Molecular characterization of full genome hepatitis b virus sequences from an urban hospital cohort in Pretoria, South Africa

Louis Stephanus Le Clercq

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

MOLECULAR CHARACTERIZATION OF FULL GENOME HEPATITIS B VIRUS SEQUENCES FROM AN URBAN HOSPITAL COHORT IN PRETORIA, SOUTH AFRICA.

By

LOUIS STEPHANUS LE CLERCQ

Submitted in partial fulfilment of the requirements for the degree

M.Sc. Medical Virology

in the

Faculty of Health Sciences

School of Medicine

Department of Medical Virology

University of Pretoria

Pretoria

South Africa

18 June 2014

DECLARATION

I, the undersigned, declare that the dissertation hereby submitted to the University of Pretoria for the degree M.Sc. Medical Virology and the work contained herein is my own original work and has not previously, in its entirety or in part, been submitted to any university for a degree.

Signature: …….

Date: …….

QUOTE

"It is the mark of an educated mind to be able to entertain a thought without accepting it."

- Aristotle

ACKNOWLEDGEMENTS

I would firstly like to acknowledge an thank my former mentor and role model **Dr Duncan Cromarty**, from the **Department of Pharmacology**, for taking in a wide-eyed undergraduate and, in the matter of a year, turning him into a scientist of some worth. The intellectual acumen, theoretical knowledge and practical know-how I've gained from you has proven invaluable to me. I can but only hope and aspire to be a scientist of your calibre one day.

To my current mentor, **Dr Sheila Bowyer** from the **Department of Medical Virology**, I would firstly like to thank you for taking a chance on me and welcoming me to your research group. I would also like to thank you for giving me yet another perspective on research and guiding me in understanding new and complex molecular concepts. In addition to making me somewhat of a virologist you have also given me the opportunity to truly learn and understand bioinformatics in a hands-on manner and challenged me with the analysis of highly complex data that most virologists haven't even dealt with. I have undoubtedly acquired a unique subset of skills that will serve me well for the rest of my career and have overcome any initial fears of bioinformatics.

I would also like to thank **Dr Sim Mayaphi,** also from the **Department of Medical Virology**, for entrusting his specimens to me and helping me find my feet in a new learning environment by teaching me about working in a "wet lab" and performing molecular techniques like PCR. This was my first time working with a clinician and your insights towards the clinical relevance of our research has been enriching.

I would also like to acknowledge and thank the **Poliomyelitis Research Foundation** and the **National Research Fund** for financial contributions to the project and my studies.

Lastly, I would like to acknowledge and thank my mother, **Petro Smit**. Your unwavering support throughout the long and winding road to becoming a scientist has been the foundation for me to stand on. I know the financial burden of taking care of two students singlehandedly has been very stressful and yet you always stayed strong and committed to a better future for us. Donc, cet un est pour vous Maman!

MOLECULAR CHARACTERIZATION OF FULL GENOME HEPATITIS B VIRUS (HBV) SEQUENCES FROM AN URBAN HOSPITAL COHORT IN PRETORIA, SOUTH AFRICA.

By

LOUIS STEPHANUS LE CLERCQ

SUPERVISOR: DR S.M. BOWYER

CO-SUPERVISOR: DR S.H. MAYAPHI

DEPARTMENT: MEDICAL VIROLOGY

DEGREE: M.Sc. MEDICAL VIROLOGY

SUMMARY

Hepatitis B Virus (HBV) is a DNA virus and belongs to the genus *Orthohepadnavirus* of the *Hepadnaviridae* family which represents one of two animal viruses with a DNA genome which replicates by reverse transcription of a viral RNA intermediate. Nucleotide variation led to further sub-classification into 8 genotypes (A to H). The reverse transcription step within its life cycle is prone to the introduction of errors and recombination when dually infected. This leads to a viral quasispecies which forms during the course of infection with many minor population variants; such variants can however only be detected by means of ultra-deep sequencing. A recent study in the Department of Medical Virology (UP) by Mayaphi et al. identified a number of the specimens that partitioned away from the typical subgenotype A1 clades with high bootstrap values and longer branch lengths. Thus, the main objective of the current study was to characterize the full genome of all variants for the outliers observed in the aforementioned

study, inclusive of potential recombination, dual infection and minor populations. Twenty samples were selected from a previous cohort for purposes of the present study. The viral DNA was extracted and amplified by PCR according to the methods described by Günther et al. with modified primer sets. Nineteen of the samples were successfully amplified and 15 of these were sequenced. Specimens were sequenced by NGS on the Illumina MiSeq™ sequencer and sequence data used to reconstruct the viral quasispecies of each specimen. Further analyses of the reconstructed variants included molecular characterization as well as phylogenetic analysis and screening for recombination and drug resistance mutations. Full genome coverage was obtained for twelve of the fifteen samples and full genome variants reconstructed, generating nearly 40 full genomes. Phylogenetic analysis showed that the majority of the samples are of genotype A, more specifically of subgenotype A1, differing by less than 4% from known sequences. The phylogenetic analysis revealed a similar clade of outliers, where four samples clustered together with significant bootstrap support (75%) and a fifth sample partitioned separate from, yet close to, this clade, away from the typical African A1 clade. This clade was assigned to genogroup III. Three samples were of the Asian A1 clade (genogroup I) with remaining specimens grouping within genotype D and E. The variants showed low diversity within each specimen with some differing at but a few positions across the genome while even the most diverse quasispecies differed by less than a percentage (32 positions). Several unique and atypical positional variations were observed amongst study samples of which some were present in but one of the variants for that sample. Twenty-six lead to shared amino acid changes. Some observed changes, such as

A1762T/G1764A and G1896A, could explain the serological patterns such as HBeAg negativity while others, such as C2002T, were previously implicated in disease progression and severity. Sample N199 presented a longer branch length and revealed short regions within the genome that display evidence of recombination between HBV/A1 and HBV/A2. The results illustrate the utility of NGS technology in characterizing viral variants.

Keywords: Hepatitis B virus, genotypes, variants, quasispecies, Next Generation Sequencing, QuRe, GALAXY, phylogenetics, recombination.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

MOLEKULÊRE KARAKTERISERING VAN VOLLEDIGE GENOOM HEPATITIS B-VIRUS VOLGORDES UIT 'N STEDELIKE HOSPITAAL STUDIEGROEP IN PRETORIA, SUID-AFRIKA.

Deur

LOUIS STEPAHANUS LE CLERCQ

PROMOTOR: DR S.M. BOWYER

MEDE-PROMOTOR: DR S.H. MAYAPHI

DEPARTEMENT: GENEESKUNDIGE VIROLOGIE

GRAAD: M.Sc. GENEESKUNDIGE VIROLOGIE

OPSOMMING

Die hepatitis B virus (HBV) is 'n DNS virus ontdek in die vroëe 1960's. Hierdie virus behoort tot die *Orthohepadnavirus* genus van die *Hepadnaviridae* familie en verteenwoordig een van twee dierevirusse met 'n DNS genoom met die kenmerk om te repliseer deur middel van omgekeerde transkripsie van 'n virale RNS tussenganger. Dit is welom bekend dat HBV, 'n virus wat staatmaak op 'n omgekeerde transkripsie proses tydens sy lewensiklus , genuig is tot die induksie van foute en herkombinering weens waarskynlike dubbele infeksies. Dit gee aanleiding tot die vorming van 'n virale kwasi-spesies gedurende die gang van 'n infeksie, met verskeie minderheids populasie variante. 'n Onlangse studie in die navorsingslaboratorium van die Departement van Geneeskundige Virologie, Universiteit van Pretoria, het verskeie monsters identifiseer wat weg verdeel vanaf die tipiese subgenotiep A1 klade met hoë 'bootstrap' waardes en lang taklengtes. Dus, was dit hoofsaaklik die doel van die huidige studie om die

volledige genoom variante van die uitskieters geidentifiseer in die voorafgenoemde studie te karakteriseer, insluitend van potensiële herkombinering, dubbelle infeksies en minderheids populasie variante. Twintig monsters was geselekteer vanuit die vorige studiegroep vir die doeleindes van die huidige studie. Die virale DNS van die monsters was geïsoleer en gebruik in PKR amplifisering volgens die metodes beskryf deur Günther et al. (1995) met aangepasde peiler stelle. Geamplifiseerde monsters was gestuur vir NGV volgordebepaling op die Illumina MiSeq™ volgordebepaler. Volgorde data was gebruik om die virale kwasispesies van elke monster te rekonstrueer. Verdere analises van die rekonstrueerde variante sluit in molekulêre karakterisering sowel as filogenetiese analise en om te skerm vir herkombinering en weerstand bieddende mutasies. Negentien van die studie monsters was suksesvol geamplifiseer en 15 van die was gebruik in volgordebepaling. Volledige dekking van die genoom was behaal in twaalf van die vyftien monsters en die volgenoom variante gerekonstrueer, bykans 40 volledige genome was so gegeneer. Filogenetiese analise het getoon dat die meerderheid van die monsters van genotiep A is, meer spesifiek van subgenotiep A1, met minder as 4% verskil vanaf bestaande volgordes. Die twee oorbleiwende monsters het saam genotiep D en E groepeer. Die variante het ‘n lae vlak van diversiteit getoon binne elke monster, waar sommiges veranderings toon by enkele posisies oor die volle genoom terwyl ander wat heelwat meer diversiteit toon in die kwasi-spesies steed by minder as ‘n persentasie van mekaar verskil. Verskeie unieke en ongewone posisionele variasies was waargeneem en gedeel tussen studiemonsters waarvan sommiges slegs teenwoordig in enkele variante van die betrokke monster was. Die

waargenome veranderings, soos A1762T/G1764A en G1896A, kon serologiese patrone soos HBeAg negatiwiteit verklaar terwyl ander, soos S2002T, voormalig betrek is by siektestoestand ontwikkeling en erns. Een studiemonster, N199, het ‘n langer taklengte getoon en het tydens herkombineringsanalise kort streke in die volle genoom onthul wat op tekens van herkombinering bedui. Die resultate hiermee gelewer illustreer die gebruiklikheid van NGV tegnologie in die karakterisering van virale variante op ‘n molekulêre vlak.

Sleutelwoorde: Hepatitis B virus, genotipes, variante, kwasispesies, Nuwe Generasie Volgordebepalings, QuRe, GALAXY, filogenetika, herkombinering.

