and S-S_SS_P f =33).. The S_SS_P type of mutation were not frequently encountered in the PS and S regions (Table 4).

Mutations causing change in amino acid of PS and S gene but not in overlapping P gene- N_SS_P

Mutations resulting in non-syn mutation in PS and S gene but not in their overlapping P gene regions were predominantly observed at position S2P3 in all classes of codons and rarely at position S1P2 (Figs.5 & 6). All amino acid classes showed variable tolerance in overlapping spacer region (H_P-=21, M_P=14 and L_P=32) to mutations in PS (Total PS-S2P3 f=67; H f=33, M f=27 and L f=7). Similarly non–syn mutations in S (Total S-S2P3 f=80; H f=30, M f=28 and L f=22), showed variable levels of tolerance in RT region across all classes of codons (H_P-=27, M_P=21 and L_P=32). The N_SS_P type of mutation was the third most frequent in PS and the second most frequently encountered in S region (Table 4&Fig. 5).

Mutations causing no change in amino acid of PS and S gene but causing amino acid change in respective overlapping Spacer and RT domain of P gene - S_SN_P

The most frequent type of mutation in PS and S was the S_SN_P type, and predominantly seen at S3P1 positions of codon in the PS and S regions, and rarely at S1P2 position (Figs 5&6). These mutations occur across all classes of codons in PS and S region. However the difference is in the proportion of different classes of codon affected in the overlapping P regions and tolerated by codon classes of the PS and S region. In the spacer region, the majority of the non-syn mutations were in the M_P (f=145) followed by H_P (f=141) and L_P (f=82) classes, and tolerated by corresponding classes in the following decreasing order of M (f=159), L (f=139), and H (f=70). Moreover, it the syn mutations in the L and M classes were mostly affecting the H_P and M_P classes of amino acids in the spacer region of the P protein (Table 4&Fig. 6).

In relation to the RT region, the most affected class by syn mutations in S region was the L class of amino acids (L_P-f=84, H_P-f=29 and M_P-f=13), tolerated mostly by the M class, followed by the H class of amino acid codons (H f=31, M f=81 & L f=14) in ORF of the S region (Table 4&Fig. 5).

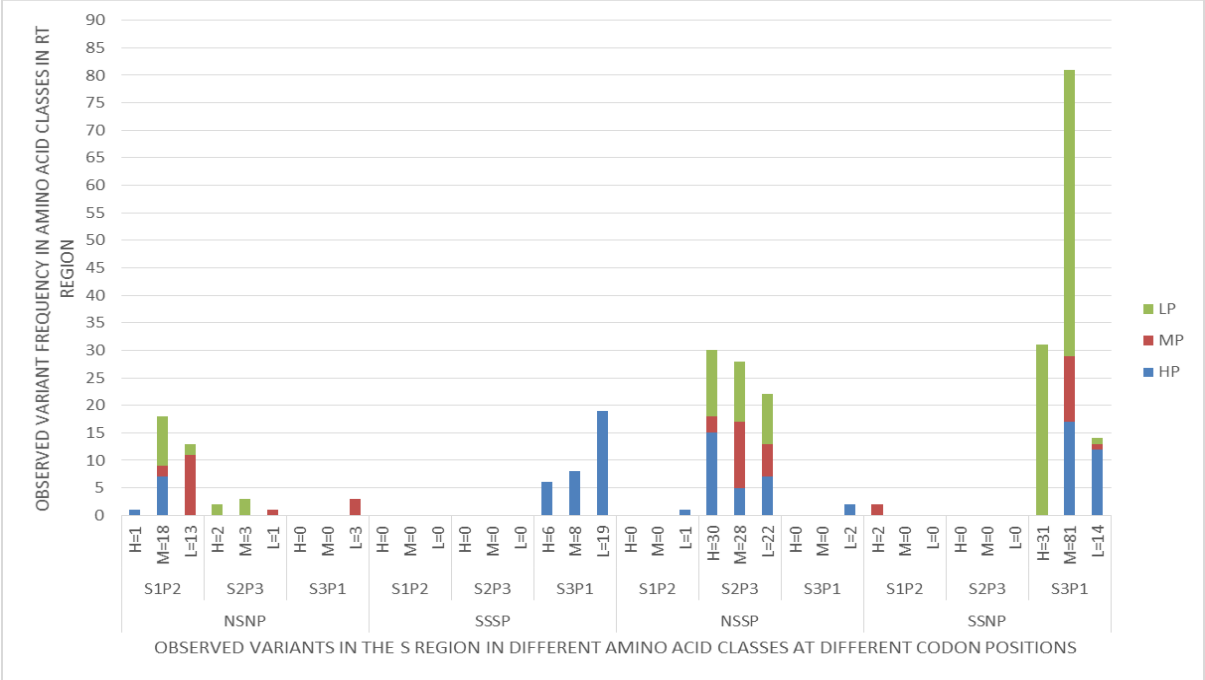

Fig. 5. Mutational landscape of overlapping amino acid classes in the S and RT region of HBV

