Author/s:
Wagner, J;Yuen, L;Littlejohn, M;Sozzi, V;Jackson, K;Suri, V;Tan, S;Feierbach, B;Gaggar, A;Marcellin, P;Buti Ferret, M;Janssen, HLA;Gane, E;Chan, HLY;Colledge, D;Rosenberg, G;Bayliss, J;Howden, BP;Locarnini, SA;Wong, D;Thompson, AT;Revill, PA

Title:
Analysis of Hepatitis B Virus Haplotype Diversity Detects Striking Sequence Conservation Across Genotypes and Chronic Disease Phase

Date:
2021-05-01

Citation:
Wagner, J., Yuen, L., Littlejohn, M., Sozzi, V., Jackson, K., Suri, V., Tan, S., Feierbach, B., Gaggar, A., Marcellin, P., Buti Ferret, M., Janssen, H. L. A., Gane, E., Chan, H. L. Y., Colledge, D., Rosenberg, G., Bayliss, J., Howden, B. P., Locarnini, S. A. ,... Revill, P. A. (2021). Analysis of Hepatitis B Virus Haplotype Diversity Detects Striking Sequence Conservation Across Genotypes and Chronic Disease Phase. Hepatology, 73 (5), pp.1652-1670. https://doi.org/10.1002/hep.31516.

Article type : Original

Analysis of Hepatitis B virus haplotype diversity detects striking sequence conservation across genotypes and chronic disease phase

[1]Josef Wagner[#], [1]Lilly Yuen[#], [1]Margaret Littlejohn[#], [1]Vitina Sozzi, [1]Kathy Jackson, [2]Vithika Suri; [2]Susanna Tan, [2] Becket Feierbach, [2]Anuj Gaggar, [4]Patrick Marcellin, [5]Maria Buti Ferret, [6]Harry L. A. Janssen,[7]Ed Gane, [8]Henry L.Y. Chan, [1]Danni Colledge, [1]Gillian Rosenberg, [1]Julianne Bayliss, [3]Benjamin P Howden, [1]Stephen A. Locarnini, [1,9]Darren Wong, [9]Alexander T. Thompson and [1]Peter A. Revill*.

[1]Division of Molecular Research and Development, Victorian Infectious Diseases Reference Laboratory, Peter Doherty Institute for Infection and Immunity, Melbourne Healthy, University of Melbourne, Melbourne, Australia

[2]Gilead Sciences, Foster City, California, USA.

[3]Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, Melbourne at The Peter Doherty Institute for Infection and Immunity, Australia

[4]Hôpital Beaujon, University of Paris, Clichy, France

[5]Liver Unit, Valle d'Hebron University Hospital, Ciberehd del Insituto Carlos III Barcelona, Spain

[6]Toronto Center for Liver Diseases, Toronto General Hospital, University Health Network, University of Toronto, Canada

[7]New Zealand Liver Transplant Unit, Auckland City Hospital, Auckland, New Zealand

[8]Department of Medicine and Therapeutics, The Chinese University of Hong Kong

[9]Department of Gastroenterology, St. Vincent's Hospital, Melbourne, Victoria,

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/HEP.31516

Australia

[#] Equal contribution.

*** Corresponding author:**

Professor Peter Revill

The Peter Doherty Institute for Infection and Immunity

792 Elizabeth St, Melbourne VIC 3000

W: +61 (0) 3 9342 9604

Peter.Revill@vidrl.org.au

Keywords:

Haplotype reconstruction, Chronic Hepatitis B virus infection, genetic distance

Abstract

Background

We conducted haplotype analysis of complete hepatitis B virus (HBV) genomes following deep sequencing from 368 patients across multiple phases of chronic Hepatitis B (CHB) infection from 4 major genotypes (A to D), analysing 4110 haplotypes to identify viral variants associated with treatment outcome and disease progression.

Results:

Between 18.2% and 41.8% of nucleotides and between 5.9% and 34.3% of amino acids were 100% conserved in all genotypes and phases examined, depending on the region analysed. HBeAg loss by week 192 was associated with different haplotype populations at baseline. Haplotype populations differed across the HBV genome and CHB history, this being most pronounced in the precore/core gene. Mean number of haplotypes (frequency) per patient

was higher in immune active HBeAg-positive chronic hepatitis Phase II (11.8) and HBeAg-negative chronic hepatitis Phase IV (16.2) compared to subjects in the "immune tolerant" HBeAg-positive chronic infection Phase I (4.3, $p<0.0001$). Haplotype frequency was lowest in genotype B (6.2, $p<0.0001$) compared to the other genotypes (A = 11.8, C = 11.8, D = 13.6). Haplotype genetic diversity increased over the course of CHB history, being lowest in Phase I, increasing in Phase II and was highest in Phase IV in all genotypes except genotype C. HBeAg loss by week 192 of tenofovir therapy was associated with different haplotype populations at baseline.

Conclusion:

Despite a previously unrecognised degree of HBV haplotype diversity and heterogeneity across the phases of CHB natural history, novel highly conserved sequences in key genes and regulatory regions were identified in multiple HBV genotypes that should be further investigated as targets for new antiviral therapies and predictors of treatment response.

The hepatitis B virus (HBV) infecting patients with chronic hepatitis B virus disease (CHB) is highly diverse, even within an individual patient. Different HBV variants present within the same patients are known as virus haplotypes. We have shown that haplotype diversity increases across the natural history of chronic hepatitis B (CHB) virus infection and that haplotype diversity differs across ethnicities and genotypes. Strikingly, different haplotype populations were observed in patients that subsequently lost hepatitis e antigen (HBeAg) compared to patients who remained HBeAg positive. We also identified highly conserved sequences in the coding and regulatory regions of the HBV genome and across multiple phases of CHB natural history and major genotypes which may represent suitable targets for novel pan-genotypic antiviral therapies effective.

INTRODUCTION

HBV exists as 9 different genotypes, with genotypes A to D, the most common worldwide (1,2), which are associated with differences in chronic hepatitis B (CHB) natural history, disease progression and treatment response. CHB is itself highly complex, and in persons infected at birth or in early childhood, commences with a high viral replication phase with little evidence of liver damage [HBeAg-positive chronic infection-Phase I, (3) also known as the immune tolerant phase (4)], followed by a period of immune activation associated with inflammatory activity in the liver [HBeAg-positive chronic hepatitis-Phase II, (3) also known as the HBeAg-positive immune-active phase (4)]. Following HBeAg seroconversion, a period of seeming inactivity is characterised by low viral replication and little evidence of liver damage [HBeAg-negative chronic infection-Phase III, (3) also known as the inactive CHB phase (4)]. A fourth phase emerges characterised by, increased viral replication due to emergence of HBeAg negative HBV variants (G1896A PC variants) and reactivation of hepatic necro-inflammation [HBeAg-negative chronic hepatitis-Phase IV,(3) also known as the HBeAg-negative immune reactivation phase (4)]. In a small proportion of patients, this progresses to a fifth phase associated with hepatitis B surface antigen (HBsAg) loss and or seroclearance, which is analogous to HBV functional cure (5); the primary aim of many current efforts to cure chronic HBV infection (6).

One of the mechanisms by which viruses such as HIV, influenza A and HBV evade the host immune system is through the generation of haplotypes (individual sequences) that alter their antigenic profile (7–9). Historically, viral haplotypes were investigated by conventional cloning and Sanger sequencing which limited analysis to a small number of samples due to the laborious nature of this approach (10). Next generation sequencing and viral haplotype reconstruction programmes now facilitate deeper characterization of the viral haplotype repertoire. However reconstructing haplotypes from short read sequencing data has been challenging. This limitation has been recently overcome through the development of haplotype reconstruction software (11–15) permitting haplotype construction for a wide variety of viruses, such as HIV (16), HCV (17) and influenza viruses (18,19). Haplotype analysis has identified immune escape mutations (16), cell-specific infectivity (17), and transmission of haplotypes between individuals (18,19).

Haplotype analysis performed for HBV to date has been limited to small studies of the HBx gene (20,21). González *et al.* identified two hyper-conserved nucleotide regions in the HBx

promoter and in the 5' coding region (20) and Mei *et al.* used haplotype analysis to identify high-risk mutations in HBx associated with hepatocellular carcinoma (21). There have been no large-scale haplotype studies of the complete HBV genome across multiple genotypes and phases of CHB. The aim of this study was to analyse HBV haplotype frequency and diversity over the complete HBV genome, including all coding and regulatory regions, to identify hyper-conserved regions across three phases (I, II, and IV) of CHB for all major genotypes (A to D).

PATIENTS AND METHODS

Patient Cohorts

This study utilised baseline samples from patients enrolled in Gilead trials "G101" GS-US-203-0101 (HBeAg-positive chronic infection; Phase I; n =96), "G103" GS-US-174-0103 (HBeAg-positive chronic hepatitis; Phase II; n = 159) and "G102" GS-US-174-0102 (HBeAg-negative chronic hepatitis; Phase IV; n = 113) (22,23). Patients were not available from Phase 3 (HBeAg- negative chronic infection).

Clinical trial number:

See following Clinical Trials Gov Identifier for the three Gilead trials.

ClinicalTrials.gov Identifier: NCT00507507 for Gilead trial GS-US-203-0101,
https://clinicaltrials.gov/ct2/show/record/NCT00507507

ClinicalTrials.gov Identifier: NCT00116805 for Gilead trial GS-US-174-0103
https://clinicaltrials.gov/ct2/show/record/NCT00116805

ClinicalTrials.gov Identifier: NCT00117676 for Gilead trial GS-US-174-0102
https://clinicaltrials.gov/ct2/show/record/NCT00117676

Throughout the manuscript we refer to Phase I (PH I), Phase II (PH II), and Phase IV (PH IV). Approximately 50% of patients in Phase II lost HBeAg over the course of 192 weeks of tenofovir treatment (24). All patients signed an informed consent form prior to screening and in accordance with local regulatory and ethics committee requirements. Experimental protocol in these trials was approved by Gilead Sciences and all local regulatory agencies. See following Clinical Trials Gov Identifier for the three Gilead trials, NCT00507507 for

Gilead trial GS-US-203-0101, NCT00116805 for Gilead trial GS-US-174-0103, and NCT00117676 for Gilead trial GS-US-174-0102

Patients in the phase I were limited to HBV genotypes B and C, whereas patients in phase II and IV included samples from subjects infected with genotypes A, B, C ~~to~~ and D.

Methods

DNA extraction

DNA was extracted from patient serum and full-length HBV genomes sequenced using Illumina MiSeq as previously described (24).