TABLE OF CONTENTS

ABBREVIATIONS

3’	= three prime
5’	= five prime
A (Ala)	= Alanine
A	= Adenine (nucleic acid – purine)
aa	= amino acids
ALT	= serum transaminase
anti-HBc	= anti-HBcAg antibody
anti-HBe	= anti-HBeAg antibody
anti-HBs	= anti-HBsAg antibody
BAM	= Binary Alignement/Map format
BCP	= Basic Core Promoter
BIC	= Bayesian Information Criterion
bp	= base pairs
BWA	= Burrows-Wheeler Aligner
C (Cys)	= Cystein
C	= Core region
C	= Cytosine (nucleic acid – pyrimidine)
cccDNA	= covalently closed circular DNA
CD	= Cluster of Differentiation
CDC	= Centre for Disease Control
CHB	= Chronic Hepatitis B
CPD	= Carboxipeptidase
C-terminal	= carboxyl terminal
D (Asp)	= Aspartic acid
DHBV	= Duck HBV
DNA	= Deoxyribonucleic acid
EcoR1	= *E. coli* restriction site 1

ER = Endoplasmic Reticulum

ε-signal = encapsidation signal

et al. = et alia (and others)

F (Phe) = Phenylalanine

Fasta = Fast all (DNA and Protein) format

Fastq = Fast all with quality format

G (Gly) = Glycine

G = Guanine (nucleic acid – purine)

H (His) = Histidine

HBc = Hepatitis B core protein

HBeAb = anti-HBeAg antibody

HBeAg = Hepatitis B e-Antigen

HBsAb = anti-HBsAg antibody

HBsAg = Hepatitis B surface Antigen

HBV = Hepatitis B Virus

HBx = Hepatitis B x protein

HCC = Hepatocellular Carcinoma

HIV = Human Immunodeficiency Virus

I (Ile) = Isoleucine

IDT = Integrated DNA Technologies

IgG = Immunoglobulin *gamma*

IgM = Immunoglobulin *mu*

IGV = Integrated Genome Viewer

INF = Interferon

K (Lys) = Lysine

kbp = kilo base pairs

kDa = kilo Dalton

L (Leu) = Leucine

LHBs = Large Hepatitis B surface protein

M (Met) = Methionine

mAb = monoclonal Antibody

MHBs = Middle Hepatitis B surface protein

MHC = Major Histocompatibility Class

mRNA = messenger RNA

N (Asn) = Asparagine

NA (na) = Nucleic Acid

NCBI = National Centre for Biotechnology Information

NF-κB = Nuclear Factor *kappa* B

NGS = Next Generation Sequencing

NHLS = National Health Laboratory Services

NICD = National Institute for Communicable Diseases

nm = Nanometres

NRF = National Research Fund

N-terminal = amino terminal

ORF = Open Reading Frame

P (Pro) = Proline

P1 = forward primer

P2 = reverse primer

PCR = Polymerase Chain Reaction

PEG = Polyethylene Glycol

pgRNA = pre-genomic RNA

pol = polymerase

pre-C = pre-Core

PRF = Poliomyelitis Research Foundation

QC = Quality Control

R (Arg) = Arginine

RNA = Ribonucleic acid

S (Ser) = Serine

S = Surface region

SAM = Sequence Alignement/Map format

sgRNA = small genomic RNA

siRNA = small inhibiting RNA

SHBs = Small Hepatitis B surface protein

SNP = Single Nucleotide Polymorphism

T (Thr) = Threonine

T = Temperature

T = Thymine (nucleic acid – pyrimidine)

TBE = Tris/Borate/EDTA

TLM = Translocation motif

Tm = melting temperature

TM1 = Transmembrane region 1

TNF = Tumour Necrosis Factor

UP = University of Pretoria

USA = United States of America

V (Val) = Valine

W (Trp) = Tryptophan

Y (Tyr) = Tyrosine

YMDD = tyrosine-methionine-aspartic acid-aspartic acid motif

α = *alpha*

β = *beta*

ΔG = *delta* G/ Gibbs free energy

γ = *gamma*

µL = micro (*mu*) litre

ºC = degrees Celsius

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW

1.1 CLASSIFICATION

Hepatitis B Virus (HBV) is a DNA virus and was first identified in the 1960s. According to the ICTV classification, this virus belongs to the genus *Orthohepadnavirus* of the *Hepadnaviridae* family and, along with the *Spumaretrovirinae* subfamily of the *Retroviridae* family, represents the only other animal virus with a DNA genome known to replicate by the reverse transcription of a viral RNA intermediate (Norder et al. 2004; Seeger et al. 2007). The Hepatitis B Virus is a blood-borne virus and roughly 75 – 200 times more infectious than HIV (Bowyer et al. 2011).

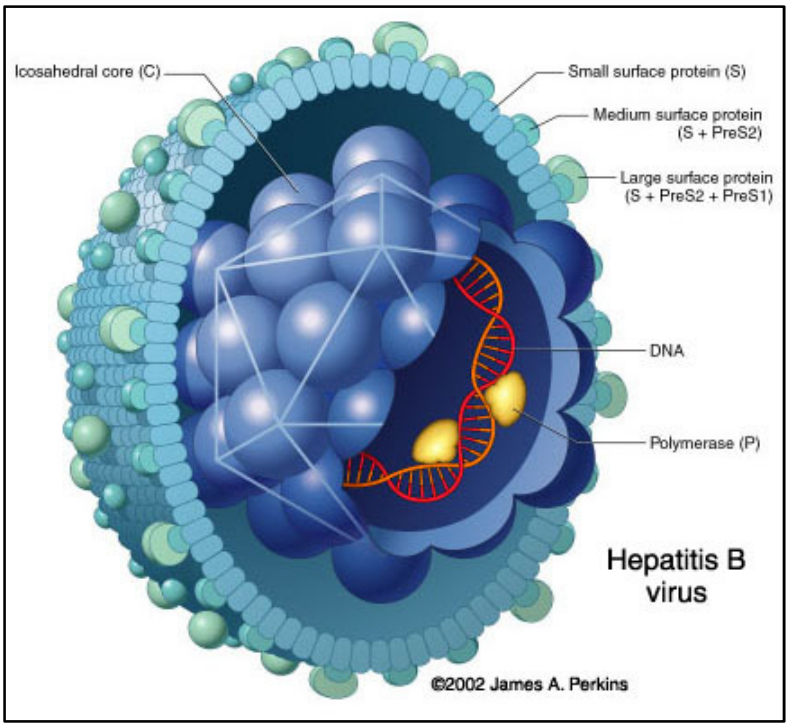
1.2 HBV - THE VIRION

The Dane particle of HBV is a spherical lipid-containing structure of approximately 42 to 47nm. The virion (figure 1.1) consists of a viral envelope, nucleocapsid and a single copy of the partially double-stranded DNA genome. The nucleocapsid is comprised of 120 dimers of core protein and is covered by a capsid membrane embedded with 3 viral envelope proteins, the large (L), middle (M) and small (S) surface proteins (Seeger et al. 2007).

The partially double-stranded DNA genome consists of a minus-strand, which spans the full genome, and a plus-strand of DNA spanning roughly two thirds of the genome. Upon infection of the liver cells, the genome is converted to

covalently closed circular DNA (cccDNA) of which the plus strand is used for the transcription of viral proteins (Bowyer and Sim 2000; Seeger et al. 2007).

A



B

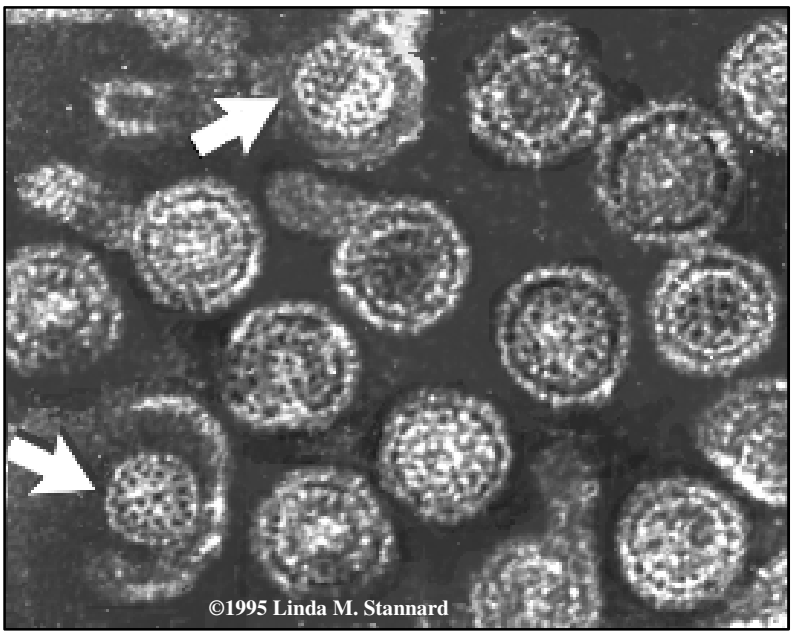


Figure 1.1: **(A)** Schematic representation of the HBV virion and its organization. The darker blue beads represent the core proteins of the nucleocapsid; the light blue to turquoise beads

represents the surface proteins (S) and M and L HBs (Perkins 2002)**. (B)** Electron micrograph of HBV virions in infected liver tissue (Stannard 1995).

The genomic organization of HBV is best depicted as a circular genome (figure 1.2) to better elucidate its overlapping gene and regulatory regions. Because viral replication takes place via an RNA intermediate and uses reverse transcriptase, an enzyme which lacks proof-reading and is known to have a high error rate, the nucleotide exchange rate is 10^4 fold higher than that of typical DNA genomes and estimated to be between 0.1 and 0.7 per annum (Bowyer and Sim 2000; Zhu et al. 2010).

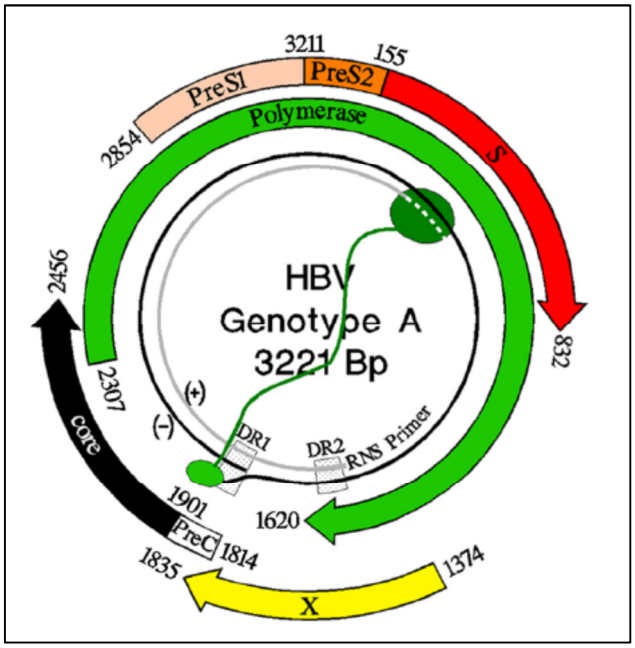


Figure 1.2: Schematic representation of the genomic organization of HBV, indicating nucleotide numbering (from EcoR1 site) of the four overlapping reading frames and the proteins for which they encode (Gerlich 2013).

There are 4 open reading frames, four promoters and two enhancer elements to regulate the transcription of viral RNA. Northern blotting experiments have revealed the four major transcripts of 3.5 kb, 2.4 kb, 2.1 kb and 0.7kb; which are

termed the pre-C/C, pre-S, S and X mRNAs (Seeger et al. 2007), respectively. These transcription regions have been mapped to specific nucleic acid positions across the genome. As mapping positions can vary from strain to strain, it is important to specify which sequence is used by different authors as it greatly affects primer design as well as the interpretation of whole genome sequences in comparison to the chosen standard. One of the most frequently used maps; Genbank accession number X02763 sequenced and reported by Valenzuela et al., will serve as the standard reference sequence for this discussion, with numbering from the EcoR1 site (Seeger et al. 2007; Bowyer et al. 1997; Bowyer and Sim 2000).