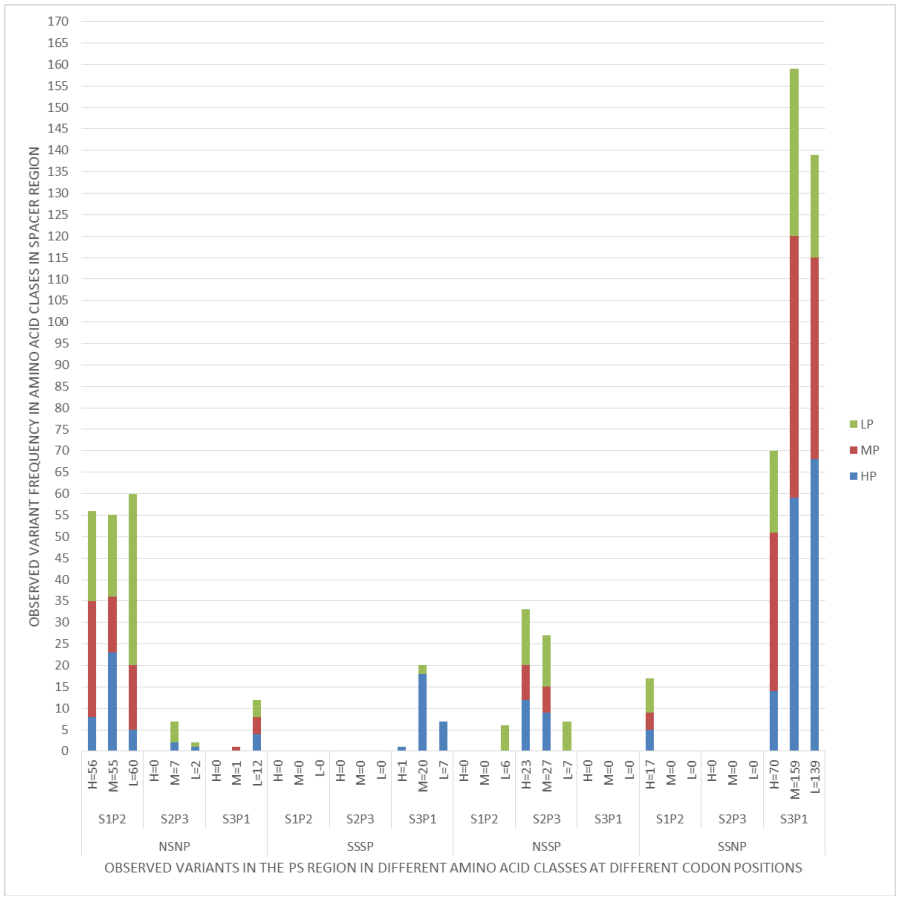

Figure 6. Mutational landscape of overlapping amino acid classes in the PS and Spacer region of HBV

Comparative analysis of codon usage in PS and S regions and the overlapping P gene

Spearman's correlation analysis of relative codon usage (RSCU) frequency of amino acids across regions showed no significant correlation between the RSCU of amino acids in PS versus Spacer region (p=0.823) and between the RSCU of amino acids in PS versus S gene (p=0.14) (Table 5). On the other hand, there was a significant correlation (p<0.01) between the RSCU of amino acids of small S gene versus its overlapping RT domain of polymerase (P-ORF) and between spacer region and RT domain of P ORF (Table 5).The base usage, at third position of codon favored the G and C by the polymerase. The usage of codon having A or T at third position was more frequent in S than that in PS region (Table 6).

Table 5. Spearman's correlation of codon usage between PS and S region and their overlapping P gene regions

Correlation of codon usage between regions		N	Correlation coefficient	P value
PS	SPACER	59	0.029	0.823
S	RT	59	0.438	0.000**
PS	S	59	O.187	0.140
SPACER	RT	59	0.403	0.001**
**. Correlation is significant at the P<0.01 level (2-tailed).				