Sequence quality filtering, sequence mapping and haplotype reconstructions

High-quality filtered reads were used for alignment against NCBI reference genomes for genotype A, B, C, and D, respectively, using SMALT (version 0.7.4). GenBank accession numbers of the genotype A, B, C, and D references were X02763, D00330, AB033556, and X02496, respectively. SMALT generated SAM files were used for haplotype reconstruction using CliqueSNV program (version 1.4.8). Ten datasets were generated in total, two for Phase I (genotypes B and C), and four each of Phase II and IV (genotypes A to D). Only haplotypes present at an abundance of greater than 1% were included in the analysis.

Haplotype analysis

4110 correct aligned haplotypes were used for downstream analysis. The proportional abundance for each haplotype per sample were visualised in a stacked bar chart. Phylogenetic analysis of aligned haplotype sequences was done by creating a maximum likelihood tree and visualised using iTOL software. All functional domains were analysed (Pol TP, Pol SP, Pol RT, Pol Rnase H, Pres, Pres2, HBsAg, X gene and precore and Core gene) and all regulatory regions (PreS promotor, S promotor, X enhancer, and Upstream Regulatory Region (URR) (**Supplementary Fig. 1**). Nucleotide and amino acid residues that were 100% conserved across the different phases of CHB for HBV genotypes A to D were calculated in EXCEL and presented as a stacked column chart. The number of haplotypes from each sample were correlated with four serological markers; alanine transaminase (ALT), HBV DNA (viral load), HBsAg, and HBeAg levels. The correlations were plotted as a scatter plot including a linear trend line. The association between number of haplotypes and serological markers was statistically assessed using linear regression for trend line and correlation analysis using the Spearman method. The genetic divergence of haplotypes within each

group was described by estimation of average evolutionary divergence over sequence pairs within groups.

Multivariate analysis of haplotypes using Principal Coordinate Analysis (PCoA)

PCoA was used to investigate whether different haplotype populations could be visually identified within genotype and sub-genotype clustering. The genotype and sub-genotype clusters are surrounded by an ordi-hull which places a line around the clusters. Within each cluster, we colored the haplotypes according to natural history of CHB. Constrained Correspondence Analysis (CCA) was used to investigate whether HBeAg loss by week 192 was associated with different haplotype populations at baseline.

Statistical analysis

All statistical tests were conducted in Prism 7 for Mac OS X (version 7.0e) except the PERMANOVA test where PAST3 (25) was utilised.

For further details regarding the methods, please refer to the "supplementary online methods and analysis".

RESULTS

Patient demographics

The breakdown of patients into the three different CHB phases and genotypes are shown in Table 1.

Haplotype frequency differed by HBV genotype and ethnic background

We performed deep coverage sequencing of the whole HBV genome isolated from baseline samples of 96 Phase I, 159 Phase II, and 113 Phase IV patients. The mean coverage ranged between ~ 17,000 and ~ 23,000 - fold (**Supplementary Table 1**). A total of 13,469 full-length HBV haplotypes were constructed from 368 patients and 4110 haplotypes with a minimum abundance of 1% were aligned to the genotypes A, B, C, and D reference genomes. Numbers of samples and haplotypes from the different disease phases, separated by genotype are shown in Supplementary Table 2. The number of haplotypes per sample was lowest in the Phase I group with 49 % of the samples (47/96) having ≤2 haplotypes (**Fig 1.a)**. In the Phase II group, 11% of samples (18/159) had ≤2 haplotypes (**Fig 1.b**) and in the Phase IV group only 1.8% of samples (2/113) had ≤2 haplotypes (**Fig 1.c**). An ethnic and/or genotype difference was identified in the Phase II group. In the genotype B and C HBV from

Asian ethnicities none of the samples had ≤2 haplotypes (**Fig 1.b**). Whereas, in the genotype A and D HBV from Caucasian ethnicities 22% (18/83) of the samples had ≤2 haplotypes (**Fig 1.b**).

Haplotype frequency per patient was statistically higher in the Phase II group ($P < 0.0001$) and Phase IV group ($P < 0.0001$) compared to the Phase I group and between the phase II and phase IV groups ($P < 0.0001$) (**Fig 2.a).** Between genotype analysis revealed that the haplotype frequencies for genotypes A, C and D, were statistically higher compare to genotype B (0.0005, < 0.0001, < 0.0001, respectively) (**Fig 2.b).** We further tested whether the significant differences seen between the Phase II and Phase IV group (**Fig 2.a**) were a global phenomenon or influenced by ethnicity. Indeed, the difference seen between the Phase II and Phase IV group was only seen in the Caucasian population ($P < 0.0001$) (**Fig 2.c**). In the Asian population the haplotype frequency was not statistically different between the Phase II and Phase IV group (**Fig 2.d**). The comparison of Asian and Caucasian haplotype frequencies within the Phase II and Phase IV groups, revealed a significantly lower haplotype frequency in the Caucasian population compare to the Asian population in the Phase II group ($P < 0.0001$) (**Fig 2.e**). We also detected significantly lower haplotype frequencies in genotype B samples compare to genotype C samples in the Asian Phase I group ($P = 0.0003$) (**Fig 2.f**).

Phylogenetic analysis confirmed the haplotype species inferred for each of the samples belonged to the same HBV genotype, and no mixed genotype infections were observed in a given sample, indicated by all sequences clustering with the expected sub genotype in the maximum likelihood tree (**Supplementary Figs 2.a to 2. c).**

Genetic divergence varied over the complete HBV genome

Analysis of genetic divergence (nucleotide diversity) over the complete HBV genome, and the individual functional genes and regulatory regions (**Fig 3.a to 3.d**) showed a statistically significant increase over the course of CHB from Phase I, to Phase II and Phase IV for genotypes A, B, and D **(Supplementary Table 3**). This was not seen for genotype C however, where a decrease in genetic divergence was observed over the course of CHB, except for the PC and Core gene, although this was not statistically significant.

The increased genetic distance observed in the POL rt domain compared to the HBsAg domain, despite a large overlap between these two domains, was expected because the additional 300bp of the POL RT domain is highly diverse. Thus, we also truncated the POL RT domain to a 700bp fragment size to enable direct comparison with the overlapping HBsAg domain and observed near identical results for each domain (Supplementary Table 5).

Haplotype nucleotide conservation is different between phases of CHB and genotype

Next, we determined if haplotype analysis could identify regions of the HBV genome that were highly conserved over the three main phases of CHB. Strikingly, over the complete HBV genome, 32.4% of all nucleotides were 100% conserved across all 4110 aligned haplotypes from 368 patients (**Table 2).**

As observed for haplotype frequencies (**Fig 1**), differences were observed in the abundance of conserved nucleotides in haplotypes across genotypes and phases of CHB. For genotypes A and D, the proportion of conserved nucleotides was lowest in Phase IV compared to Phase II (**Table 2**), whereas, for genotypes B and C, nucleotide conservation was lowest in Phase II (**Table 2**). Overall, nucleotide conservation was lower in the regulatory regions ranging from 15.2% to 38.3 % for the S promoter and X enhancer, respectively (**Table 2)** compared to between 18.2% to 41.8% for the PreS2 gene and Pol Rnase H gene.

Haplotype amino acid conservation across individual functional genes

At the amino acid (aa) level (**Table 2**, second column, value in brackets), 100% aa conservation in all 4110 haplotypes varied between 5.9% and 34.3% of residues, depending on the region analysed. Highest aa conservation was identified in the Pol RNase domain and in the Pol RT domain, with 37.3% and 34.3% of these two functional domains being 100% conserved across all 4110 haplotypes respectively. In contrast, lowest aa conservation was identified in the Pol Sp domain and the HBsAg gene, with 5.9% and 9.7% of these two functional regions being 100% conserved across all 4110 haplotypes respectively.

Polymerase gene

For the polymerase gene, as expected there was a ~~As expected, there was a~~ high degree of sequence conservation across the known functional domains in the Pol RT region, with less conservation in the A/B interdomain region (**Fig 4.a**). However, additional areas with high conservation were present outside these known active domains, e.g. aa 17 - 35 and aa 57-74 (**Fig 4.a).**

HBsAg gene

In contrast, for the HBsAg gene (**Fig 4.b**), fewer areas were highly conserved. Low levels of conservation were identified in the major immune-dominant regions. The known N-linked glycosylation site at sN146 was conserved and loop 2 within the immunodominant region encoded a number of conserved residues. In contrast, loop 1 in this region including mini loop aa 120-124, did not encode any 100% conserved residues. In addition, no cysteine residues, including those involved in disulphide bond formation were 100% conserved across all genotypes and disease phases, although most were conserved at 99.9% alignment threshold (data not shown).

X gene

For the HBX gene, 100% aa sequence conservation was highest in regions known to be essential for HBV replication, including the Direct Repeat 1 (DR1) which overlaps the X gene at aa151-154 and the PC mRNA start site at aa 139-142, which also overlaps the glucocorticoid response element (GRE) (**Fig 4.b)**. Some residues were conserved in the known DDB1 binding site (aa 88-100) (26). Of interest, the complete aa sequence conservation observed for the DR1 overlap was not evident for the DR2 region (X residues aa 73-76). To elucidate this further, a detailed analysis specifically for the DR2 and DR1 region was conducted at the NT and aa level (**Supplementary Table 11**). At the NT level four NTs were not 100% conserved at 100% stringency across all 4110 haplotypes in the DR2 region. However, when the stringency was decreased to 98% (i.e. 98% of all 4110 haplotypes must have the same NT at a given position) then we observed 100% conservation. The same was observed at the aa level. 100% aa conservation for the DR2 region was not observed at 100% or 99% stringency but at 98% stringency.

PC and Core genes

For the precore and core genes, 100% aa sequence conservation was highest in the N and C terminal regions (**Fig 4.b**). As expected, 100% conservation was seen at the overlapping DR1 site (PC aa 4-7), across most of the epsilon region, an RNA element essential for viral replication, and in the polyA signal region (core aa 6-7). The cysteine residue at core 61, involved in di-sulphide bond dimerization of the core protein was 100% conserved, however other cysteine residues involved in di-sulphide bond formation (core aa 48, 107, 183) (27,28) were not 100% conserved. A number of residues across amino acids 120-143 known to form crucial interactions that stabilize capsid structure (29) were conserved, as were residues important in pgRNA encapsidation (aa 162-164) and second strand DNA synthesis (aa171-172).