1.3 VIRAL GENOME AND ITS TRANSCRIPTS

1.3.1 HBc

The pre-C/C or Core (1814-2456) mRNAs encode the core and pre-core proteins. Core protein is a cytoplasmic, basic phosphoprotein whose antigenicity has been exploited from early days for the detection and monitoring of ongoing or resolved infections (Seeger et al. 2007). The full-length HBcAg (core antigen) is a 183 aa polypeptide. After post-translational modification by enzymatic cleavage, the amino acid residues at position 149 (N-terminal residue) assemble to form dimeric capsid proteins. At the secondary structure level, each HBcAg dimer has four α-helix bundles (figure 1.3) flanked by an α-helix domain on either side. The quaternary structure of the capsid is a lattice of triangles (dimers and trimmers) which assemble (Packianathan et al. 2010) from 90-120 of these dimers where the

four α-helix bundles project outwards as spikes and the flanking α-helices covalently link adjacent dimers (Watts et al. 2010).

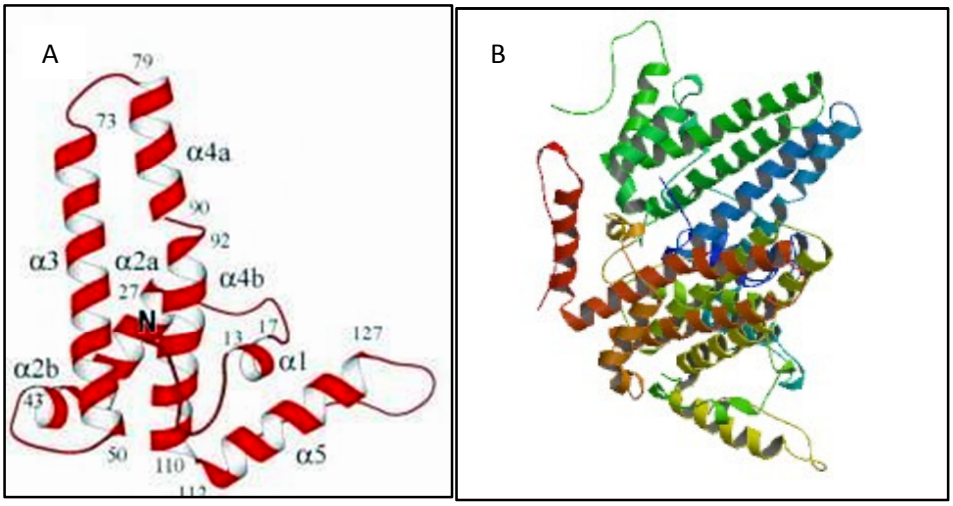


Figure 1.3: **(A)** Ribbon structure of the tertiary structure for HBV Core protein (monomer) with two α-helices projecting up and one to the side (Wynne et al. 1999). **(B)** Three dimensional structure of the core protein (dimerised) based on the x-ray crystallography protein databank file 1QGT.

1.3.2 HBe

The pre-C protein, or e-antigen (HBeAg) as it is serologically termed, is a shorter excreted soluble protein whose exact function is unknown although it is thought to be associated with the regulation of the host immune response in HBV infection (Seeger et al. 2007). Despite close similarities at the sequence level, HBcAg and HBeAg differ in solubility, assembly properties, function, infection kinetics and antigenic specificity (Watts et al. 2010).

Following translation, the e-antigen has an additional 10 N-terminal amino acid residues not present in the core antigen. A 19 amino acid signal peptide at the amino end of the protein targets it for post-translational modification in the

endoplasmic reticulum where the protein is cleaved at a fixed site, the signal peptide cleavage site (amino acid 19), while the cleavage of the C-terminal is variable, but HBeAg is always shorter than the core antigen (Watts et al. 2010; Seeger et al. 2007). HBeAg is water soluble and is secreted into the serum. On the other hand, cytoplasmic HBeAg can enter both the Major Histocompatibility Complex (MHC) class I and class II pathways and be presented at the hepatocyte membrane, targeting infected cells for destruction (Ribeiro et al. 2010). However, studies have shown that the tolerogenic effect of free secreted e-antigen dominates any additional immune responses it might elicit (Ribeiro et al. 2010).

1.3.3 DNA Polymerase (P protein)

A *pol* gene (2307-1620) encodes the viral DNA polymerase. The product has three functional domains: a terminal protein (TP), which acts as a primer for minus-strand DNA synthesis; a reverse transcriptase domain for transcription and a downstream RNAseH domain (Seeger et al. 2007).

1.3.4 HBx

The smallest gene, found only in animal hepadnaviruses, is the gene encoding the hepatitis B x antigen, HBxAg or X protein (1374-1835). This protein predominantly occurs as a soluble cytoplasmic protein but has also been found in association with the cytoskeleton and in the nucleus. HBx is associated with the activation of transcription by interacting with cellular promoters such as NF-κB, AP-1/2, c/EBP, ATF/CREB or NFAT binding sites (Seeger et al. 2007) and non-synonymous change within this gene has recently been implicated in propagating hepatocarcinogenesis (Toh et al. 2013).

1.3.5 HBs

The pre-S/S genes (2854-832) encode the three transmembrane glycoproteins (figure1.4) of the viral membrane. L-protein (PDB 1KCR) is a myristilated polypeptide translated from the first initiation codon of the S open reading frame and is coded by the pre-S1, pre-S2 and S domains. This protein provides a ligand for the viral receptor on hepatocytes. The M-protein represents a form that is larger (3211-832) than HBsAg, but smaller than the L-protein, and is translated from an in-frame initiation codon. The 55 amino acid extension at the N-terminal of M-protein represents the pre-S2 domain. The exact function of this protein is unknown as it does not appear to have a prominent function in virion assembly.

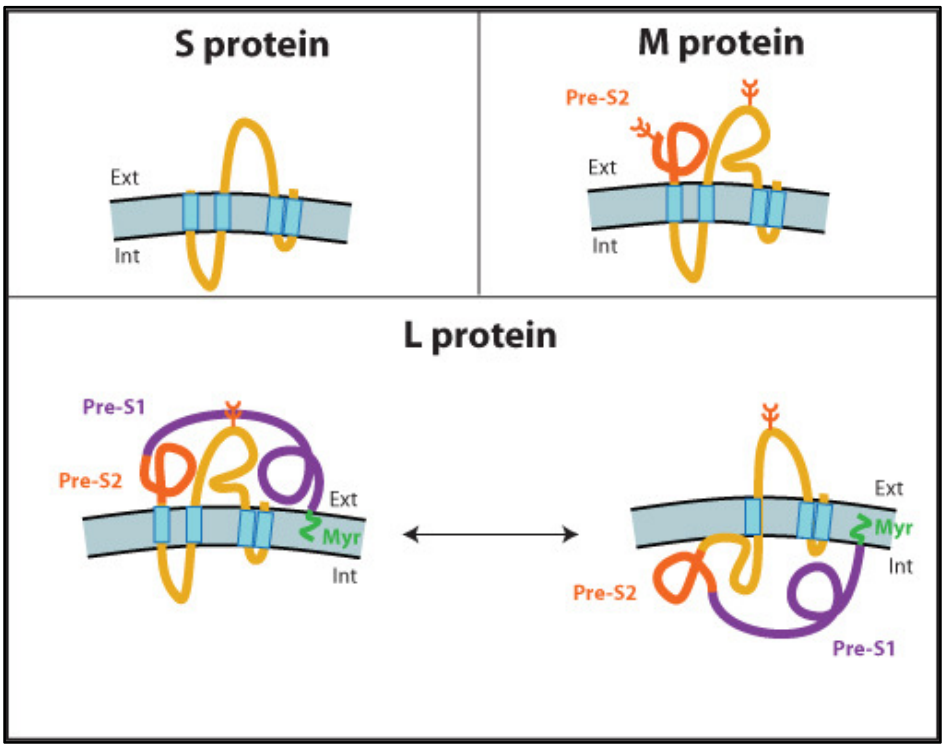


Figure 1.4: Diagrammatic representation of the small (S), middle (M) and large (L) surface proteins, showing both conformations of L HBs (Viral Zone 2011).

The shortest, S-protein, or S-antigen (HBsAg) is translated from a second in-frame initiation codon (155-832) and contains the major antigenic determinants that led to the discovery of HBV and is the basis of diagnostic tests for active infections and vaccines against HBV infection (Seeger et al. 2007).

Before the emergence of commercial scale DNA technologies the HBsAg was used to serologically classify HBV into different strains based on primary structure, amino acid sequence differences. These nine major serotypes are denoted *ayw1, ayw2, ayw3, ayw4, ayr, adw2, adw4q-, adrq+* and *adrq-*. The a-determinant comprises amino acids 124 through 147 with two antigenic loops, the first spanning 124 to 137 and the second 138 to 147 (Locarnini and Yuen 2010). This antigenic determinant is usually highly conserved and shared by all serotypes, representing the major antigenic determinant for HBsAg (Locarnini and Yuen 2010). The *y/d* and *w/r* variations were shown to be a result of Lys/Arg substitutions at the 122 and 160 amino acid residues respectively. Residue 127 was found to be important for the *w1-4* differences; *w1/2* encoded Pro; *w3* encoded Thr; and *w4* encoded Leu. The *q*-determinant, which is expressed by most strains, is defined by amino acid residues 177 and 178 (Norder et al. 2004). The HBsAg is depicted in figure 1.5.

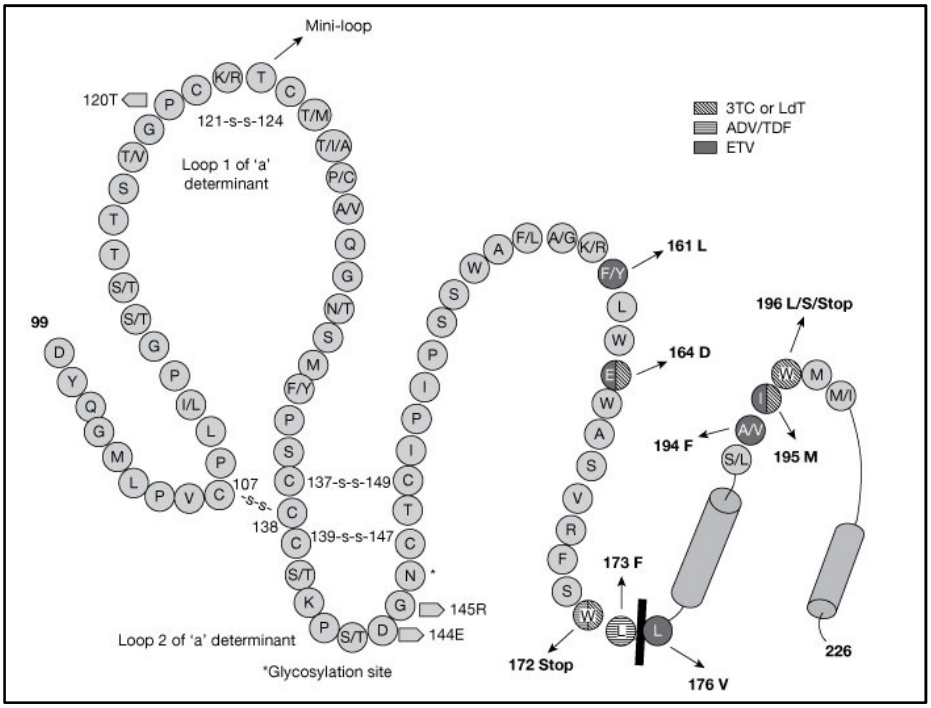


Figure 1.5: Schematic representation of the HBV surface antigen depicting both the major and minor loops of the main antigenic determinant (a-determinant), the legend indicates change in shaded regions associated with drug resistance mutations in the overlapping polymerase gene (Locarnini and Yuen 2010).