Table 6.Frequency of nucleotide base at third position of most favoured codon in different genic region of HBV

Nucleotide at the third site in codon	hydrogen bonding	REGION			
		PS	SPACER	S	RT
A/U(T)	2	8	6	10	6
G/C	3	10	12	8	12

Comparative analysis of positional variabilities in B/T cell epitope coding and non coding regions of PS and S genes

In the PS region the 123 amino acid lie in the B or T cell epitope determining regions. The epitope determining regions in the PS are interrupted by the non-epitope coding regions that account for 41 amino acids in the obtained sequences. The positional variability in the B/T epitope coding region and non-epitope coding regions of PS did not show any significant (Table 7 and 8; Fig. 7).

In the S region, there were only 51 amino acids that lie in the B/T epitope coding sites and 81 amino acids that do not lie in it. There was a significantly greater ($p < 0.05$) variability observed in the B/T epitope coding sites at positions 1, 2 and 3 of codon compared to the corresponding codon positions in the non-epitope coding codons of the S region (Tables 9 and 10; Fig. 8).

Table 7. Descriptive of Shannon entropy in PS gene categorized under B/T epitope coding and non-coding regions

B/T epitope	Δ in codon	N	Min	25% Percentile	Median	75% Percentile	Max	Mean	SD	SEM	Lower 95% CI	Upper 95% CI
coding	1	123	0	0.000	0.061	0.205	0.918	0.129	0.144	0.013	0.103	0.154
not coding	1	41	0	0.000	0.106	0.166	0.592	0.136	0.158	0.025	0.086	0.186
coding	2	123	0	0.000	0.106	0.213	0.778	0.141	0.155	0.014	0.114	0.169
not coding	2	41	0	0.000	0.106	0.167	0.761	0.137	0.157	0.025	0.087	0.186
coding	3	123	0	0.106	0.211	0.331	1.145	0.250	0.216	0.020	0.212	0.289
not coding	3	41	0	0.061	0.166	0.389	0.567	0.206	0.178	0.028	0.150	0.262

B/T epitope- B or T cell epitope, Δ- position, N – number of observations, CI- confidence interval, SD- standard deviation from mean, SEM- standard error of mean, min- minimum, max- maximum.

Table 8. Analysis of positional variability of nucleotides in B/T epitope coding and non-coding regions of PS gene				
Δ in Codon	Mean rank of B/T epitope coding	Mean rank of B/T epitope non coding	Mann-Whitney U	P value
1	82.12	83.63	2475	0.8589
2	82.96	81.13	2466	0.8303
3	84.75	75.74	2245	0.2934

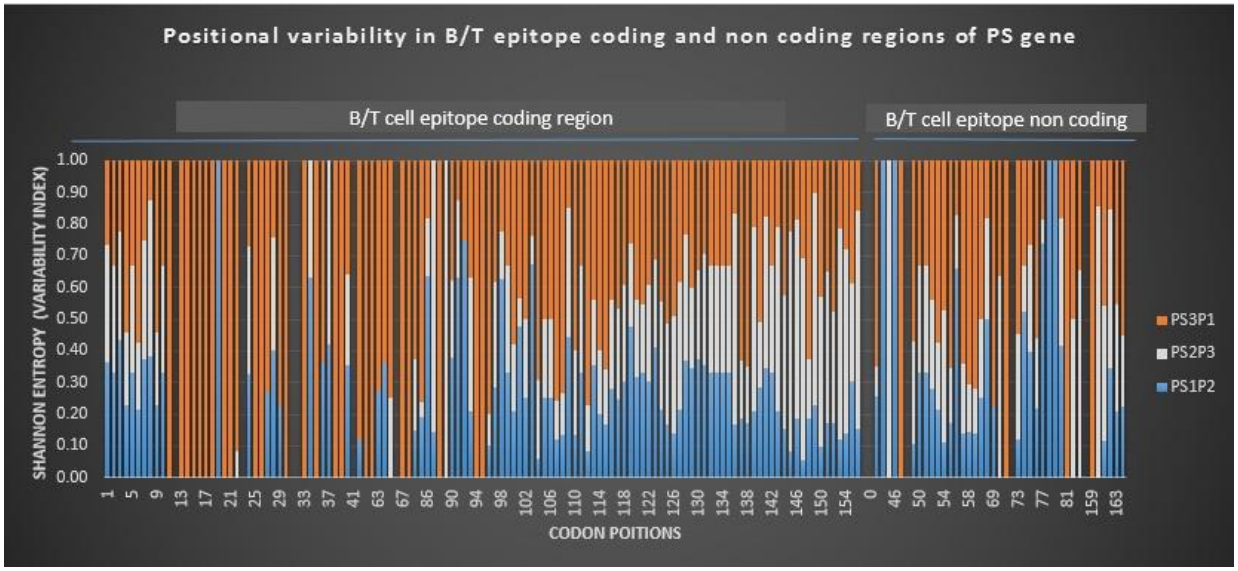

Figure 7. Positional variability within codon sites in the B/T epitope coding and non-coding regions of PS gene