Conserved nucleotide composition across regulatory regions

Conservation was high at the known binding sites for transcription factors in the liver specific PreS promoter, however was not apparent in the non-liver specific S promoter (**Figure 4.c**). The HBx promoter/Enhancer I sequence was highly conserved. This region overlaps with the polymerase gene (end of RT region and RNase H region). Sequence conservation did not appear to be clustered at known transcription factor binding sites. The Enhancer II/ core promoter showed high conservation for the region of the BCP involved in the precore mRNA start site and the GRE motif, however the rest of the BCP showed very little conservation, possibly due to the inclusion of samples from across both HBeAg positive and negative phases of disease. In contrast, the core upstream regulatory sequence showed much higher nucleotide conservation (Figure 4.c).

Haplotype frequency correlated with clinical serological markers

The patient cohorts in this study accurately represented the different disease phases as determined by their CHB clinical markers. Viral DNA level was highest in patients from Phase I (mean 8.4 log10 IU/ml) compare to Phase II (8.02 log10 IU/ml) and Phase IV (6.39 log 10 IU/ml) with statistically significant differences between the three Phases, **Table 1**. HBeAg titre was also highest in Phase I patients compared to Phase II patients (mean 3.45 log 10 PE IU/ml versus 2.88 log10 PE IU/ml) which was also statistically significant. All Phase IV patients were negative for HBeAg, **Table 1**. HBsAg titre was highest in Phase I patients (mean 4.72 log 10 IU/ml) compared to Phase II patients (mean 4.44 log 10 IU/ml) and Phase

IV patients (mean 3.75 log 10 IU/ml) with statistically significant differences between the three Phases, **Table 1**. The ALT enzyme level was lowest in Phase I patients (mean 26.61 IU/L) and highest in Phase II patients (mean 163 IU/L). Phase IV patients had a slightly lower ALT enzyme compare to Phase II patients (mean 150.32 IU/L). The ALT enzyme level was statistically significantly different between Phase I & II and between Phase I & IV, **Table 1**. Fibrosis score was very similar between Phase II (mean 3.31) and Phase IV (mean 3.21) patients and was not statistically significantly different.

Correlation analysis between haplotype frequencies and clinical markers revealed that haplotype frequencies were negatively correlated (i.e. lower haplotype frequencies) with higher HBsAg titre, HBeAg titre, and HBV viral load, although this varied by HBV genotype and phases of CHB (**Supplementary Table 4, Supplementary Fig 3 –5).** A negative correlation for HBsAg titre (**Supplementary Fig 3**) was observed in Phase I for genotypes B and C (Spearman P = 0.0429 and 0.0082) and in Phase II for genotypes A, C, and D (Spearman P = <0.0001, 0.0017, and 0.0002). In Phase IV a negative correlation was observed only for genotype D (Spearman P = 0.0247). For HBeAg titre, a negative correlation (**Supplementary Fig 4**) was observed in Phase II for genotypes A, C, and D (Spearman P = < 0.0001, 0.0097 and < 0.0001), whereas in Phase I, the correlation only reached statistical significance for genotype C samples (Spearman P = 0.0182). For HBV DNA, a negative correlation (**Supplementary Fig 5**) was observed in Phase I for genotype C samples (Spearman P = 0.00242) and in Phase II for genotype A and C (Spearman P = 0.01 and 0.0056). In contrast, for ALT, we observed a negative correlation in Phase IV for genotype A, (Spearman value = 0.0388) (**Supplementary Fig 6).** Between genotype analysis using PERMANOVA test showed that differences between the different phases of CHB and within each genotype were statistically significant (**Supplementary Fig 3 – 6,** P value line, above the different phases of CHB for each genotype).

We further investigated if the clinical markers were independently associated with the three different disease phases after adjusting for confounders. Linear regression analysis using all three Phases (I, II, and IV) as Y dependent variables and haplotypes numbers & all four clinical variables (viral load, HBsAg titre, HBeAg titre, and ALT enzyme level) as X independent variables showed that viral load was a confounder for HBsAg titre and HBeAg

was a confounder for viral load, HBsAg, and number of haplotypes but not for ALT enzyme. However, the number of haplotypes, HBsAg, or ALT enzyme alone or in combination with each other were independently statistically associated with the three different disease phases after adjusting or controlling for each of the variables.

Haplotypes clustered at the genotype level

PCoA enabled haplotype grouping based on sequence similarities with differences plotted into two-dimensional space. First, we grouped haplotypes by genotype which showed that for the complete HBV genome and functional genes, haplotypes separated into distinct groups based on genotype as expected (**Supplementary Fig 7.a – 7.i**), except for the precore and core gene **(Supplementary Fig 7.j).** The genotype separation diminished for some regulatory regions, particularly the URR regions **(Supplementary Fig 8.d – 8.f),** indicating that haplotype sequence was similar for the negative regulatory element, the core upstream regulatory sequence, and for the basal core promotor. It was not possible to distinguish clear separation of haplotypes across the different phases of CHB within each genotype cluster. Thus, we conducted further PCoA analysis by separating the datasets into sub-genotypes.

Haplotypes clustered according to CHB phase at the sub genotype level

PCoA at the sub genotype level revealed that haplotypes generally separated according to the phases of CHB. This was most apparent for the pre-core / core gene particularly for genotype A2, B2 and B4, C1 and C2 and genotype D1 and D2 samples (**Fig 5.a**). In all cases Phase IV haplotypes clustered furthest from the other two phases of CHB analysed.

For the HBx gene, strong separation of haplotypes was observed between Phase II and Phase IV haplotypes for genotype A2 and D3 (**Fig 5.b**), between Phase I and Phase II haplotypes for genotype C3 (**Fig 5.b**), and between the Phase IV haplotypes from the other two phases for genotype B2 (**Fig 5.b**).

Haplotype separation was less dominant for the Pol domains, terminal protein, spacer, reverse transcriptase**,** and RNase H gene **(Supplementary Fig 9.a to 9.d).** The same was observed for surface genes (PreS1, PreS2, and HBsAg), (**Supplementary Fig 10.a to 10.c**), though some distinction of CHB phases was detected in A2 genotype and in some of the C

and D genotype samples.

HBeAg loss by week 192 was associated with different haplotype populations at baseline

HBeAg loss was only detected in patients from CHB Phase II thus, only haplotypes from Phase II were used for HBeAg loss analysis. Constrained Correspondence Analysis showed that ~~significantly~~ different haplotype populations were detected at baseline for patients who lost HBeAg compare to those that remained HBeAg positive by week 192 of antiviral treatment, across multiple genotypes and subgenotypes (**Fig 6.a to 6.e**). Permutational Analysis Of Variance (PERMANOA) analysis showed that the haplotype populations differ significantly between patients at baseline who lost HBeAg compare to those that remained HBeAg positive at week 192. This was true for all genotypes from which HBeAg data were available (**Fig 6.f**).

DISCUSSION

This is the first in depth haplotype analysis of HBV genome length sequences across four major genotypes spanning multiple phases of CHB. Haplotype frequencies per sample varied by HBV genotype (A, C and D > B), and CHB Phase, which differed by ethnic background (Caucasian Phase II < Phase IV, but Asian Phase II = Phase IV), reflecting previous findings using cloning techniques (10). Genetic divergence of haplotype sequences increased across the course of CHB for genotypes A, B, and D. In contrast, this was not observed for genotype C, except in the PC and Core gene. A high level of amino acid conservation was detected in functional genes with some regions being more conserved then others. Part of this sequence conservation may be explained by the constraints imposed on the virus by the presence of multiple overlapping reading frames for the functional genes, as well as essential regulatory regions that also overlap the functional genes. However, the high level of amino acid conservation has identified novel regions, some of which are of unknown function, that should be further investigated as potential novel targets for HBV specific antiviral therapies. Current direct acting antivirals for HBV only target the polymerase, and there is an urgent need to develop an amoury of antivirals targeting different aspects of the HBV "life-cycle". Approaches targeting viral entry, nucleocapsid formation and stability, cccDNA expression and HBsAg production/egress are in development, with many in preclinical or early phase clinical trial (30,31). It will be important that these and other

antiviral agents are pan-genotypic, and effective across multiple phases of natural history. The sequence analysis we have performed should provide a useul tool to assist development of such approaches, for the major genotypes A to D. Ideally these studies should be extended to include HBV genotype E, which is common in Africa where HBV is highly endemic (2).

Clustering of haplotypes by phases of CHB was most dominant for the PC and Core gene, across multiple sub genotypes. In addition, different haplotype populations were identified at baseline in patients from Phase 2, dependent on HBeAg status at week 192. This suggests it may be possible to predict patients who will achieve HBeAg loss on therapy. Our studies are currently being extended to determine if we can also identify haplotype populations predictive of HBsAg loss (functional cure).

The observed increasing number of haplotypes per sample through disease progression was influenced by HBV genotypes and patient ethnicity with significant differences between Phase II and Phase IV patients in the Caucasian population (genotypes A and D) but not in the Asian population (genotypes B and C).

Despite the high level of sequence diversity observed, highly conserved nucleotide sequences and amino acid residues were present in all HBV genotypes and across CHB Phases I, II, and IV. At the nucleotide level, this was particularly evident in the Enhancer I/HBx promoter, and highly conserved nucleotides were also observed in the PreS promoter and the URR. We have previously shown that sequence differences in the URR contribute to differences in HBV replication observed across genotypes (32,33), however the high level of sequence conservation observed in 368 patients was unexpected, due to (i) the intrinsically high error rate in the HBV replication cycle, (ii) known sequence variability across the different HBV genotypes, and (iii) the different immune pressures associated with different phases of CHB natural history (e.g. “Phase I” versus “Phase IV”). However we were surprised to observe that not all sequences in key HBV regulatory regions were 100% conserved. This was particularly evident for DR2, which is critical for positive strand DNA synthesis and formation of RC DNA (34). Complementarity between the HBV RNA primer and DR2 sequence is required for efficient initiation of positive sense DNA synthesis at DR2, although some mismatch can be tolerated (35). Failure to initiate priming at DR2 leads to preferential formation of duplex linear DNA (DL DNA) instead of RC DNA. Since DL DNA is the template

for HBV DNA integration, it remains to be determined if some of the haplotypes identified in our study favour DL DNA formation and integration.