1.4 REPLICATION AND INFECTIVE CYCLE

As with all viruses, the first step of infection involves the attachment of the virus to the host cell surface. In the case of HBV this is achieved by the binding of a defined sequence of Pre-S, aa 30-115, to carboxipeptidase D (CPD) in a duck HBV model (Schultz et al. 2004). More recent studies have however reported other modes of attachment; the N-terminal of the S-domain (aa 1-23, transmembrane region 1 [TM1]) and a membrane permeable peptide of the Pre-S2 domain (translocation motif [TLM]) (Schädler and Hildt 2009) in humans.

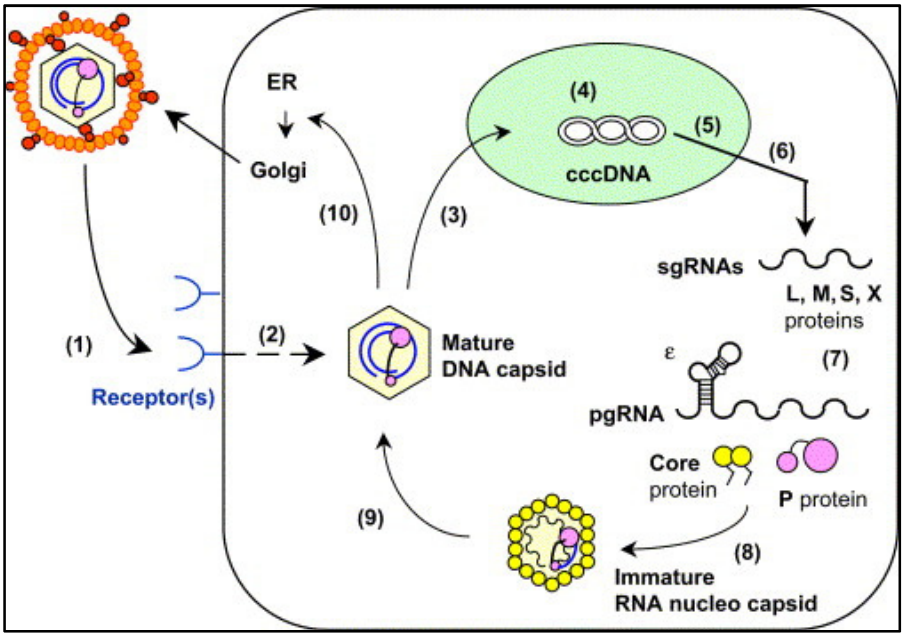


Figure 1.6: Simplified schematic representation of the HBV infective cycle (Schultz et al. 2004) showing attachment (1) and entry (2), uncoating of capsid and nuclear import (3), cccDNA formation (4) followed by transcription (5) and nuclear export (6) of mRNA. In the cytoplasm the mRNA is translated (7) and the different viral components used to assemble new virions (8).

Studies have shown that HBV remains infective if one or two of these modes of attachment and internalization is blocked, indicating that they may act independent of each other or only be utilized at specific internalization and/or export steps (Schädler and Hildt 2009). Subsequent steps of infection are depicted in figure 1.6.

After internalization, the nucleocapsid is released into the cytoplasm (step 2) and transferred to the nucleus by means of nuclear localization signals in the Arginine-rich C-terminal of the core protein and importin α/β (Schultz et al. 2004).

In step 4 the relaxed circular DNA of the genome is converted to cccDNA which then acts as a transcriptional template (step 5) for cellular RNA polymerase II

(Schultz et al. 2004; Seeger et al. 2007; Schädler and Hildt 2009). Transcripts, both sub-genomic and pre-genomic RNA, are then exported to the cytoplasm (step 6) via a post-transcriptional regulatory element (Seeger et al. 2007).

As previously mentioned, the sub-genomic mRNA from the 4 ORF's is translated (step 7) into HBcAg, HBeAg, HBsAg, HBx and P protein. Pre-genomic RNA has a stem-loop structure, the ε-signal (Kramvis and Kew 2002), that enables target recognition for encapsidation (step 8) via chaperone mediated interactions with the P protein. This pgRNA - P protein complex then acts as a nucleation centre, where core protein monomers will be assembled into dimers and polymers to form an immature, RNA containing, nucleocapsid (Seeger et al. 2007; Schultz et al. 2004; Schädler and Hildt 2009).

The immature nucleocapsid (at step 8) undergoes maturation in the cytoplasm when the ε-signal is recognized as an origin of replication and the RNA begins the complex process of reverse transcription (step 9) to DNA. Once mature, the newly formed nucleocapsid can either be transported to the nucleus once again (step 3) and continue the replication cycle or will be exported from the cell (step 10) through interactions of the pre-S domains of the large surface protein with the endoplasmic reticulum and golgi-complex (Schultz et al. 2004; Schädler and Hildt 2009).

It is known that HBV, a virus that relies upon a reverse transcription step within its life cycle, is prone to the introduction of errors (Capobianchi, Giombini et al. 2013) and recombination (Bowyer, Sim 2000, Simmonds, Midgley 2005) most likely due to duel infections. This leads to a viral quasispecies that forms during

the course of infection with many minor population variants (Beerenwinkel, Günthard et al. 2012) which remain infective and are thus transmittable.

1.5 CLINICAL ASPECTS

HBV infections have become a public health problem worldwide, with approximately 2 billion people with markers of past infection and an estimated 240 million patients who are currently chronically infected (Seeger et al. 2007; World Health Organization 2013).

1.5.1 Laboratory Diagnosis

Diagnostic tests to determine HBV infection and monitor disease progression measure three viral components found in serum samples; (1) HBV DNA, (2) HBsAg (s-antigen), (3) HBeAg (e-antigen) as well as non-viral components such as antibodies to the respective antigens, including HBcAb, and host serum transaminase (ALT) levels. Of these markers, HBsAg, HBsAb and HBcAb are usually the primary markers screened to establish a diagnosis and core antibodies may be subtyped to distinguish between acute (IgM class) and chronic (IgG class). Secondary markers most frequently used are HBeAg and HBeAb; however ALT levels and molecular tests (quantitative PCR of HBV DNA/Viral load) may also be used. These markers vary in titre and may all but disappear, depending on the stage (acute vs. chronic) and the phase of persistent infection.

A data mining study conducted at the National Institute for Virology, South Africa (NIV; now renamed the National Institute for Communicable Diseases, NICD, a division of the NHLS) and reported recently (Bowyer et al. 2011) examined 39 774 HBV serology records, encompassing infections from 1985 to

1992 and calculated the frequency of the 8 (2^3) possible combinations (present or absent) of the 3 primary serological markers. The aim this study was to generate suitable commentary and all possible diagnoses for computer based laboratory reports (see table 1.1). For example, three (II, III and V) of the 8 primary screen combinations require a secondary screen and two of these (namely stage II and III) remain ambiguous although ALT and HBV DNA levels can assist in assessing the stage of disease.

Table 1.1: Profile and interpretation of primary and secondary serology markers

	Primary				Secondary		Interpretation
	HBsAg	**HBsAb**	**HBcAb**		**HBeAg**	**HBeAb**	
			IgM	**IgG**			
	-	-	-	-	-	-	Susceptible[1], Stage I[3]
	+	-	-	-	-	-	Early phase infection[1,2], Stage II[3]
	-	+	-	-	-	-	Immune (vaccination)[1], Recovery with loss of HBcAb[2] Stage VII[3]
Acute	-	-	+	-	-	-	Acute infection[2]
	+	-	+	+	+	-	Acute - Symptomatic phase[2], Stage III[3]
	-	+	+	+	-	+	Acute - Asymptomatic phase[2], Stage VI[3]
	+	-	+/-	+	-	+	Acute - Healthy HBsAg carrier[2], Stage III[3]
Chronic	+	-	+/-	+	+	-	Chronic - Persistent carrier[2], Stage III[3]
	-	+	+/-	+	-	+	Chronic - Recent past/convalescence[2], Stage VI[3]
	-	+/-	-	+	-	-	Chronic[2] - Distant past/recovery - HBeAg negative CHB - Occult - Stage V[3]

1. Based on primary markers (Center for Disease Control and Prevention 2012)

2. Based on primary and secondary markers (Previsani and Lavanchy 2002)

3. Based on primary markers (Bowyer et al. 2011)

During the normal course of infection, HBsAg titres usually decrease as the infection clears. In a small number of infections HBsAg levels become non-detectable, suggesting a cleared infection, however small quantities of HBV DNA are still found in the serum – associated with an established infection. These cases are classified as 'occult' infections. With the establishment of PCR amplification assays of the whole genome (Gunther et al. 1995) we are only now able to detect minute amounts of HBV DNA in persistent infections within the liver tissue, serum, peripheral blood mononuclear cells and other lymphoid tissues. Studies conducted on these cases showed mutations in both the Pre-S1/S2 as well as the S region, as a result of recombination between genotype A and D. In South Africa approximately 1 in every 4 067 blood donors present with an occult infection (Allain et al. 2009) which would have gone unnoticed had HBsAg been used as the sole serological marker for infection. The typical serology of different diseased states will be discussed in the subsequent sections.

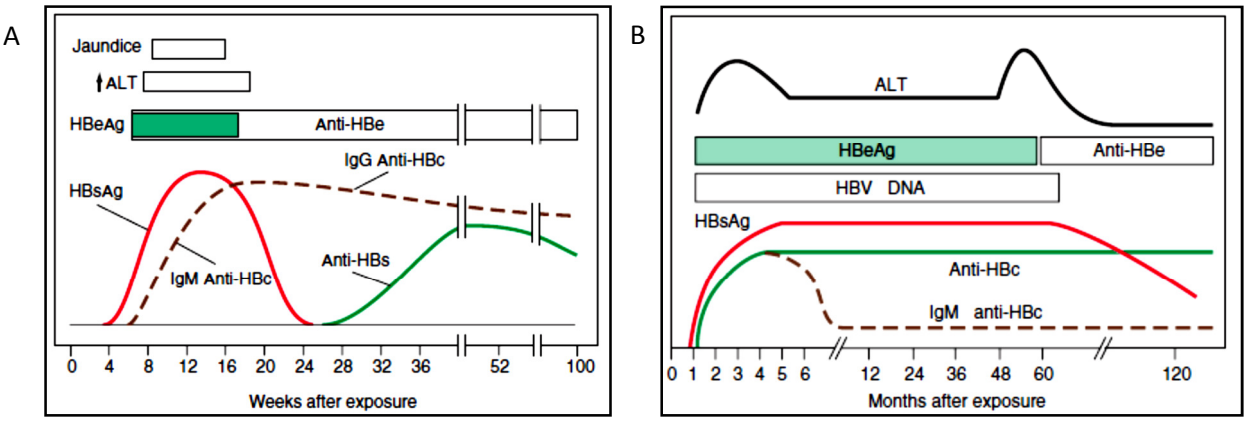
1.5.2 Disease states

1.5.2.1 Acute Viral Hepatitis B

Acute Hepatitis infections have a 1 month (4-6 weeks) to as long as 6 months incubation period after transmission as the virus spreads within the liver. In approximately 65% of acute infections the infection and resolution is clinically silent. Symptoms that are clinically recognized in the remaining cases include decreased appetite, nausea and vomiting, fatigue and abdominal pain as well as jaundice in the more severe cases. These symptoms most often result from

increased production of pro-inflammatory cytokines such as INF-γ or TNF-α (Seeger et al. 2007).