Table 9. Descriptive statistics of Shannon entropy in S gene categorized under B/T epitope coding and non-coding regions												
B/T epitope	Δ in codon	N	Min	25% Percentile	Median	75% Percentile	Max	Mean	SD	SEM	Lower 95% CI	Upper 95% CI
coding	1	51	0	0	0.063	0.127	0.608	0.090	0.141	0.020	0.050	0.130
not coding	1	81	0	0	0.000	0.000	0.357	0.023	0.063	0.007	0.009	0.037
coding	2	51	0	0	0.063	0.151	0.602	0.100	0.135	0.019	0.062	0.138
not coding	2	81	0	0	0.000	0.086	0.441	0.060	0.091	0.010	0.040	0.080
coding	3	51	0	0	0.063	0.151	1.018	0.146	0.242	0.034	0.078	0.214
not coding	3	81	0	0	0.000	0.063	1.003	0.079	0.170	0.019	0.041	0.117
NOTE: B/T epitope- B or T cell epitope, Δ- position, N – number of observations, CI- confidence interval, SD- standard deviation from mean, SEM- standard error of mean, min- minimum, max- maximum.												

Table 10. Analysis of positional variability of nucleotides in B/T epitope coding and non-coding regions of S gene				
Δ in Codon	Mean rank of B/T epitope coding	Mean rank of B/T epitope non coding	Mann-Whitney U	P value[a]
1	82.53	56.41	1248	< 0.0001**
2	75.74	60.69	1595	0.0188*
3	76.98	59.9	1531	0.0052*
P value significant at α <0.05*, α<0.001**				

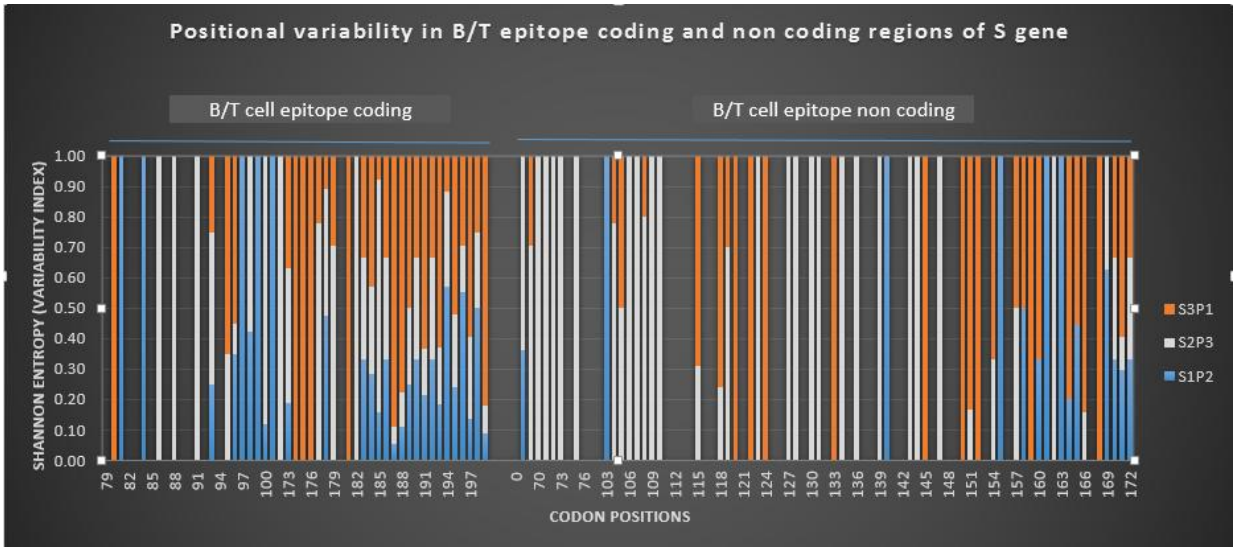

Fig. 8. Positional variability within codon sites in the B/T epitope coding and non-coding regions of S gene