At the amino acid level, highly conserved regions were observed in the HBx gene, Pol RT, HBsAg and the precore and core genes. Some of these findings were expected, due to overlap with important regulatory regions, such as amino acids 151-154 in the HBx gene and amino acids 4-7 in the pre-core gene which overlap with the DR1, a key nucleotide sequence motif required for primer translocation and plus and minus strand DNA synthesis during HBV replication (36). However conserved residues were also observed at positions in HBx which do not overlap with other reading frames or known regulatory regions. Sequence conservation observed in 10 amino acid residues between amino acid positions 1 and 73 in the HBx gene may be due to the constraints resulting from the overlap with the RNAse H region in Pol. However conserved residues were also observed at positions 89, 93, 100, 104, 108, 139, 140, and 142 in HBx, which do not overlap with any HBV reading frame or known regulatory region. Three of these residues (89, 93, 100) are located within a known HBx-DNA Damage Binding 1 (DDB1) binding element, critical for regulation of HBV cccDNA transcription through degradation of the structural maintenance of chromosomes (SMC) 5/6 complex (37,38).

Of particular interest are the number of residues and regions that are highly conserved, but have not previously been described as essential for viral processes, for example the amino acid region 48-51 in the precore/core gene, amino acids 53-59 in the X gene, and the regions in polymerase gene located outside of the known functional domains. To compare the level of sequence conservation we identified with other haplotype studies is difficult since the only HBV haplotype studies reported to date were for the HBx gene (20,21). One study reported a maximum NT conservation between 70% and 100% depending on the region of HBx gene analysed (20). However, this study did not use a haplotype specific program to generate haplotypes, rather haplotypes were generated using individual short sequences and further haplotype studies using complete HBV genomes are required.

CCA analysis at the sub genotype level suggests evolution of different viral populations across Phases I, II, and IV of CHB. This was particularly evident between the HBeAg positive

(phase I and phase II) and HBeAg negative (phase IV) phases. Further examination of the HBeAg positive patients (Phase II) showed different haplotype populations existed at baseline in those patients who lost HBeAg by W192 on treatment compared to those who remained HBeAg positive. This is an important finding as HBeAg loss is a key intermediate step towards HBsAg loss and functional cure and is itself an accepted treatment endpoint. Further studies are required to tease out the actual sequence differences between the HBV sequences from these two distinct groups. This may lead to an ability to predict those patients likely to undergo HBeAg or HBsAg loss (functional cure) on antiviral therapy.

Negative correlations were observed between haplotype frequency and HBV markers (HBsAg, HBeAg, serum HBV DNA), and was dependent on HBV genotype. Correlations also varied by the phase of CHB. For example, for HBV genotype C, HBsAg titre, HBeAg titre and HBV viral load were all negatively associated with haplotype frequency in phase I and phase II, but not with phase IV. Genotype D was the only genotype to show a negative correlation between haplotype frequency and HBsAg titre, in the HBeAg-negative Phase IV. This suggests that haplotype frequency is not a useful surrogate for HBsAg and HBeAg titre, or serum HBV DNA.

There are some limitations with our study. First, our data was not continuous for individual patients as they progressed through the different phases of CHB (only baseline samples were available, prior to commencement of treatment). Therefore, we cannot definitively conclude whether HBV haplotypes arise during the course of the disease or whether their appearance may be influenced by other factors. Our findings were also limited by the unavailability of samples for genotypes A and D from Phase I, or from patients with HBeAg-negative HBV infection (Phase III). Future studies using longitudinal on-treatment samples from individual patients are planned, although will be limited by the high level of viral suppression observed on therapy which restricts the ability to perform deep sequencing of complete HBV genomes.

CONCLUSION

We have carried out the first in depth study of whole genome HBV haplotype frequency and diversity across the three main Phases (I, II, and IV) of CHB, for the major HBV genotypes A-

D, using bioinformatics to examine the relationship of the viral quasispecies to clinical markers. We observed a striking level of sequence conservation in HBV genes and regulatory regions in genotypes A to D, and across multiple phases of CHB natural history, allowing identification of conserved regions that may represent novel pan-genotypic antiviral targets.

Our study showed that most of the clinical markers were negatively associated with haplotype numbers. We also observed differences in haplotype poplulations in patients who achieved HBeAg serconversion on therapy, compared to patients who did not achieve this important treatment outcome. Together, our study shows that deep haplotype analysis has identified conserved "hotspots" in the HBV genome that should be investigated for their impact on HBV replication and their potential as novel targets for direct acting antiviral therapies.

Sequence Data availability

This study utilised three large clinical sample cohorts from Gilead Sciences (Foster City, California, USA) and Gilead Sciences is unable to make the sequence data publicly available.

FIGURE LEGENDS

Figure 1. Haplotype proportion from the phase I, phase II, and phase IV group

Haplotype proportion in terms of natural history of disease, ethnic background and HBV genotype. Samples were sorted from least number of haplotype diversity (top y axis) to the highest number of haplotype diversity (bottom y axis) separated by genotype and ethnic background. The X axis shows the proportion of each haplotype in the individual samples. Black bar denotes haplotypes below 1% abundance which were removed from analysis. Abbreviations: A, B, C, D = genotype A, B, C D, AS = Asian, Cau = Caucasian, PI = Pacific Islander, AA = African American, OT = other

Figure 2. Scatter plot presentations showing haplotype frequencies for (2.a) phase I, phase II and phase IV of CHB. **(2.b)** HBV genotypes A to D, **(2.c)** Phase II and phase IV group in the Caucasian population. **(2.d)** Phase I, phase II and phase IV in the Asian population, (**2.e**) Grouped by phase II and phases IV of CHB. Between the two phases the statistical analysis was conducted

between the Asian (AS) and Caucasian (Cau) population and (**2.f)** the Asian phase I group between genotypes B and C. The Y axis shows the number of haplotypes per sample. Read lines and bars represent means with standard deviations. Statistically significant differences are shown with FDR (false discovery rate) adjusted P values if more than two groups were compared.

Figure 3. Genetic divergence of haplotype sequences grouped by phases of CHB and genotypes showing (a)

Full genome, Pol TP, Pol SP, Pol RT, and Pol Rnase H genes, (b) PreS1, Pres2, HBsAg, X gene and precore and core gene, (c) PreS promotor, S promotor, and X enhancer and (d) the upper regulatory regions and its components NRE, CURS, and BCP.

Figure 4. Conserved amino acid and nucleotides in functional genes in regulatory regions

The conserved amino acid (AA) composition for functional genes and the conserved nucleotide (NT) composition for regulatory regions are shown, specific motifs highlighted by arrows. AA and NT positions that were less than 100% conserved are indicated in black. AA and NT positions that were 100% conserved are shown in colours according to the legend. **(4.a)** Pol TP, Pol SP, Pol RT and Pol Rnase H genes, **(4.b)**PreS1, PreS2, HBsAg, X gene, and precore and core gene; **(4.c)** PreS promotor, S promotor, X enhancer, and upper regulatory region.

Figure 5. PCoA analysis of the precore and core gene (5.a) and of the X gene (5.b) at the sub genotype level showing the haplotype variation based on genetic distances for genotype A, B, C, and C. PH I = phase I, PH II = phase II, PH IV = phase IV

Figure 6. HBeAg loss by week 192 was associated with different haplotype populations

CCA analysis of haplotype populations between patients who lost HBeAg at week 192 and those who did not loss HBeAg at week 192, for genotypes A2 (**6.a**), B2 (**6.b**), C1 (**6.c**), C2 (**6.d**) and D1 (**6.e**). The number of patients and haplotypes for each genotype and groups are shown in (**6.f**).

References

1. Kim BK, Revill PA, Ahn SH. HBV genotypes: relevance to natural history, pathogenesis and treatment of chronic hepatitis B. Antivir. Ther. [Internet]. 2011;16:1169–1186. Available from: http://www.intmedpress.com/journals/avt/abstract.cfm?id=1982&pid=48
2. Kramvis A. Genotypes and Genetic Variability of Hepatitis B Virus. Intervirology [Internet]. 2014;57:141–150. Available from: https://www.karger.com/Article/FullText/360947
3. Lampertico P, Agarwal K, Berg T, Buti M, Janssen HLA, Papatheodoridis G, et al. EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. J. Hepatol. [Internet]. 2017;67:370–398. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016882781730185X
4. Terrault NA, Lok ASF, McMahon BJ, Chang K-M, Hwang JP, Jonas MM, et al. Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. Hepatology [Internet]. 2018;67:1560–1599. Available from: http://doi.wiley.com/10.1002/hep.29800
5. Cornberg M, Lok AS-F, Terrault NA, Zoulim F, Berg T, Brunetto MR, et al. Guidance for design and endpoints of clinical trials in chronic hepatitis B - Report from the 2019 EASL-AASLD HBV Treatment Endpoints Conference‡. J. Hepatol. [Internet]. 2020;72:539–557. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168827819306713
6. Revill PA, Chisari F V, Block JM, Dandri M, Gehring AJ, Guo H, et al. A global scientific strategy to cure hepatitis B. Lancet Gastroenterol. Hepatol. [Internet]. 2019;4:545–558. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2468125319301190
7. Andino R, Domingo E. Viral quasispecies. Virology [Internet]. 2015;479–480:46–51. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0042682215001580
8. Domingo E, Holland JJ. RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL. Annu. Rev. Microbiol. [Internet]. 1997;51:151–178. Available from: http://www.annualreviews.org/doi/10.1146/annurev.micro.51.1.151
9. Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol. Mol. Biol. Rev. [Internet]. 2012;76:159–216. Available from:

http://mmbr.asm.org/cgi/doi/10.1128/MMBR.05023-11

10. Yang Z-T, Huang S-Y, Chen L, Liu F, Cai X-H, Guo Y-F, et al. Characterization of Full-Length Genomes of Hepatitis B Virus Quasispecies in Sera of Patients at Different Phases of Infection. J. Clin. Microbiol. [Internet]. 2015;53:2203–2214. Available from: http://jcm.asm.org/lookup/doi/10.1128/JCM.00068-15
11. Baaijens JA, Aabidine AZ El, Rivals E, Schönhuth A. De novo assembly of viral quasispecies using overlap graphs. Genome Res. [Internet]. 2017;27:835–848. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.215038.116
12. Baaijens J, Roest B van der, Koester J, Stougie L, Schoenhuth A. Full-length de novo viral quasispecies assembly through variation graph construction (prior posting). bioRxiv. 2018;
13. Chen J, Zhao Y, Sun Y. De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding. Bioinformatics [Internet]. 2018;34:2927–2935. Available from: https://academic.oup.com/bioinformatics/article/34/17/2927/4959151
14. Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. Virus Evol. [Internet]. 2018;4. Available from: https://academic.oup.com/ve/article/doi/10.1093/ve/vey007/4999822
15. Knyazev S, Tsyvina V, Melnyk A, Malygina T, Porozov YB, Campbell E, et al. CliqueSNV : Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads. bioRxiv. 2018;1–8.
16. Pandit A, de Boer RJ. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. Retrovirology [Internet]. 2014;11:56. Available from: https://retrovirology.biomedcentral.com/articles/10.1186/1742-4690-11-56
17. Fukuhara T, Yamamoto S, Ono C, Nakamura S, Motooka D, Mori H, et al. Quasispecies of Hepatitis C Virus Participate in Cell-Specific Infectivity. Sci. Rep. [Internet]. 2017;7:45228. Available from: http://www.nature.com/articles/srep45228
18. López-Labrador FX, Natividad-Sancho A, Pisareva M, Komissarov A, Salvatierra K, Fadeev A, et al. Genetic characterization of influenza viruses from influenza-related hospital admissions in the St. Petersburg and Valencia sites of the Global Influenza