The first serological marker to become detectable during infection is the HBsAg, which usually becomes detectable at 8-12 weeks post-infection, assuming a 1 month incubation. This marker typically precedes an elevation of serum ALT levels and symptoms of hepatitis by 2 to 6 weeks and remains detectable throughout the symptomatic phase. After the onset of jaundice, HBsAg titres gradually decrease and usually and become undetectable after 2 to 6 months. Shortly thereafter antibodies against S-antigen (Anti-HBs) become detectable in the serum and may remain detectable indefinitely (Dienstag 2010). This algorithm is represented in figure 1.7 A.



↑ Increased ‖ Longer time period

Figure 1.7: Graphs representing the algorithms used when analysing the progressive serology in acute **(A)** and chronic **(B)** viral hepatitis (Dienstag 2010). For acute (A) the red line represents HBsAg levels which disappear around 24 weeks when Anti-HBs appears; the brown dash line represents Anti-HBc. For chronic (B) the red line indicates HBsAg levels which drop around 120 months post exposure whilst the green line represents IgG Anti-HBc and the brown dash line IgM anti-HBc.

HBcAg is not normally found in the serum as it is either intracellular in hepatocytes or sequestered within the virion. Anti-HBc is however detectable in the serum within a week or two after the appearance of HBsAg and remains for weeks to months before Anti-HBs is detectable. As is illustrated in figure 1.7 B, a class switch occurs in the immunoglobulins from IgM to IgG around 6 months post infection, where IgG becomes the main class of antibody detected whilst IgM levels wain to below detection limits. This feature can be used to differentiate between recent and more remote infections (Dienstag 2010).

A third serological marker, HBeAg, is readily detectable either concurrently or shortly after the S-antigen. This marker is associated with a period of high levels of virus replication, more circulating intact virions and detectable levels of HBV DNA in plasma samples. In self-limited cases, HBeAg levels decrease and become undetectable shortly after the characteristic peak in serum ALT activity. This coincides with the appearance of Anti-HBe and a period of lower infectivity with little to undetectable HBV DNA levels (Dienstag 2010; Hadziyannis and Vassilopoulos 2001).

The most severe cases of acute infection (± 0.1-1%) lead to complete liver failure and are termed fulminant hepatitis. This form of acute infection is serologically distinguished from others by a 100-fold increase in serum transaminase levels (ALT) in contrast to the 10-fold increase found in non-fulminant cases (Seeger et al. 2007). Of those acutely infected, 5-10% of adults, 90% of neonates and 25-30% of children will develop a persistent or chronic infection (Bowyer et al. 2011).

1.5.2.2 Chronic Viral Hepatitis B

Chronic Hepatitis B, or the persistence of HBsAg and HBV disease for more than 6 months, is host and virus dependant and presents in several distinct phases based on differing levels of viral replication and intensity of the immune response. Carriers experience an initial immune tolerant phase characterised by near normal levels of ALT, high levels of HBV DNA and both HBsAg and HBeAg positivity (Dienstag 2010; Seeger et al. 2007). This phase ends when the immune system matures (in younger carriers) or recovers and begins to control and clear the virus. The end of the immune clearance (or immune active) phase is often marked by HBeAg seroconversion when HBeAg levels become undetectable and Anti-HBe antibodies appear. This is considered a good clinical sign and marks the beginning of an inactive carrier state because high HBeAg levels are indicative of high viral replication and infectivity, whereas high Anti-HBe levels indicate a low level of viral replication with low to moderate infectivity (Seeger et al. 2007; Bowyer et al. 2011; Dienstag 2010).

In some cases, patients may fail to undergo seroconversion and remain in the immune active phase which is associated with an increase in ALT and high but variable HBV DNA titres. During this phase the virus causes more severe liver damage while the host immune system is unable to control the infection (Seeger et al. 2007). This eventually contributes to liver cirrhosis and hepatocellular carcinoma (Seeger et al. 2007; Kramvis 2008).

It should be noted, that the phases of chronic infection are not static and an active phase can move to an inactive phase and vice versa. One such “reactivation”, the

immune escape phase, occurs when infected individuals acquire HBV strains with mutations that prevent the expression of e-antigen. The most common mutations that stop HBeAg expression occur in the pre-C and Basal Core Promoter (BCP) region, first characterized in patients with genotype D from Mediterranean countries, and are acquired late in the natural history of infection (Dienstag 2011). The term HBeAg-negative chronic hepatitis carrier now has a wider geographic distribution and refers to all chronic carriers with hepatitis B with mutations which diminish or abolish HBeAg production (Hadziyannis and Vassilopoulos 2001).This phenomenon has been well studied and reported in South African Negroid populations who are infected with subgenotype A1 (Kramvis 2008; Araujo et al. 2011) but these reports have largely been ignored in global discussions on HBeAg-negative chronic hepatitis B (Funk et al. 2002) from regions where subgenotype A2 is prevalent.

In summary, these infections are characterised as being HBsAg positive, HBeAg negative and elevated ALT in the serum with $>10^4$ copies/mL but fluctuating HBV DNA (Seeger et al. 2007). This is referred to as HBeAg negative (with or without HBeAb) chronic hepatitis B and is largely considered to be a fourth or reactivation phase, associated with a larger degree of liver damage (Hadziyannis and Vassilopoulos 2001; Hadziyannis and Papatheodoridis 2006; Hadziyannis 2011).

Since mutations can confound routine diagnostics (Bowyer et al. 2011) and can lead to more aggressive disease which requires unique management strategies, HBeAg negative chronic hepatitis B infection and its causes clearly warrants further investigation at a molecular/sequence level to determine and better

characterise the various combinations of mutations underlying serological abnormalities.

1.5.3 Treatments

As of yet, seven drugs have been implemented as treatment for chronic hepatitis B viral infection. These agents are injectable Interferon-α (INF-α), pegylated interferon (PEG-INF-α) and the oral agents lamivudine, adefovir dipivoxil, entecavir, telbivudine and tenofovir (Dienstag 2011; Ayoub and Keeffe 2011).

1.5.3.1 INFα and PEG-INFα

INF-α was the first drug approved for the treatment of chronic HBV but has largely been replaced by PEG-INF-α as it is long-acting and dosing intervals can be increased from once a week to once every three weeks (Dienstag 2011; Billioud et al. 2011). The utility of immunomodulatory agents in treating HBV is however overshadowed by the vast amount of side effects which may include systemic “flu-like” symptoms, bone marrow suppression, emotional (irritability, depression, anxiety) and autoimmune reactions, alopecia, rashes, diarrhoea and numbness/tingling of the extremities (Dienstag 2011; Billioud et al. 2011; Wang et al. 2009; Ayoub and Keeffe 2011).

1.5.3.2 Lamivudine

The dideoxynucleoside lamivudine, the first nucleoside analogue to be approved, is a potent and effective agent for treating retroviruses such as HIV and HBV through inhibiting the reverse transcriptase enzyme. Treatment of HBV patients with lamivudine has been associated with HBeAg loss (32-33%), HBeAg

seroconversion (16-21%), normalized ALT levels (40-75%), improved histology (50-60%), delayed fibrosis (20-30%) and has been shown to prevent progression to cirrhosis (Dienstag 2011).

In spite of the markedly better tolerance and side effects profile, long-term monotherapy with nucleoside analogues such as lamivudine may lead to resistance mutations (Dienstag 2011; Billioud et al. 2011). Such mutations include a methionine to valine/isoleucine mutation at amino acid 204 (M204V/I) in the tyrosine-methionine-aspartate-aspartate (YMDD) motif of HBV DNA polymerase. These mutations occur in 15-30% of patients over the course of the first year on lamivudine treatment, and increases with each subsequent year of treatment to up to 70% in the 5th year (Dienstag 2011).

At present, the use of lamivudine in the USA and Europe has largely been replaced by more potent antivirals that have superior resistance profiles (Dienstag 2011; Ayoub and Keeffe 2011).

1.5.3.3 Adefovir Dipivoxil

Adefovir Dipivoxil, the prodrug of adefovir, is an acyclic nucleotide analogue. HBeAg positive patients on a 48 week course of adefovir showed a 23% loss of HBeAg, seroconversion in 12%, normalized ALT levels in approximately 50% (along with improved histology and reduced fibrosis) and plasma DNA levels below PCR detection limits in 13-21%. Those with HBeAg negative chronic hepatitis B, under the same treatment regimens, showed normalized ALT levels in 75% of the group, improved histology in 66% and suppressed HBV DNA in 50-66% (Dienstag 2011).

No incidence of resistance mutations has been reported within the first year of treatment, which is a great improvement upon the 15-30% of patients on lamivudine. There are however two mutations, asparagine to threonine (N236T) and alanine to valine/threonine (A181V/T), that occur at a rate of 2.5% and 29% after 5 years of therapy, respectively. As these mutations are located on a different part of the genome adefovir still remains an excellent candidate for treating lamivudine resistant strains (Dienstag 2011; Billioud et al. 2011; Ayoub and Keeffe 2011).

1.5.3.4 Entecavir

Entecavir is currently a first line treatment for patients with chronic HBV infections as it has a high potency with a corresponding high barrier to resistance while being just as well tolerated as lamivudine. This oral cyclopentyl guanosine analogue polymerase inhibitor requires both an YMDD mutation (seen in lamivudine resistance) as well as a second mutation at one of several sites (T184A, S202G/I or M250V) to establish resistance, which occurs in a mere 1.2% of patients after 5 years of treatment (Dienstag 2011; Ayoub and Keeffe 2011). Due to the shared YMDD mutation this drug is however not as attractive a choice for treating lamivudine resistant strains as adefovir or tenofovir (Dienstag 2011).

1.5.3.5 Telbivudine

Telbivudine, a cytosine analogue, closely resembles entecavir with regards to efficacy but is slightly less powerful in reducing HBV DNA. Telbivudine resistance is established by a M204I mutation which occurs less frequent after a

year of treatment than lamivudine, reaching an incidence of 22% after two years (Dienstag 2011).

1.5.3.6 Tenofovir

Another potent antiviral used in the treatment of both HIV and HBV is the acyclic nucleotide analogue tenofovir, with strong similarity to adefovir yet more effective at suppressing HBV DNA levels and inducing HBeAg responses. Treatment of HBV with tenofovir has been associated with HBeAg seroconversion in 21%, normalized ALT levels in 68% of HBeAg positive and 76% of HBeAg negative patients as well as suppression of HBV DNA to undetectable levels in 76% and 93% of HBeAg-positive and –negative patients, respectively (Dienstag 2011).