Discussion

HBV is documented having the highest mutation rate among double stranded DNA viruses, but not the RNA viruses (Echevarria and Avellon, 2006; Osiowy et al., 2006). This is partly due to error prone HBV viral replication by HBV polymerase that lacks proof reading activity (Horvat, 2011). The dynamics of HBV evolution vary in different regions of its genome (Bouckaert et al., 2013; Datta et al., 2007). The mutation rate in the three nucleotides of a codon vary due to degeneracy in the genetic code(Yang et al., 1995). The HBV, being a compact virus with overlapping genes, adds constraint to the evolutionary sites of its codon. A syn or a non–syn mutation in one frame may cause a syn or a non-syn mutation in the overlapping region(Torres et al., 2013b). The difference in HBV genotypes according to geographic distribution allows for a difference in selection pressures, resulting in an independent evolution of the virus. This situation urged researchers to explore the factors that determine the evolution of HBV in a specific geographic region. There are hardly any studies in Saudi Arabia that have explored the variability and mutational landscape of different genomic regions of this virus. The identified HBV genotype D was the most prevalent in the patients included in this study, which is in agreement with a previous documented research (Kumar et al., 2011). There was greater variability in the PS region compared to the S region of the HBV genome. Several studies hypothesized that the greater variability in PS region could be due to its overlap with the functionally dispensable spacer region of HBV polymerase(Kim et al., 1999), while the S region overlaps the functionally indispensable RT domain of polymerase. Other factors that may determine the variability and mutability of a region could be the composition of the amino acids, the overlapping amino acid pairs, and the codon usage bias between regions and the overlapping genes(Torres et al., 2013a). Another factor that may determine the variability could be the immune selection pressure (Xu et al., 2013).Some roles of these factors were explored, while others need further investigations.

The variant analysis of the studied sequences revealed that there was an overall high variability (entropy) in PS region, both in magnitude and density, in comparison to variability (entropy) in the S region of HBV (Tables 1

and 2; Figs 2 and 3). Comparative analysis of Hx at each position has shown both positional variations in codons in the same region and regional variations when comparing codon position in PS region to its counterpart codon position in the S region.

In PS region the third position in the codon was found to have significantly high variability as against position 2 and 3 of a codon. However there was no significant difference found in variability of sites in codon in S region. When comparing variability of counterpart position in codon of PS and S regions, all sites were significantly more variable ($p<0.001$) in PS than those in S region .Majority of substitutions in PS (f=413/679) and S (f=161/285) were found to be of syn type (Table 4). A very high proportion of syn mutations in PS and S genes, resulted in a non-syn mutation in the overlapping spacer region (in PS-SsNp/SsNp+SsSp=385/413) and RT domains (in S SsNp/SsNp+SsSp =128/161) of polymerase (Table 4, Figure5 & 6). This seems to be more due to positional effect as most of the syn mutations occurred at position S3P1 of a codon (Figure5 & 6). The non-syn substitutions in PS seem to be not so well tolerated by the overlapping spacer region (in PS NsSp/NsNp+NsSp=73/266) as compared to toleration of non-syn mutations in S by the overlapping RT domain (in S NsSp/NsNp+NsSp 83/124) (Table 4). The estimation of transition/transversion bias revealed, that there is greater rate of transitions occurring than transversions at positions S2P3 and S3P1 in both PS and S regions. It is intriguing that there is a greater rate of transversions occurring at position S1P2 in both PS and S. Interestingly data on mutation frequencies at position S1P2 show almost all mutations falling under N_SN_P group (Figure5 & 6) which affect both overlapping proteins and therefore have a high chance of negative selection (Schmidt et al., 2008).

Compositional analysis of amino acids in different regions and their mutability considering the overlapping amino acid pairs has revealed an interesting patterns of distribution of mutation frequencies in different groups (N_SN_P, S_SS_P, N_SS_P and S_SN_P) and codon classes (H, M and L) at different codon positions (S1P2, S2P3 and S3P1) (Table 4 and Figure5 & 6) in the PS and S regions of HBV.

Most of the mutations in N_SN_P were encountered at position S1P2. This may be because positions 1 and 2 in codon are most critical and mutations in these positions will result in an amino acid change, which is seen affecting both the overlapping genes in both PS and S regions (spacer and RT domain of pol respectively). The pattern in the PS region that non-syn mutations in L class amino acid mostly affect the overlapping Lp class amino acid and the non-syn mutations in M class amino acid affect the overlapping Hp class amino acid and vice versa, shows a kind of homogeneity in the distribution of amino acid classes in PS and Pol ORF that result in higher NsNp type mutations in PS region as compared to S region. Therefore it can be said that most non-syn mutations in the PS region would result in a non-syn mutations in spacer region of pol (NSNp/NsNp + NsSp =171/228). It can also be said that most non-syn mutations in PS and S region occurring at position S1P2 (or PS1P2) would most probably also affect the P protein.