Hospital Surveillance Network during the 2013/14 influenza season. J. Clin. Virol. [Internet]. 2016;84:32–38. Available from: https://linkinghub.elsevier.com/retrieve/pii/S138665321630539X

19. Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying influenza virus diversity and transmission in humans. Nat. Genet. [Internet]. 2016;48:195–200. Available from: http://www.nature.com/articles/ng.3479
20. González C, Tabernero D, Cortese MF, Gregori J, Casillas R, Riveiro-Barciela M, et al. Detection of hyper-conserved regions in hepatitis B virus X gene potentially useful for gene therapy. World J. Gastroenterol. [Internet]. 2018;24:2095–2107. Available from: http://www.wjgnet.com/1007-9327/full/v24/i19/2095.htm
21. Mei F, Ren J, Long L, Li J, Li K, Liu H, et al. Analysis of HBV X gene quasispecies characteristics by next-generation sequencing and cloning-based sequencing and its association with hepatocellular carcinoma progression. J. Med. Virol. [Internet]. 2019;91:1087–1096. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25421
22. Marcellin P, Heathcote EJ, Buti M, Gane E, de Man RA, Krastev Z, et al. Tenofovir Disoproxil Fumarate versus Adefovir Dipivoxil for Chronic Hepatitis B. N. Engl. J. Med. [Internet]. 2008;359:2442–2455. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa0802878
23. Chan HLY, Chan CK, Hui AJ, Chan S, Poordad F, Chang T-T, et al. Effects of Tenofovir Disoproxil Fumarate in Hepatitis B e Antigen-Positive Patients With Normal Levels of Alanine Aminotransferase and High Levels of Hepatitis B Virus DNA. Gastroenterology [Internet]. 2014;146:1240–1248. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0016508514001036
24. Bayliss J, Yuen L, Rosenberg G, Wong D, Littlejohn M, Jackson K, et al. Deep sequencing shows that HBV basal core promoter and precore variants reduce the likelihood of HBsAg loss following tenofovir disoproxil fumarate therapy in HBeAg-positive chronic hepatitis B. Gut [Internet]. 2017;66:2013–2023. Available from: http://gut.bmj.com/lookup/doi/10.1136/gutjnl-2015-309300
25. Hammer Ø, Harper DAT, Ryan PD. PAST: Paleontological statistics sofware package for education and data analysis. Palaeontol. Electron. 2001;4:1–9.
26. Li T, Robert EI, van Breugel PC, Strubin M, Zheng N. A promiscuous α-helical motif

anchors viral hijackers and substrate receptors to the CUL4–DDB1 ubiquitin ligase machinery. Nat. Struct. Mol. Biol. [Internet]. 2010;17:105–111. Available from: http://www.nature.com/articles/nsmb.1719

27. Yu X, Jin L, Jih J, Shih C, Hong Zhou Z. 3.5Å cryoEM Structure of Hepatitis B Virus Core Assembled from Full-Length Core Protein. PLoS One [Internet]. 2013;8:e69729. Available from: https://dx.plos.org/10.1371/journal.pone.0069729
28. Nassal M. The arginine-rich domain of the hepatitis B virus core protein is required for pregenome encapsidation and productive viral positive-strand DNA synthesis but not for virus assembly. J. Virol. [Internet]. 1992;66:4107–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1602535
29. Wynne S., Crowther R., Leslie AG. The Crystal Structure of the Human Hepatitis B Virus Capsid. Mol. Cell [Internet]. 1999;3:771–780. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1097276501800095
30. Fanning GC, Zoulim F, Hou J, Bertoletti A. Therapeutic strategies for hepatitis B virus infection: towards a cure. Nat. Rev. Drug Discov. 2019;
31. Revill PA, Penicaud C, Brechot C, Zoulim F. Meeting the Challenge of Eliminating Chronic Hepatitis b Infection. Genes (Basel). 2019;
32. Sozzi V, Walsh R, Littlejohn M, Colledge D, Jackson K, Warner N, et al. In Vitro Studies Show that Sequence Variability Contributes to Marked Variation in Hepatitis B Virus Replication, Protein Expression, and Function Observed across Genotypes. J. Virol. [Internet]. 2016;90:10054–10064. Available from: http://jvi.asm.org/lookup/doi/10.1128/JVI.01293-16
33. Sozzi V, Shen F, Chen J, Colledge D, Jackson K, Locarnini S, et al. In vitro studies identify a low replication phenotype for hepatitis B virus genotype H generally associated with occult HBV and less severe liver disease. Virology [Internet]. 2018;519:190–196. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0042682218301338
34. Loeb DD, Hirsch RC, Ganem D. Sequence-independent RNA cleavages generate the primers for plus strand DNA synthesis in hepatitis B viruses: implications for other reverse transcribing elements. EMBO J. 1991;
35. Haines KM, Loeb DD. The Sequence of the RNA Primer and the DNA Template Influence the Initiation of Plus-strand DNA Synthesis in Hepatitis B Virus. J. Mol. Biol.

2007;

36. Beck J, Nassal M. Hepatitis B virus replication. World J. Gastroenterol. [Internet]. 2007;13:48–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17206754

37. Murphy CM, Xu Y, Li F, Nio K, Reszka-Blanco N, Li X, et al. Hepatitis B Virus X Protein Promotes Degradation of SMC5/6 to Enhance HBV Replication. Cell Rep. [Internet]. 2016;16:2846–2854. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211124716310907

38. Decorsière A, Mueller H, van Breugel PC, Abdul F, Gerossier L, Beran RK, et al. Hepatitis B virus X protein identifies the Smc5/6 complex as a host restriction factor. Nature [Internet]. 2016;531:386–389. Available from: http://www.nature.com/articles/nature17170

*** Corresponding author:**

Professor Peter Revill

The Peter Doherty Institute for Infection and Immunity

792 Elizabeth St, Melbourne VIC 3000

W: +61 (0) 3 9342 9604

Peter.Revill@vidrl.org.au

List of Abbreviations

HBV = hepatitis B virus

CHB = chronic Hepatitis B

A, B, C, D = genotype A, B, C D

AS = Asian

Cau = Caucasian

PI = Pacific Islander

AA = African American

OT = other

Pol TP = polymerase terminal protein

Pol SP = polymerase terminal spacer

Pol RT = polymerase reverse transcriptase

Pol Rnase H genes = polymerase Rnase H genes

PreS1 = pre surface 1 gene

Pres2 = pre surface 2 gene

HBsAg = Hepatitis B surface antigen

HBeAg = Hepatitis B e antigen

URR = the upper regulatory regions and its components NRE, CURS, and BCP

NRE = negative regulatory element

CURS = core upstream regulatory sequence

BCP = basal core promotor

Financial support statement:

This study was funded by Gilead Sciences Inc., Victorian Infectious Disease Reference Laboratory, Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, and the Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne.

Table 1. Patient characteristics and demographics

	Total patients	Phase I	Phase II	Phase IV
Number	368	96	159	113
Mean age (years)	37.6 (18-64)	34.6 (18-62)	34.7 (18-64)	44.3 (20-63)
Female	126 (34.2%)	51 (53.13%)	49 (30.63%)	26 (23.01%)
Genotype A	59	0	41	18
Genotype B	79	52	20	7
Genotype C	105	44	50	11
Genotype D	125	0	48	77
Ethnicity		Patient number by genotype	Patient number by genotype	Patient number by genotype
Asian	180 (49%)	91 (52B, 39C)	68 (1A, 19B, 45C, 3D)	21 (2A, 7B, 10C, 2D)
Caucasian	178 (48%)	1 (C)	84 (37A, 1B, 1C, 45D)	92 (16A, 1C, 75D)
Pacific Islander	7	3 (C)	4 (1A, 3C)	0
African American	2	0	2 (A)	0
Other	2	1 (C)	1 (C)	0
HB viral DNA mean (min-max, SD) (log10 IU/ml)	7.62 (3.61-9.76, 1.23)	8.4 (7.24-8.98, 0.36) Phase I & II P value = 0.0268	8.02 (5.24-9.76, 0.9) Phase II & IV P value <0.0001	6.39 (3.61-8.77, 1.21) Phase IV & 1 P value <0.0001
HBeAg status mean (min-max, SD) (log10 PE IU/ml)	3.09 (0.1-3.87, SD 0.75)	3.45 (1.33-3.79, SD 0.37) Phase I & II P value <0.0001	2.88 (0.1-3.87, SD 0.84)	all negative
HBsAg level, mean (min-max, SD) (log10 IU/ml)	4.3 (1.01-5.4, 0.71)	4.72 (2.73-5.2, 0.4) Phase I & II P value = 0.0019	4.44 (1.01-5.4, 0.71) Phase II & IV P value <0.0001	3.75 (1.73-5.29, 0.59) Phase IV & 1 P value <0.0001
ALT enzyme mean (min-max, SD) (IU/L)	123.03 (6-884, 115.6)	26.61 (6-43, 7.91) Phase I & II P value <0.0001	163 (40-742, 111)	150.32 (39-884, 123.42) Phase IV & I P value <0.0001
Fibrosis stage mean (min-max, SD)		not tested	3.31 (1-6, 1.55)	3.21 (1-6, 1.47)

Abbreviations: min = minimum, max = maximum, SD = standard deviation, ALT = alanine aminotransferase, HB = hepatitis B. The Kruskal-Wallis test with Dunn's multiple comparisons test was used for the statistical analysis between all three disease Phases. The Mann-Whitney test was used for the statistical analysis between two disease Phases. The statistical analysis was done in GraphPad Prism 8 for macOS version 8.3.0.