Tenofovir has no recorded resistance profile and a negligible side effects profile, making it a favourable first line treatment above adefovir for the treatment of chronic hepatitis B (Dienstag 2011; Ayoub and Keeffe 2011).

1.5.3.7 Current antiviral research

Current treatment for hepatitis B infection focusses on HBV DNA suppression to avoid the severe sequelae of cirrhosis and hepatocellular carcinoma and not on achieving HBsAg seroconversion, although some patients do eventually clear HBsAg (Gish and Adams 2009). Furthermore, since immunomodulators are poorly tolerated and nucleoside analogues pose a risk of resistance, research into novel antiviral agents targeting different steps of viral proliferation is ongoing. Present research mostly involves *in vitro* experiments in the HepG2 transfected

cell line as well as some *in vivo* work in either duckling (Duck Hepatitis B Virus [DHBV]) or chimeric mice models.

The first approach to pharmacological research, the synthesis of new compounds, has focused on creating agents that act by specifically inhibiting encapsidation by binding free HBV core particles. Such compounds include BAY-41 4109, a heteroaryldihydropyrimidine (Weber et al. 2002), and AT-61/AT-130, both of which are derivatives of phenylpropenamide (Perni et al. 2000).

BAY-41 primarily acts by impairing the formation of complete nucleocapsids and has proven efficacy in both *in vitro* (Deres et al. 2003) and *in vivo* (Weber et al. 2002) experiments, including strains with resistance mutations (Billioud et al. 2011). Phenylpropenamides, on the other hand, act by favouring protein-protein interactions to the detriment of pre-genomic RNA (pgRNA) – leading to the formation of empty nucleocapsids (Feld et al. 2007). Related studies have also been evaluating the utility of novel delivery systems such as artificial recombinant cell-penetrating peptides (Pan et al. 2011) as well as gene therapy by small interfering mRNA's (Giladi et al. 2003).

Another route of research, namely phytomedicinal or medicine of botanical origin, has also been successful at identifying extracts from traditionally used plants that show significant antiviral activity (Herrmann et al. 2011). Such agents include chlorogenic-, quinic- and caffeic acid (Wang et al. 2009) from the leaves and fruit of dicotyledonous plants such as coffee beans as well as HD-03/ES, an Indian herbal medicine (Kar et al. 2009). There are however several key issues that still need to be addressed by further research such as determining the specific

component in extracts responsible for the antiviral activity, possible cross reactivity between components (inhibition or synergism) and the actual action mechanism behind the observed antiviral effects.

Clearly, both avenues of scientific inquiry give great promise to providing novel, better tolerated and cost effective medicines with action mechanisms that may circumvent the vast genetic diversity and mutability of HBV, be it as mono- or adjunct therapy.

1.5.4 HBV Vaccine

The HBV vaccine consists of a yeast-derived recombinant HBsAg protein (Engerix-B) and is effective at producing protection in up to 95% of immunocompetent recipients (Sheldon and Soriano 2008; Keating and Noble 2003; Machida and Nakamura 1991).

The first vaccine experiments, observing immuno-protection elicited by immunization with short HBsAg irrespective of subtype showed effective protection (Purcell and Gerin 1975; Schaefer 2005). However, the majority of anti-HBsAg antibodies of the primary immune response were type specific. Immunization of *Homo sapiens* (Legler et al. 1983) and chimpanzees (Purcell and Gerin 1975) with SHBsAg of serotype *adw*, first gave rise to *d*-specific IgM antibodies. As this response was broadened, somatic hypermutation and epitope maturation lead to the inclusion of the *a*-determinant. Thus, for more rapid protection, HBV serotypes may be of importance (Schaefer 2005).

The first evidence of vaccine escape mutants was found in a study group vaccinated with genotype A in a region where genotype D was the prominent

circulating strain (Carman et al. 1990). Furthermore, it has also been found that vaccine escape was more commonly seen in cases with the *y* than the *d* determinant (Wong et al. 1984).

Vaccine escape mutants have two primary origins (see figure 1.8); the first is due to immune pressure and as an attempt to avoid immuno-surveillance and the second is due to drug resistance mutations in the overlapping *pol* gene (Yamamoto et al. 1994; Mimms 2005; Sheldon et al. 2007; Sheldon and Soriano 2008). Such mutations normally occur within the a-determinants' two antigenic loops spanning amino acids 121 to 147 and may abrogate the disulphide bridges where a Cys to Ser change occurs (Sheldon and Soriano 2008). Common mutations associated with immune evasion include a sG145R change (Yamamoto et al. 1994) while those associated with lamivudine therapy include a sE164D and sI195M change (Sheldon et al. 2007; Sheldon and Soriano 2008). Interestingly, mutations resulting from adefovir and tenofovir treatment aren't associated with detrimental changes in the antigenic loops of the surface protein (Sheldon et al. 2007).

While the HBV vaccine has been found to be both safe and efficient, researchers do suggest that vaccination be conducted with the subtypes that are predominant in a specific region (Schaefer 2005), taking all variants into consideration.

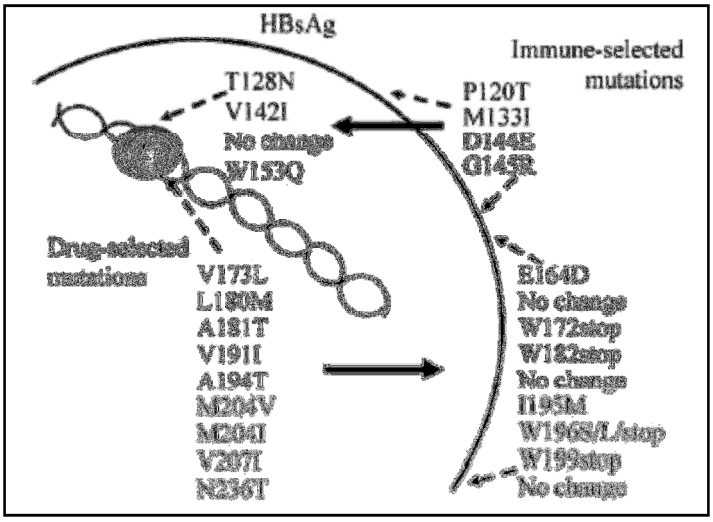


Figure 1.8: Amino acid changes in the surface and polymerase genes as a consequence of antiviral therapy or immune pressures (Sheldon and Soriano 2008).

1.6 GENOTYPES

The hepatitis B virus currently encompasses eight genotypes with several genotypes comprising multiple subgenotypes (Gerlich 2013; Mason et al. 2012). This vast genetic diversity is largely attributed to the lack of proofreading of the viral polymerase (Capobianchi et al. 2013) as well as recombination between established genotypes (Simmonds and Midgley 2005) and even non-human primate strains (Kurbanov et al. 2008). Although there is some correlation between serological subtypes or serotypes (1.3.5) and DNA genotypes, several s-antigen serotypes are represented in more than one genotype (Norder et al. 2004).

Studies on the nucleic acid sequence of HBV DNA has led to the classification of HBV into eight widely accepted genotypes or genetic subtypes, denoted A to H, based on pair wise differences >8 and <17% across the full genome (Seeger et al. 2007; Norder et al. 2004) according to the latest report from the International Committee on Taxonomy of Viruses (Mason et al. 2012). Another two genotypes,

I and J, have also been proposed (Kurbanov et al. 2008; Tatematsu et al. 2009). This classification system has become the standard reference nomenclature for distinguishing different strains of HBV. Each genotype, with the exception of E, G and H, can also be subdivided into subgenotypes based on pair-wise differences >4% but less than 8% and in the absence of evidence of recombination across the full genome.

The eight HBV genotypes (A-H) show a markedly conserved geographical distribution (see figure 1.9), which has been impacted by ancient human migration as well as more recent migrations such as the 15-16th century travellers (Kramvis and Kew 2007) and the slave trade (Andernach et al. 2009) In Africa, genotype A is found in southern Africa, including South Africa, Zimbabwe and Malawi. Genotype D, which is often found in co-infection with HBV A, is found throughout Africa but is most prevalent in North African countries on the Mediterranean. Genotype A and D coexist in South Africa (Kimbi et al. 2004). The HBV E genotype has a more restricted distribution that was limited to the West African countries such as Cameroon (Kramvis 2008; Kurbanov et al. 2005) but has also been identified in recent years in cohorts from Angola, Namibia and South Africa (Kramvis et al. 2005; Mayaphi et al. 2013).

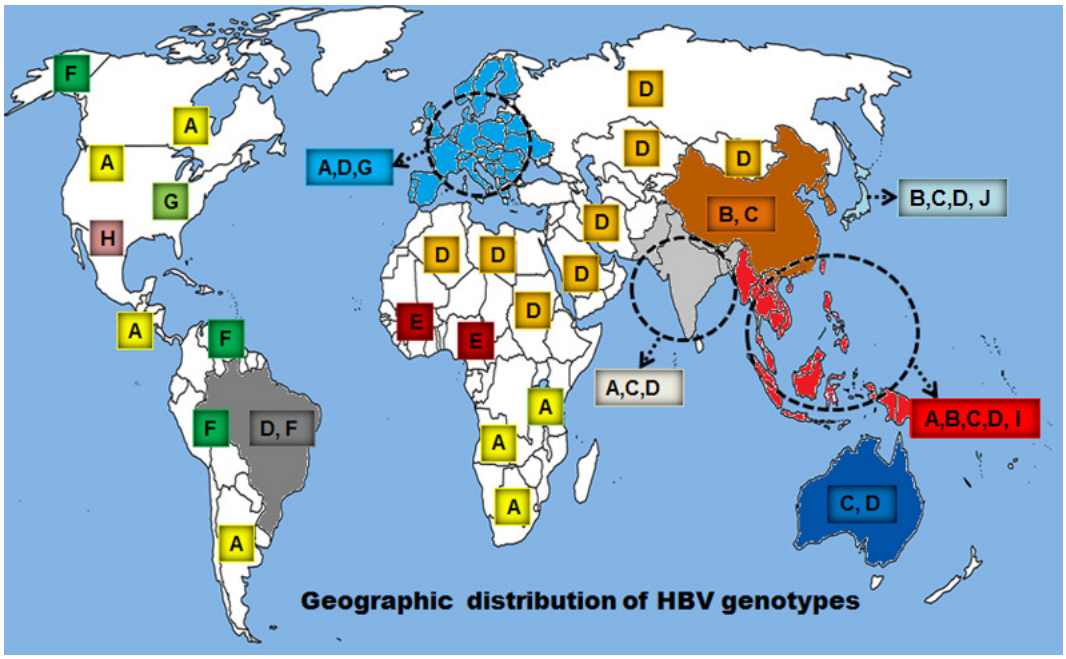


Figure 1.9: Geographic distribution of the main hepatitis B virus genotypes (A-H) as well as the putative genotypes I and J (Hussain 2013).