There were very few mutations in the group S_SS_P, where none of the overlapping genes resulted in amino acid change. Interestingly all the S_SS_P mutations were found at position S3P1. Therefore it can be said that a syn mutation in PS or S gene at position S3P1 (or PS3P1) would rarely not affect the P protein. Another interesting finding was that almost all mutations in S_SS_P group were tolerated by H_P class of codons at position 1 of pol. This supports the positional and compositional effect on the variability of the overlapping gene.

Groups wherein mutation affected only one of the overlapping gene are N_SS_P and S_SN_P. These type of mutations occurred variably in all classes of codons and were variably tolerated by all classes of codons. In the N_SS_P group where amino acid change occurs only in PS/S genes, mutations were seen at position S2P3 and rarely at position S1P2. Position 2 is critical in codon for coding amino acid and therefore affect the PS/S gene but not the P gene since a variable position 3 of codon in P is involved. The N_SS_P type of mutation are third most frequent in PS and second most frequently encountered in S region (Table 4&Figure 5). Therefore it seems more unlikely that a non-syn mutation in S region would affect P proteins. The SsNp group of mutation are syn mutations occurring in PS (SsNp/SsSp+ SsNp =368/408) and S (SsNp/SsSp+ SsNp =126/161) gene that only affect their respective overlapping spacer and RT regions in the pol protein. SsNp group of mutations are the most commonly encountered mutations in PS and S gene, majority of which occurring at codon position S3P1. SsNp mutations in PS region affects all classes of amino acids in the spacer region. However in the S region SsNp mutations mostly affect the Lp class in the RT domain. Therefore it can be said that majority of the syn mutations in the PS and S region affect the P gene and the most affected amino acid class in the RT region of Pol is the Lp class.

We also explored the role of codon usage in the observed differences in variability and mutability between PS, S and their respective overlapping spacer and RT regions. There was a codon usage bias found between overlapping PS ORF and spacer domain of pol ORF. Difference was also seen in codon usage in PS versus S genes and spacer domain versus RT domain of pol gene. On analysing the most favoured codon usage for each amino acid, it was observed that PS and pol gene favour codon usage which have G or C at third base (positions S3P1 and S2P3)meaning they are evolving towards more stable dsDNA. On the contrary GC rich sites have been shown to

have higher methylation of cytosine and hence higher mutation rates in humans and hence may contribute to higher mutability at S3P1 site in HBV (Mugal and Ellegren, 2011). On the other hand S gene favoured more codons having A or U (T) at their third base(positions S3P1 and S2P3). This may explain high rates of mutations at S3P1 in S_SN_P and mutations at S2P3 in N_SS_P(Gorlov et al., 2008). But the codon usage may mostly explains the mutability at the third codon positions and to some extent the first position in amino acids that are coded by six different codons (Palidwor et al., 2010).

To determine the role of immune selection pressure we compared the variability of B/T epitope coding regions with B/T epitope non coding regions in both PS and S genes. In the PS region most of the sequence length determines the epitope for the B/T cell as most of its length is exposed to extracellular side of the viral particle as compared to S protein that has several trans-membrane domains. In the PS region there was no significant difference in the variability at 369 nucleotide sites (coding for 123 amino acid) that coded for B/T epitope determining regions and 123 nucleotide sites that did not code for B/T epitope sites. This implies that variability in the PS region may not be under immune selection pressure but rather due to its amino acid overlap, compositional effect and to some extent due to codon usage bias between PS and S gene. In the S region however there was a significant higher variability ($p<0.05$) in the B/T epitope coding regions as compared to B/T epitope non coding regions. Therefore S region may be evolving under immune selection pressure as HBV antigenic 'a' determinant site is present in the major hydrophilic region (MHR) of the S protein. The MHR is frequently mutated and is thought to result in immune escape HBV types that may be responsible for occult HBV infections that are unnoticed by regular serology tests that increase the chances of HBV infection during blood transfusion.

To conclude, this work like earlier studies has found higher variability in the PS region than S region of HBV. This is probably the first study in Saudi Arabia that has discussed the compositional and mutational landscape of HBV PS and S genes among treatment naïve chronic hepatitis B patient. The key findings are that the majority of the synonymous mutations in the PS and S gene occur at position S3P1 cause a non-synonymous mutation in the overlapping Pol gene. Therefore it is important not to overlook syn mutations in overlapping genes. Hence there is a need to focus on the pathophysiological effects of not only non-syn but also syn mutation in PS and S regions on the pol spacer and RT domains. The non-syn mutation in the S region are comparatively well tolerated by P gene than non syn mutations in the PS gene. There is a higher transversion occurring at S1p2 codon site in both PS and S regions. The syn and non syn mutations in the overlapping PS/S and Pol gene show positional and compositional pattern. There is a codon usage bias between the PS and S regions wherein the PS region codons with G or C at the third codon site are favoured as against A or T in the S region. There is probably no immune selective pressure acting on the evolution of PS region as compared to the S region which seems to evolve under immune selective pressure.