Table 2. Conserved nucleotide proportion (%) in the full-length HBV genome and the different sub genes and regions, grouped by genotypes and phases of CHB for functional genes and regulatory regions,

HBV region	all haplotypes	Genotype A		Genotype B			Genotype C			Genotype D	
		Ph II	Ph IV	Ph I	Ph II	Ph IV	Ph I	Ph II	Ph IV	Ph II	Ph IV
Complete genome	32.4	83.0	**73.9**	81.1	**79.2**	84.3	73.4	**62.9**	81.5	70.1	**52.9**
	Functional genes										
Pol TP	32.1 (23.6)	82.8	**74.3**	80.6	**77.6**	81.7	70.7	**65.5**	78.9	71.1	**53.7**
Pol Sp	20.3 (5.9)	79.7	**68.8**	72.6	**72.4**	86.0	67.7	**50.9**	78.5	65.5	**45.0**
Pol RT	37.3 (34.3)	84.3	**75.4**	83.7	**79.3**	85.7	76.4	**64.7**	82.8	70.0	**53.6**
Pol RNase H	41.8 (32.7)	91.3	**80.1**	81.0	84.2	87.4	77.5	**67.3**	84.4	79.2	**60.0**
PreS1	22.4 (16.8)	84.0	**74.5**	**71.4**	73.1	86.0	68.9	**55.7**	76.8	68.6	**51.0**
PreS2	18.2 (3.6)	69.1	**53.3**	76.4	**72.1**	86.1	66.1	**46.1**	83.0	**14.5**	32.7
HBsAg	40.7 (9.7)	85.7	**77.0**	90.4	**83.3**	88.5	83.2	**67.7**	86.0	74.8	**57.7**
X gene	38.3 (14.3)	87.7	**81.0**	**82.0**	83.3	78.6	77.7	**68.8**	87.0	75.3	**56.3**
PC and Core	30.1 (17.8)	79.4	**69.9**	85.9	**82.9**	**82.9**	72.2	**65.7**	79.8	64.3	**51.6**
	Regulatory regions										
PreS promoter	28.7	73.0	**74.7**	**77.5**	82.0	87.6	70.8	**51.7**	78.1	63.5	**42.1**

S promotor	15.2	66.7	**57.3**	93.6	**88.9**	93.0	66.1	**61.4**	68.4	52.6	**43.3**
X enhancer	38.3	87.7	**77.8**	**75.3**	78.9	85.9	71.5	**62.1**	81.4	72.4	**57.0**
URE	25.1	80.8	**78.3**	77.8	77.3	**70.0**	71.9	**59.6**	84.7	66.0	**47.8**

Figure 1. Haplotype proportion from Phase I, Phase II, and Phase IV chronic Hepatitis B

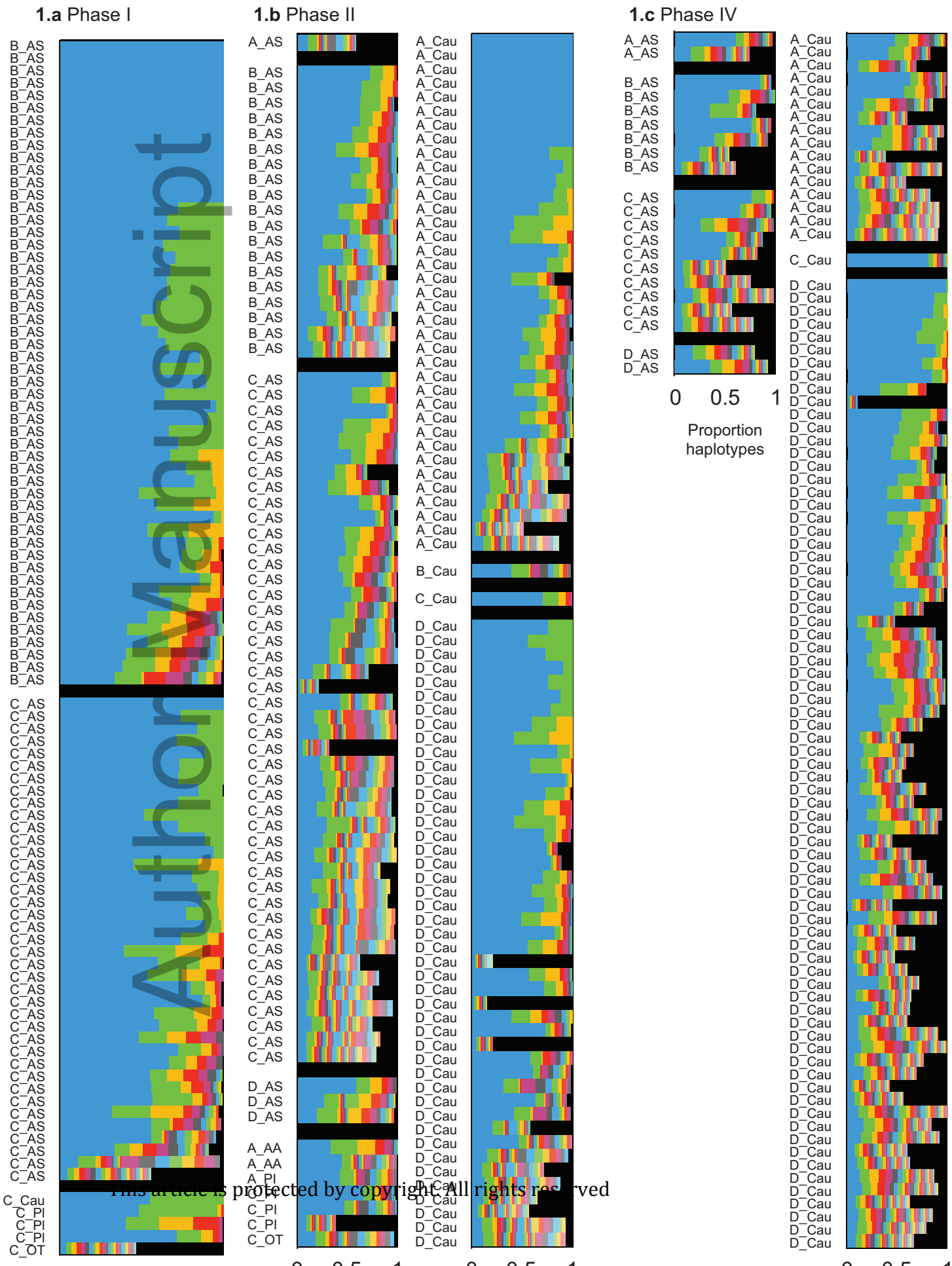
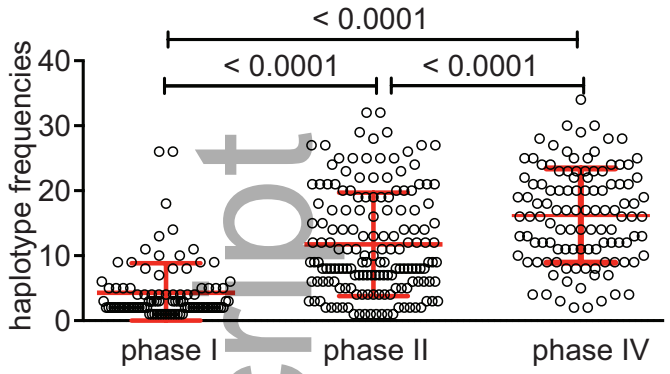
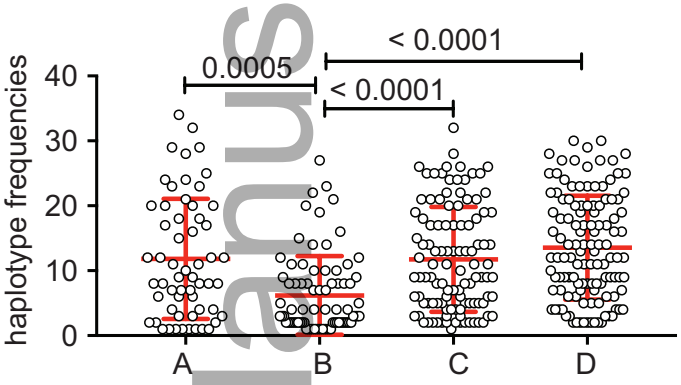


Figure 2. Haplotype frequencies differ in terms of phases of CHB, ethnicity, and genotypes

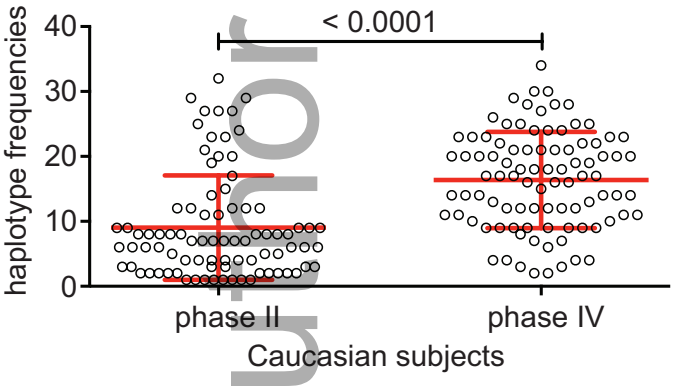
2.a Haplotype frequencies - Disease history



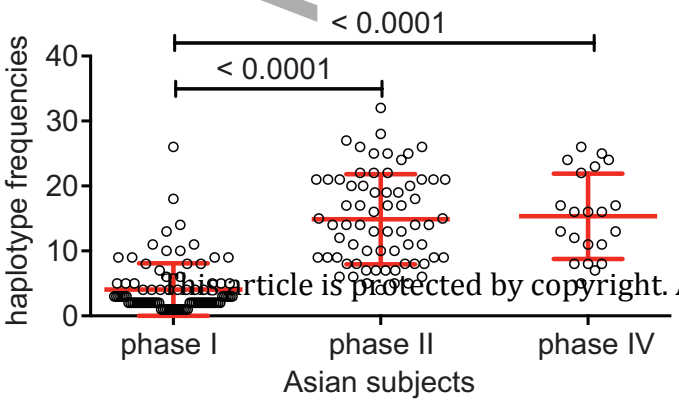
2.b Haplotype frequencies - Genotype



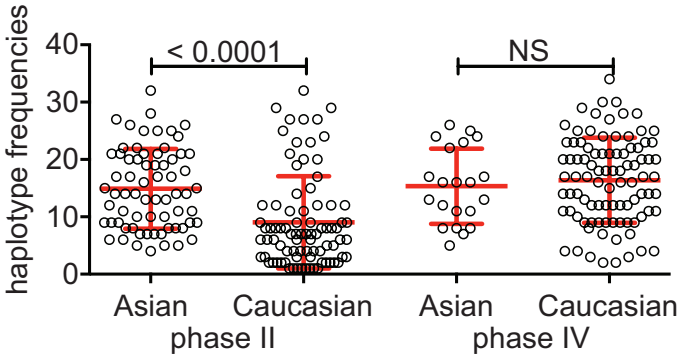
2.c Haplotype frequencies - Disease history in Caucasian subjects