1.6.1 Genotype A

Strains belonging to genotype A predominate in Europe, India, North America and Africa. This, largely African, genotype is divided into seven subgenotypes denoted A1-7 (Gerlich 2013). The first two subgenotypes of HBV denoted A1 and A2 (originally designated A' and A (Bowyer et al. 1997)) were recognised and characterised by researchers in South Africa between 1997-2002 (Norder et al. 2004; Bowyer et al. 1997; Kramvis et al. 2002). Amino acids differentiating Genotype A into two subgenotypes based on the surface gene, are localised to the pre-S1 region which slightly overlaps the spacer of the polymerase encoding region (Kimbi et al. 2004). This division is based upon differences in the deduced amino acid sequences from the S-gene region where A1 encodes Asn207 and Leu209 and A2 encodes Ser207 and Val209 (Norder et al. 2004). These changes

do not result from an adaptive change under immunological pressure as they are also present in isolates from infected children and acute hepatitis patients (Kimbi et al. 2004).

The majority of genotype A isolates from South Africa were found to be of subgenotype A1 however some belonging to subgenotype A2, also known as the 'European' subgenotype, have been encountered. It is suggested that the A2 subgenotype might in fact have originated in South Africa and was transferred to Europe by 15th century sailors who visited the continent (Kramvis and Kew 2007).

The year 2005 marked the emergence of a newly characterised subgenotype A3. This subtype was first found in the Cameroon (Kurbanov et al. 2005) and later identified in cohort studies from both Mali and Gambia (Kramvis and Kew 2007). These studies have identified nine unique amino acid substitutions as well as evidence of possible recombination between genotype A and E (Kurbanov et al. 2005). Additional studies conducted on samples from Mali, Nigeria and Haiti lead to the discovery of another three subtypes, termed subgenotype A4, A5 and A6 (Olinger et al. 2006; Kramvis and Kew 2007; Andernach et al. 2009). The most recent addition to the HBV A genotype is subgenotype A7 which was isolated in Cameroon (Hubschen et al. 2011).

To date, HBV A1 remains the dominant and endemic subgenotype among South Africans (Kramvis and Kew 2007). Infection with subgenotype A1 is associated with markedly lower levels of HBV DNA in both HBeAg and anti-HBeAg positive phases when compared with HBV A2 or D infections (Kramvis and Kew

2007). Studies have also shown that HBeAg is lost very early in the infection of the anthropologically termed (Sauer 1992; Cartmill 1998; Hinkes 2009) Negroid race of South Africans and only 5% of those infected as children still having HBeAg in adulthood. This trend is not present in South African Caucasiods or Mongoloids or in any of the other areas of the world where HBV is hyperendemic (Kramvis 2008) but may be age and genotype dependant.

One mechanism underpinning HBeAg negativity is selection for mutation, previously described for samples from Greece, Italy and the Far East (Hadziyannis 2011). Here, a G to A mutation at position 1896 converts the TGG codon for Trp to a stop codon, resulting in a truncated precursor protein and a lack of e-antigen expression (Hunt et al. 2000). Although this mutation has been noted in African A1 isolates it is relatively rare (Hadziyannis and Vassilopoulos 2001). This is attributed to the fact that most genotypes have 1858T whilst genotypes A and F mainly have 1858C, which base pairs with 1896G by non-Watson-Crick interaction in the lower stem of the encapsidation signal (Lok et al. 1994; Kramvis and Kew 2002). Thus, a G1896A transvertion in other genotypes serves to stabilize the ε-signal when 1858T is present by creating a Watson-Crick base pair. Genotype A, which mainly has 1858C, already has a stable base pair and would thus first have to undergo a C1858T transvertion before the G1896A change is necessary (Lok et al. 1994).

A review published by Kramvis in 2008 highlighted three distinct mutations that lead to HBeAg negativity in South Africa. Firstly, an A1762T and G1764A mutation was observed (Okamoto et al. 1994; Hunt et al. 2000) that affects the basic core promoter region and, through reduced transcription of pre-core mRNA,

leads to a 60% decrease in e-antigen titres. A second mechanism is variation seen at the nucleotide kozak consensus sequence for positions 1809-1812. The observed TTT and TCT triple mutations at positions 1809, 1811 and 1812 severely impaired HBeAg expression whereas T_T and A_T double mutations caused a moderate reduction. Lastly, a common missense mutation was identified at position 1862 of the pre-core region which could interfere with the initiation of reverse transcription. Furthermore, the phenotypic change to Phe at codon 17 could interfere with signal peptide cleavage during the post-translational modification of the HBeAg precursor (Kramvis 2008).

The 1862 mutation has since been observed in a cohort study on Zimbabwean blood donors (Gulube et al. 2011), however the remaining two (1762T and 1764A) are still to be substantiated by further studies on separate cohorts from South Africa.

A second mechanism which introduces diversity to HBV is recombination between genotypes. Recombination has previously been documented for HBV and in general, “mosaic blocks of sequence identical to an alternate type or subtype within a specimen of established type is considered unequivocal evidence that recombination has taken place” (Bowyer and Sim 2000). Bowyer and Sim (2000) found mosaic sequences in 14 of 65 specimens. In these cases, genotype D contained mosaics of genotype A and genotype B contained mosaics of genotype C (Bowyer and Sim 2000). As previously mentioned, genotype A and D are often found as a co-infection (Africa) and genotype B is often co-infected with genotype C (China) (Bowyer and Sim 2000; Candotti et al. 2012; Fang et al.

2011). This may lead to the acquisition of mutations previously found in other genotypes.

Lastly, another explanation for the high prevalence of HBeAg negative infections in South Africa and globally may be due to false negative laboratory results. As previously mentioned, the c-antigen and e-antigen have marked sequence similarities which makes it hard to discriminate between the two proteins using immuno-assays. Despite similarities at the primary structural level, conformational differences in the fully assembled peptides do selectively mask or expose epitopes (Watts et al. 2010). A study conducted on the high degree of antigenic cross reactivity (Watts et al. 2010) between these peptides assessed the specificity and selectivity of six frequently used monoclonal antibodies (mAb).

This study noted that only a partial antigenic distinction can be made based on the assembly state of the peptides. Exceptions were found with two mAbs; one could only detect HBeAg/β epitope (residue 124-132) in non-polymerised HBeAg dimers and another could only detect the HBcAg/β epitope in assembled capsid dimers.(Watts et al. 2010) Furthermore, binding of one mAb to its epitope on HBeAg caused some stearic hindrance, preventing other mAbs from optimally binding to their respective epitopes. Thus, a false negative test for HBeAg could result from lack of selectivity between dimeric forms of c- and e-antigen or stearic hindrance when e-antigen is bound to host anti HBeAg IgG. Future studies will also need to access the degree of specificity and sensitivity of HBeAg mAbs to HBV mutants and recombinants (Watts et al. 2010) as studies on the effect mutation has on recognition of the surface antigen have shown a marked reduction

in antigenicity with resultantly high false negatives (Weber 2005). The crystal structure of HBeAg (PDB 3V6Z) was only recently determined (DiMattia et al. 2013) and the full effect of mutations within the pre-core/core region on the conformation of the HBeAg as well as changes in the binding energy for epitope regions can now be assessed.

1.6.2 Genotype B

Genotype B partitions into two major groups, the non-recombinant forms which cluster close to the genotype (formerly called) Bj from B Japan and the recombinant forms which partition with the genotype (formerly called) Ba from mainland Asia. The Bj group comprises subgenotype B1, the major genotype of B in Japan, and B6 which is prevalent in the Arctic region amongst indigenous populations of Alaska, Canada and Greenland (McMahon 2009; Kramvis et al. 2005). Group Ba comprises subgenotype B2, the major genotype found in China, as well as B3 (Indonesia), B4 (Vietnam) and B5 (Philippines). These subgenotypes contain portions of the genotype C genome recombined into the core region of the B1 genome (McMahon 2009; Norder et al. 2004). More recently, an additional two subgenotypes have been suggested; B7 from southern China (Shen et al. 2009) and B8 from Indonesia (Mulyanto et al. 2009). B7 appears to be a recombinant of B3, B4 and B5 (Shen et al. 2009), strongly questioning the validity of their subgenotype designation. Characteristic features of infection with genotype B include an early seroconversion from HBeAg positive to Anti-HBe in B1 and a T1762/A1764 double mutation in the basal core promoter region in the Ba recombinant group, associated with a higher incidence

of HCC (McMahon 2009). With regards to distribution, it is found that this genotype often occurs as a co-infection with genotype C.

1.6.3 Genotype C

HBV genotype C is partitioned into subgenotype C1 (Japan, Korea and China), C2 (China, Thailand and Vietnam), C3 (Pacific islands), C4 (Australian aborigines) and C5 (Norder et al. 2004; McMahon 2009; Kramvis et al. 2005). Two other subtypes, C6 (Lusida et al. 2008) and C7 (Mulyanto et al. 2009) have been observed in Indonesia, along with a string of other proposed subgenotypes C8-16 (Gerlich 2013). Subgenotype C2 commonly encodes Leu53 and Asn209 as opposed to Ser53 and Ser209 seen in the other genotype C strains. C3 differs from other strains by the fact that it lacks the Ile212 to Leu212 substitution seen in genotype C (Norder et al. 2004). Subgenotype C4 is by far the most divergent of the strains belonging to genotype C and differs from other subtypes by 5.9-7.4% across the complete genome (Kramvis et al. 2005). Patients infected with HBV genotype C experience HBeAg seroconversion at a much older age and are thus more likely to be HBeAg positive (McMahon 2009).

1.6.4 Genotype D

Genotype D is the most widespread HBV genotype globally but predominates in the Mediterranean, near East and India (Norder et al. 2004). This genotype has also been found in isolates from Indonesia, Papua, France, Germany, South Africa and the USA (Norder et al. 2004; Kramvis and Kew 2007). HBV genotype D can be divided into 6 subgenotypes; D1 (Middle East), D2 (India), D3 (South Africa and Alaska), D4 (Oceania, Somalia and South Africa) (Norder et al. 2004;

Mayaphi et al. 2013), D5 (India), D6 (Indonesia) and D7-9 (Lusida et al. 2008; Gerlich 2013). As a whole, none of the subtypes seem to carry unique amino acid substitutions(Norder et al. 2004) however the entire genotype is characterized by a 33 nucleotide deletion at the N terminal of the Pre-S1 region resulting in a loss of protein expression (Kramvis et al. 2005). From a clinical perspective, patients infected with HBV D strains typically undergo HBeAg seroconversion during adolescence or early adulthood and are thus more likely to be HBeAg negative and Anti-HBe positive due to an increased propensity for stop codon mutations in the pre-Core region (McMahon 2009).

1.6.5 Genotype E

This genotype was first described in 1992 (Andernach et al. 2009) and is mainly found in the “African genotype E crescent”, encompassing countries of west and into central Africa as far apart as Mali and Namibia (Andernach et al. 2009; Kramvis and Kew 2007; McMahon 2009; Norder et al. 2004). Surprisingly, HBV E shows a markedly low genetic diversity (1.75%) over the whole genome in spite of hyperendemicity (Andernach et al. 2009; McMahon 2009; Mulders et al. 2004) and thus does not divide into subgenotypes. Upon comparing genotype E with the other established genotypes it was found that genotype D and E do not partition separately in the X and Core ORFs (Kramvis et al. 2005; Bowyer and Sim 2000) and share the characteristic expression of Ser140 seen in genotype F (Norder et al. 2004). Beyond this, HBV E appears to be more closely related to non-human strains (Andernach et al. 2009) and very little is known about the influence of this genotype on disease outcome (McMahon 2009).