Acknowledgement

This work was supported by the grant from King Abdulaziz city for science and technology (grant #-AT-34-212). We thank the KFMRC for allowing this work to be carried out in their labs. The contribution of authors are in the order of the names. This work was performed by the first author as the part of his PhD research work under the supervision of Professor Khalid Omer Abualnaja, Esam I Azhar and Elie K. Barbour. We thank Professor Taha A. Kumosani who helped in obtaining the research grant from KACST. Dr. Mai M. El-Daly helped in the collection of sample and questionnaire data. Dr. Mai also helped in the optimization of pcr conditions. Dr. Sherif helped in the procurement of reagents and in manuscript correction. We thank the participating clinicians Dr. Hisham O. Akbar, Dr. Hind B. Fallatah and Dr. Mohammed I. Dgdgi who allowed us access to patients and their clinical records.

Reference

B**ouckaert, R., Alvarado-Mora, M.V., and Pinho, J.R. (2013)**. Evolutionary rates and HBV: issues of rate estimation with Bayesian molecular methods. Antivir Ther *18*, 497-503.

Buti, M., Rodriguez-Frias, F., Jardi, R., and Esteban, R. (2005). Hepatitis B virus genome variability and disease progression: the impact of pre-core mutants and HBV genotypes. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology *34 Suppl 1*, S79-82.

Cento, V., Mirabelli, C., Dimonte, S., Salpini, R., Han, Y., Trimoulet, P., Bertoli, A., Micheli, V., Gubertini, G., and Cappiello, G. (2013). Overlapping structure of hepatitis B virus (HBV) genome and immune selection pressure are critical forces modulating HBV evolution. Journal of General Virology *94*, 143-149.

Chen, P., Gan, Y., Han, N., Fang, W., Li, J., Zhao, F., Hu, K., and Rayner, S. (2013). Computational Evolutionary Analysis of the Overlapped *Surface (S) and Polymerase (P)* Region in Hepatitis B Virus Indicates the Spacer Domain in P Is Crucial for Survival. PLoS ONE *8*, e60098.

Datta, S., Banerjee, A., Chandra, P.K., and Chakravarty, R. (2007). Selecting a Genetic Region for Molecular Analysis of Hepatitis B Virus Transmission. Journal of Clinical Microbiology *45*, 687-689.

Drexler, J.F., Geipel, A., Konig, A., Corman, V.M., van Riel, D., Leijten, L.M., Bremer, C.M., Rasche, A., Cottontail, V.M., Maganga, G.D., et al. (2013). Bats carry pathogenic hepadnaviruses antigenically related to hepatitis B virus and capable of infecting human hepatocytes. Proceedings of the National Academy of Sciences of the United States of America *110*, 16151-16156.

Echevarria, J.M., and Avellon, A. (2006). Hepatitis B virus genetic diversity. Journal of medical virology *78 Suppl 1*, S36-42.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research *32*, 1792-1797.

Gao, W., and Hu, J. (2007). Formation of Hepatitis B Virus Covalently Closed Circular DNA: Removal of Genome-Linked Protein. Journal of virology *81*, 6164-6174.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S.H. (2010). The NCBI BioSystems database. Nucleic Acids Res *38*, D492-496.

Gorlov, I.P., Gorlova, O.Y., and Amos, C.I. (2008). Relative effects of mutability and selection on single nucleotide polymorphisms in transcribed regions of the human genome. BMC Genomics *9*, 292-292.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In Nucl. Acids. Symposium, pp. 95-98.

Horvat, R.T. (2011). Diagnostic and Clinical Relevance of HBV Mutations. Lab Medicine *42*, 488-496.

Hu, J., and Ng, P. (2012). Predicting the effects of frameshifting indels. Genome Biology *13*, R9.

Hu, J., and Ng, P.C. (2013). SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. PLoS ONE *8*, e77940.

Ingman, M., Lindqvist, B., and Kidd-Ljunggren, K. (2006). Novel mutation in Hepatitis B virus preventing HBeAg production and resembling primate strains. Journal of General Virology *87*, 307-310.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics (Oxford, England) *28*, 1647-1649.

Kim, Y., Hong, Y.B., and Jung, G. (1999). Hepatitis B virus: DNA polymerase activity of deletion mutants. Biochemistry and molecular biology international *47*, 301-308.