2.d Haplotype frequencies - Disease history in Asian subjects



2.e Haplotype frequencies - phase II and phase IV between Asian and Caucasian subjects



2.f Haplotype frequencies - genotype B and genotype C differences in the phase I Asian group

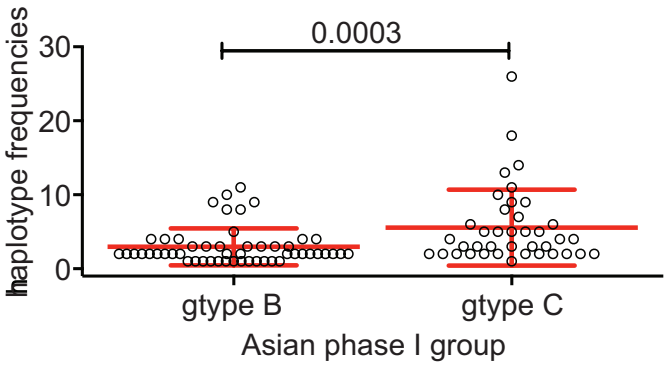


Figure 3.b Genetic divergence for PreS1, PreS2, HBsAg, X gene, and precore (PC) and core

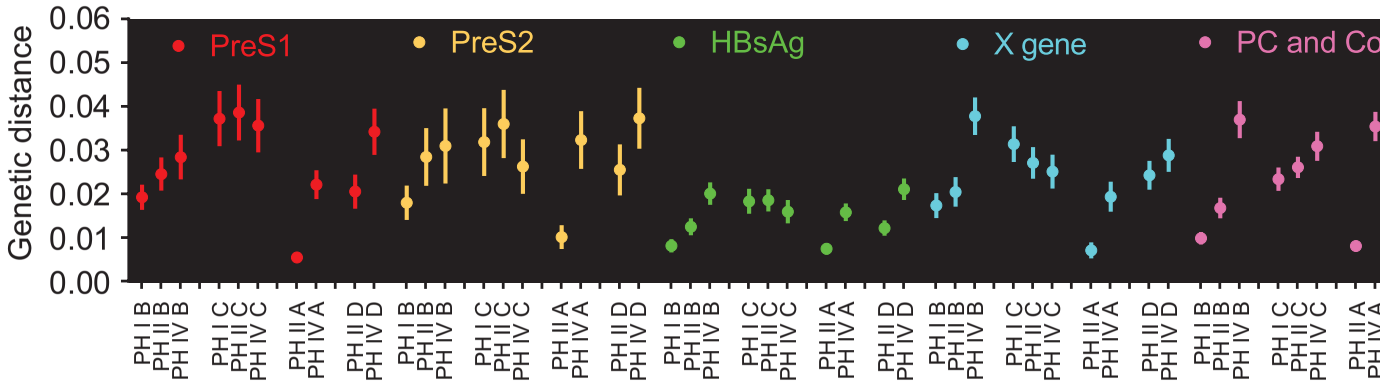


Figure 3.c Genetic divergence for PreS promoter, S promoter and X enhancer

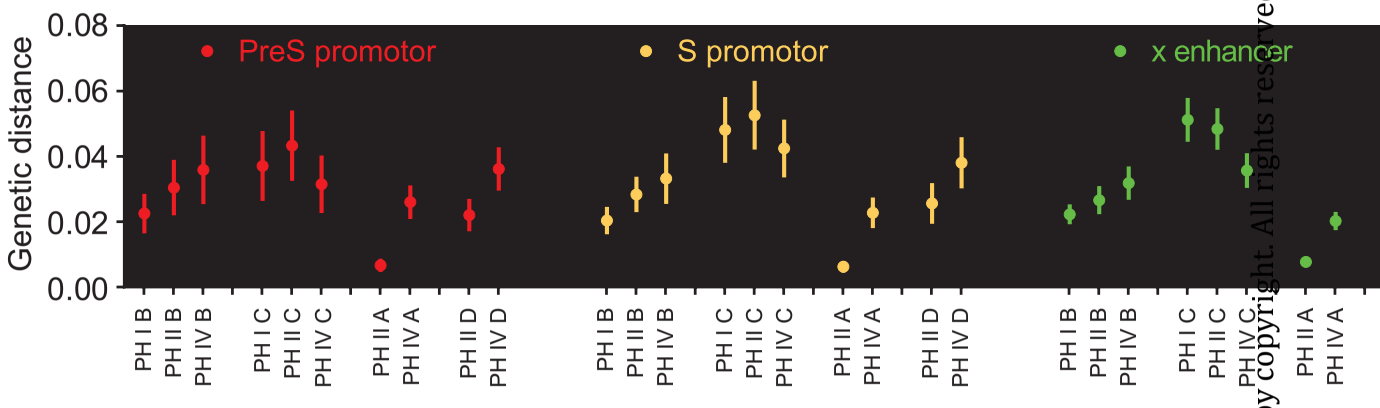


Figure 3.d Genetic divergence for upper regulatory region (URR) and its sub compotents NR
CURS, and BCP

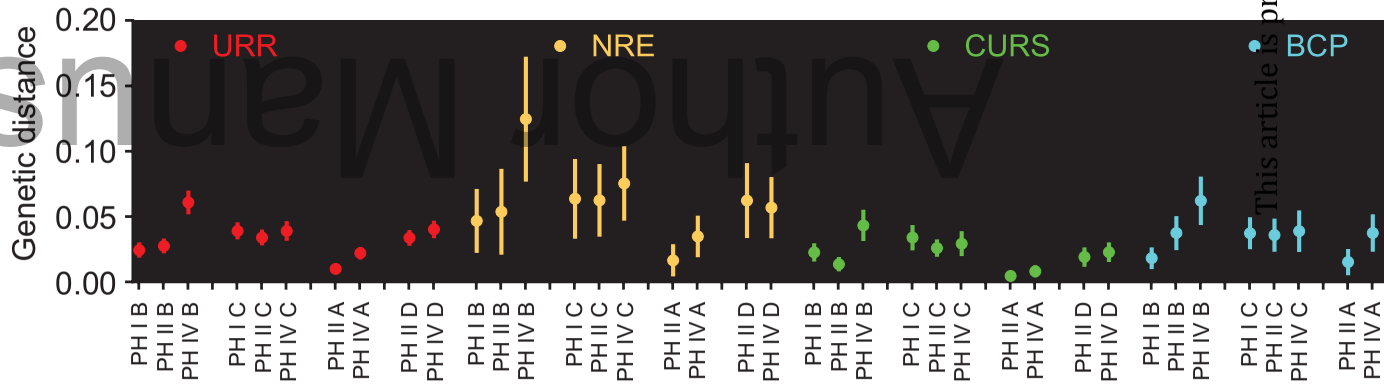


Figure 4.a Conserved amino acid composition across the polymerase (Pol) terminal protein (TP), spacer (SP), reverse transcriptase (RT), and Rnase H regions

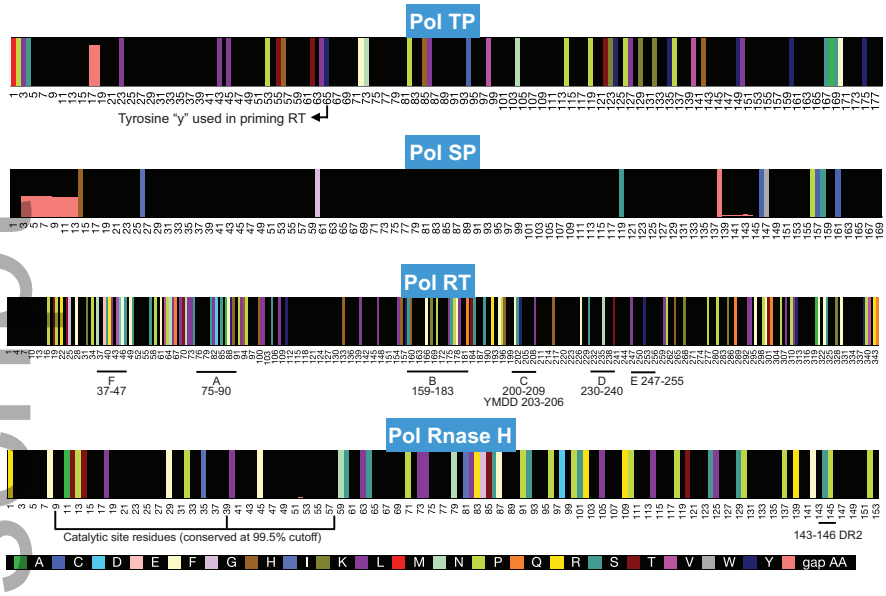


Figure 4.b Conserved amino acid composition across PreS1, PreS2, HBsAg, X gene, and PC and Core genes

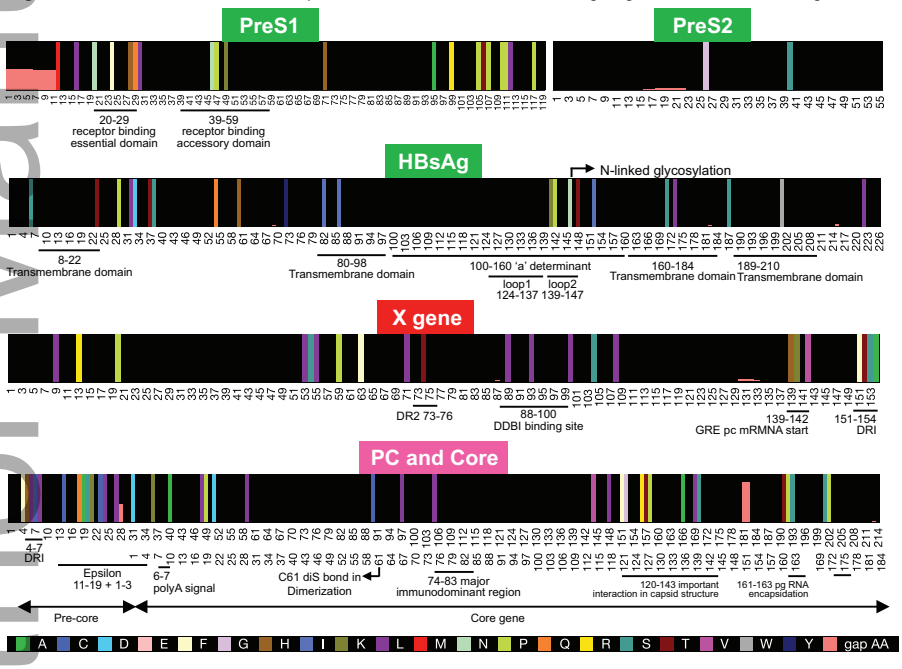


Figure 4.c Conserved nucleotide composition across four regulatory regions

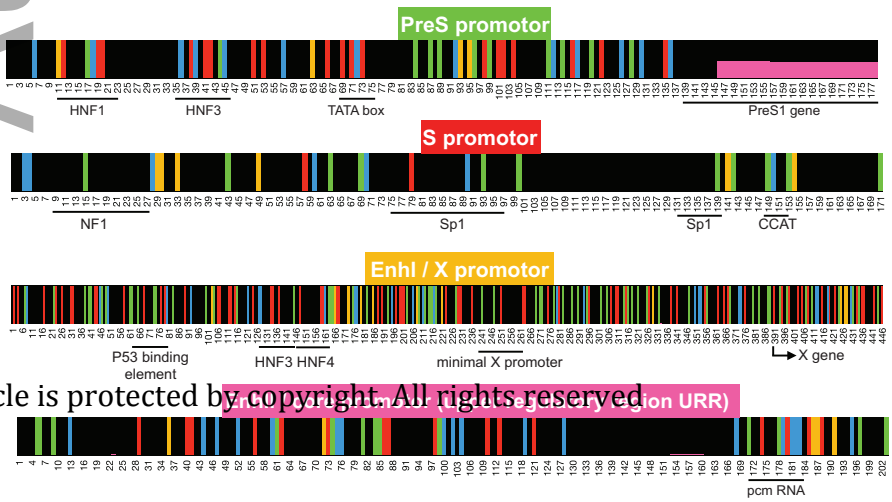


Figure 5.a PCoA analysis of the pre-core and core gene at the sub genotype level revealed different haplotype population between phases of CHB

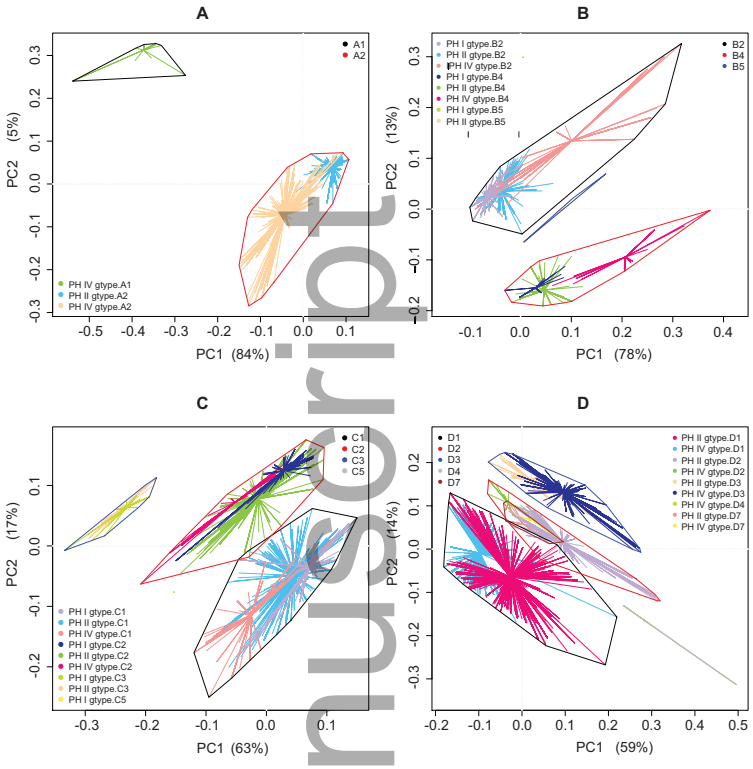
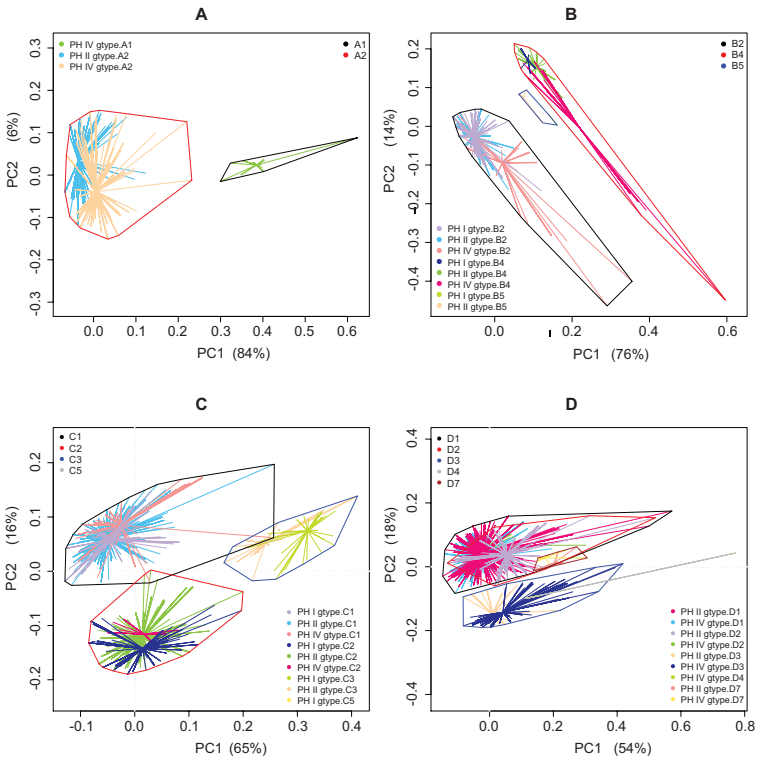
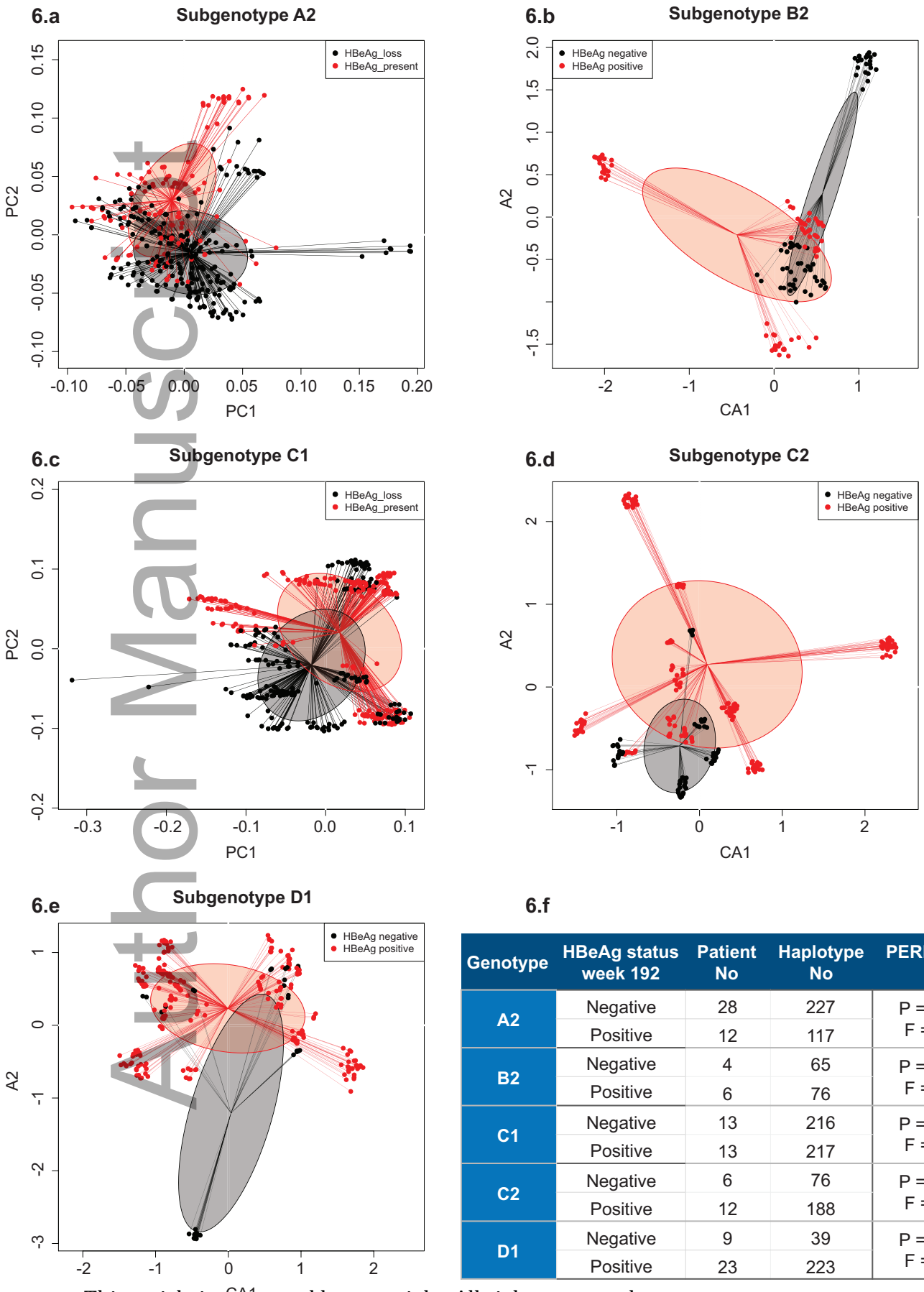


Figure 5.b PCoA analysis of the X gene at the sub genotype level revealed different haplotype population between phases of CHB



hep_31516_f5.eps

Figure 6. HBeAg loss by week 192 is associated with different haplotype populations at baseline



6.f

Genotype	HBeAg status week 192	Patient No	Haplotype No	PERMANOVA test
A2	Negative	28	227	P = 0.0001 F = 14.32
	Positive	12	117	
B2	Negative	4	65	P = 0.0001 F = 20.93
	Positive	6	76	
C1	Negative	13	216	P = 0.0001 F = 23.74
	Positive	13	217	
C2	Negative	6	76	P = 0.0001 F = 14.23
	Positive	12	188	
D1	Negative	9	39	P = 0.0001 F = 16.73
	Positive	23	223	