Krakauer, D.C. (2000). Stability and evolution of overlapping genes. Evolution *54*, 731-739.

Kramvis, A., Arakawa, K., Yu, M.C., Nogueira, R., Stram, D.O., and Kew, M.C. (2008). Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. Journal of medical virology *80*, 27-46.

Kumar, K., Kumar, M., Rahaman, S.H., Singh, T.B., Patel, S.K., and Nath, G. (2011). Distribution of Hepatitis B virus genotypes among healthy blood donors in eastern part of North India. Asian Journal of Transfusion Science *5*, 144-149.

Mugal, C., and Ellegren, H. (2011). Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome Biology *12*, R58.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. PLoS ONE *7*, e37558.

Norder, H., Couroucé, A.-M., and Magnius, L.O. (1994). Complete Genomes, Phylogenetic Relatedness, and Structural Proteins of Six Strains of the Hepatitis B Virus, Four of Which Represent Two New Genotypes. In Virology, pp. 489-503.

Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R.I., Imai, M., Miyakawa, Y., and Mayumi, M. (1988). Typing Hepatitis B Virus by Homology in Nucleotide Sequence: Comparison of Surface Antigen Subtypes. Journal of General Virology *69*, 2575-2583.

Osiowy, C., Giles, E., Tanaka, Y., Mizokami, M., and Minuk, G.Y. (2006). Molecular Evolution of Hepatitis B Virus over 25 Years. Journal of virology *80*, 10307-10314.

Palidwor, G.A., Perkins, T.J., and Xia, X. (2010). A General Model of Codon Bias Due to GC Mutational Bias. PLoS ONE *5*, e13431.

Pollicino, T., Cacciola, I., Saffioti, F., and Raimondo, G. (2014). Hepatitis B virus PreS/S gene variants: Pathobiology and clinical implications. Journal of Hepatology *61*, 408-417.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., and Tatusova, T. (2004). A web-based genotyping resource for viral sequences. Nucleic Acids Research *32*, W654-W659.

Sanger, F. (1981). Determination of nucleotide sequences in DNA. Bioscience reports *1*, 3-18.

Schmidt, S., Gerasimova, A., Kondrashov, F.A., Adzuhbei, I.A., Kondrashov, A.S., and Sunyaev, S. (2008). Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection. PLoS Genetics *4*, e1000281.

Seeger, C., and Mason, W.S. (2000). Hepatitis B Virus Biology. Microbiology and Molecular Biology Reviews *64*, 51-68.

Sunbul, M. (2014). Hepatitis B virus genotypes: Global distribution and clinical importance. World Journal of Gastroenterology : WJG *20*, 5427-5434.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution.

Torres, C., Fernandez, M.D., Flichman, D.M., Campos, R.H., and Mbayed, V.A. (2013a). Influence of overlapping genes on the evolution of human hepatitis B virus. Virology *441*, 40-48.

Torres, C., Fernández, M.D.B., Flichman, D.M., Campos, R.H., and Mbayed, V.A. (2013b). Influence of overlapping genes on the evolution of human hepatitis B virus. Virology *441*, 40-48.

Xie, Y., Liu, S., Zhao, Y., Zhang, L., Zhao, Y., Liu, B., and Guo, Z. (2015). Precore/Core Region Mutations in Hepatitis B Virus DNA Predict Postoperative Survival in Hepatocellular Carcinoma. PLoS ONE *10*, e0133393.

Xu, Z., Wu, G., Li, F., Bai, J., Xing, W., Zhang, D., and Zeng, C. (2013). Positive selection signals of hepatitis B virus and their association with disease stages and viral genotypes. Infect Genet Evol *19*, 176-187.

Yang, Z., Lauder, I.J., and Lin, H.J. (1995). Molecular evolution of the hepatitis B virus genome. J Mol Evol *41*, 587-596.

Yu, H., Yuan, Q., Ge, S.-X., Wang, H.-Y., Zhang, Y.-L., Chen, Q.-R., Zhang, J., Chen, P.-J., and Xia, N.-S. (2010). Molecular and Phylogenetic Analyses Suggest an Additional Hepatitis B Virus Genotype "I". PLoS ONE *5*, e9297.

Zhang, Q., and Cao, G. (2011). Genotypes, mutations, and viral load of hepatitis B virus and the risk of hepatocellular carcinoma. Hepat Mon *11*, 86-91.

Zhong, J., Gao, Y.-q., Sun, X.-h., Zhu, X.-j., and Li, M. (2012). High prevalence of the B2+C2 subgenotype mixture in patients with chronic hepatitis B in Eastern China. Acta Pharmacol Sin *33*, 1271-1276.