Unique features of this genotype includes a 3 nucleotide (1 amino acid) deletion at the N terminal of Pre-S1 (Kramvis et al. 2005; Kramvis and Kew 2007), signature amino acids Arg39, His45, Thr53, Met84, Lys86 and Thr109 and the introduction of an additional start codon Met83 in the pre-S1 region (Kramvis and Kew 2007). Additionally, the core region has the Golgi peptidase motif AsnThrTrp↓Arg upstream of the arginine region, instead of the ThrThrTrp↓Arg motif observed in all other HBV genotypes (Kramvis et al. 2005; Takahashi et al. 2000). This feature could be characteristic of a genotype E progenitor (Kramvis et al. 2005; Takahashi et al. 2000).

Due to the low genetic diversity of HBV/E it is hypothesised that this genotype was only introduced into the human populace more recently (Kramvis et al. 2005; Andernach et al. 2009). This is at least partially corroborated by studies tracking the spread of HBV strains during the slave trade from the early to the late 18th century (Andernach et al. 2009), which found that genotype E was only recently introduced into South American countries and was essentially absent from West Africa when and where slaves were assembled for transport.

1.6.6 Genotype F

Known as one of the “new world” genotypes, HBV F is mainly found in Central and South America. Strains originating from Central America share a characteristic T1858 and Thr45 in the surface genes while those from South America have C1858 and Leu45 (Kramvis et al. 2005; Norder et al. 2004). The complete genomic sequence of genotype F differs from that of other genotypes by approximately 14% and isolates from different geographical regions seem to

separate into four clusters (I-IV), with two clades within cluster I (Kramvis et al. 2005). Further studies reclassified genotype F into the following four subgenotypes: F1 (cluster I, Central America), F2 (clusters II and IV, South America), F3 and F4 (Kramvis et al. 2005; McMahon 2009). The remaining cluster (III) has been reclassified as a distinct genotype, HBV genotype H (Kramvis et al. 2005). The T1858C substitution in the wild type of subgenotype F2 does not favour the common G1869A Pre-C stop mutation which means patients infected with this subgenotype usually have a better clinical outcome than those infected with subgenotype F1 (McMahon 2009).

1.6.7 Genotype G

The existence of genotype G was first reported in 2000 (Stuyver et al. 2000) when the new genotype was characterised from samples collected in France and the USA and has since been isolated in Germany (Vieth et al. 2002). A characteristic feature of this genotype is a 36-nucleotide insertion at the 5' end of the Core region making it 3248 bp long and a resulting 24 kDa (12 extra amino acids) core protein instead of the usual 21 kDa. Due to this insertion, the pre-C region has two translational stop codons; one at codon 2 (TAA instead of CAA) and another at codon 28 (TAG instead of TGG). These mutations prevent the synthesis of the HBeAg (Vieth et al. 2002).

Paradoxically, HBeAg has been detected in the serum samples of patients infected with HBV genotype G (Stuyver et al. 2000; Vieth et al. 2002). Vieth *et al.* suggested that this can be ascribed to the detection of non-particulate C-protein that shares epitopes with the e-antigen or the more likely alternative, that HBeAg

is produced by a fraction of coinfecting HBV with intact pre-C regions. This escapes detection by direct (Vieth et al. 2002), Sanger sequencing but could be detected by Next Generation Sequencing (NGS).

In a NGS sequencing study conducted by Beck *et al.* to profile circulating DNA of pathogens in the plasma of 51 apparently healthy volunteers, one sample was found which contained HBV DNA (Lo and Chiu 2009; Beck et al. 2009). This study failed to explore whether NGS data correlates to that seen with traditional cloning and sequencing methods but it (and others) clearly illustrates the possible application of ultra-deep sequencing in virology research and future diagnostics (Lo and Chiu 2009).

HBV genotype G strains isolated in the USA and Canada were found to be present in co-infection with HBV genotype A only (Kato et al. 2002; Osiowy et al. 2008). As genotype A1 is hyperendemic in South Africa the possibility of co-infection with genotype G and the possibility of recombination between the two remains to be explored.

Thus far, data mapping the distribution of genotype G is very limited. However, one case of HBV G being isolated in Africa has been reported in the PhD dissertation of Lukhwareni (2008) and the sequence data of the Polymerase region submitted to Genbank (accession number EF619364)(Lukhwareni 2008). This is the first report of genotype G in South Africa but unfortunately it has not been published. Ideally, full genome sequencing should be performed to better understand the phylogenetic relatedness of this isolate to the American and

European isolates. Also of interest would be studies reporting the rates of co-infection between HBV genotype A1 and G in sub-Saharan Africa.

1.6.8 Genotype H

Strains belonging to this genotype were initially grouped with HBV genotype F, cluster III (see above) (Kramvis et al. 2005) but upon analysis of the complete genome were found to differ from that genotype by 7.5-9.6% and was thus designated as a new genotype (Arauz-Ruiz et al. 2002; Kramvis et al. 2005). Genotype H has mostly been identified in samples from Central American countries such as Mexico and Nicaragua (McMahon 2009) and is characterised by two unique amino acid substitutions, Val44 and Pro45, as well as other amino acid substitutions including Ile57, Thr140, Phe158 and Ala224 (Norder et al. 2004) of the surface gene. As this genotype is most closely related to HBV genotype F, it is believed that genotype H evolved from genotype F after it was established in the new world (Arauz-Ruiz et al. 2002; Kramvis et al. 2005; McMahon 2009). To date, there is very little information relating this genotype to disease outcome (McMahon 2009).

1.6.9 Genotypes I and J

"Due to the lack of universally accepted rules, irregularities have accumulated within the last 20 years of HBV genotype research" (Schaefer et al. 2009). This is evident within the accepted subgenotypes as well as the proposed genotypes I and J which remain controversial. A new genotype – designated genotype I - which was described in 2008 (Tran et al. 2008), was soon recognised to be the same as the recombinant described eight years earlier by Hannoun et al (2000).

Recombination is fairly common in HBV and as the tentative genotype I is a complex recombinant between several human and even gibbon HBV sequences, the designation genotype has been questioned by leading experts (Schaefer et al. 2009; Kurbanov et al. 2008). However, as more and more authors report finding strains belonging to the putative genotype I the call for recognition is getting stronger. HBV genotype J, which was characterized in a single patient, appears to be a recombinant of subgenotype C4 and non-human primate strains (Tatematsu et al. 2009) and does not yet have the same support.

1.7 GENOTYPE VS CLINICAL OUTCOME

Evidence is mounting that the patterns in global distribution of genotypes may be responsible for the differences observed in clinical outcome, response to anti-viral treatment and vaccine efficacy (Araujo et al. 2011). Furthermore, the persistence of HBV when serological tests for both e- and s-antigen are negative has many implications. Among these is: the reactivation of liver disease when patients become immunosuppressed, transmission of infections through blood and/or organ donations and an increased risk for developing cirrhosis or hepatocellular carcinoma (Owiredu et al. 2001).

1.7.1 Role of genotypes in disease progression to HCC

A study (Zhu et al. 2010) which looked at the whole genome sequence of HBV to identify mutations strongly associated with the development of HCC mapped several distinct mutations. Among these were the A1762T and G1764A mutations, previously identified in the South African population by Kramvis (2008), which showed a statistically significant correlation to developing HCC

(Zhu et al. 2010; Kramvis 2008). Zhu et al. (2010) identified five additional highly prevalent mutations in the Pre-C/C region (G1899A, C2002T, A2159G, A2189C and G2203A/T) in HCC patients, three of which lead to non-synonymous change at the amino acid level. Two deletions, one involving region 1793-1819 which codes the Core Promoter/X and Pre-C region and the other involving the Core region (2155-2229), were also observed (Zhu et al. 2010).

In South Africa the risk of developing hepatocellular carcinoma is 4.5 times higher in South Africans of the Negroid race infected with subgenotype A1 as compared to those resulting from non-A infection (Kramvis and Kew 2007). The high incidence of early loss of HBeAg, establishing HBeAg negative infections in South Africans (Tanaka et al. 2004) infected with genotype A1, is due to mutations other than G1896A (discussed fully in section 1.5.2.2 Chronic Viral Hepatitis B and 1.6.1 Genotype A), some of which are present in the wild type although not all the mutations have been fully characterized (Kramvis 2008). Zhu et al. (2010) clearly established the association of pre-S deletions and point mutations that may lead to e-antigen negativity and the development of HCC. The presence of pre-S deletions and numerous point mutations have been described in subgenotype A1 (Gopalakrishnan et al. 2013). These include: pre-S: T53C which results in F22L; wild type G1862T (Kramvis et al. 1997; Kramvis et al. 1998) and the basal core promoter pair of mutations A1762T/G1764A (Gopalakrishnan et al. 2013).

Only two of the mutations reported by Zhu et al. (2010) have been identified in South African cohorts. This can, at least partially, be explained by the fact that genotype B and C predominate in China (Zhu et al. 2010; Candotti et al. 2012;

Fang et al. 2011), whereas to date genotype A1 and D are the major endemic strains in South Africa.

1.7.2 Role of genotypes in response to therapy

Several studies have explored the possibility of genotype related responses to particular antiviral therapies. One such study, observing the response to IFN-α therapy by different genotypes (Kao et al. 2000), compared genotype B and C for their individual responses. At the end of a 72 week follow-up period 41% of patients with HBV/B and 15% of patients with HBV/C had normalized ALT levels, seroconversion of HBeAg and seroclearance of HBV DNA (Kao et al. 2000). In terms of those patients which initially presented with high ALT levels, patients in the genotype B group had a significantly higher response rate as compared to those of genotype C (50% vs. 17%). Furthermore, genotype B patients tended to have an increased rate of sustained biochemical and virological response (41%) than those of genotype C (15%) (Kao et al. 2000). Results from this study thus clearly established the existence of differential responses to antiviral therapy with immune-modulators, depending on genotype.

Another study (Lau et al. 2005) reported similar rates of seroconversion in HBeAg between genotypes B and C, but a slightly higher rate for genotypes A and D (Raimondi et al. 2010). Other studies (Janssen et al. 2005; Flink et al. 2006) reported a higher probability of HBeAg loss in genotype A as compared to C and D, whilst being higher for genotype B than C (Raimondi et al. 2010). When comparing composite endpoints (HBeAg seroconversion + PCR negativity and HBeAg seroconversion + PCR negativity + ALT normalization), genotypes A and

B generally responded better to therapy with IFN-α than C and D (Raimondi et al. 2010).

With regards to mono-therapy with the nucleotide analogue lamivudine, Zöllner et al. (2002) reported a twenty fold increase in the risk of selection for resistance between genotype D and A in a study observing the subtype specific response (as encoded by different genotypes). Other, long-term studies have reported a slightly higher risk during the first year of therapy in genotype A patients however; this difference seemed to decrease when therapy was prolonged to 2 or 3 years (Kramvis and Kew 2005).

Two studies (Marcellin et al. 2008; Liaw et al. 2009) compared genotype specific responses to different nucleot(s)ide analogues. The first study (Marcellin et al. 2008), reported a higher incidence of histological improvement in patients treated with tenofovir versus those on adefovir for all genotypes except B, with the highest difference between treatment groups being seen in genotype A for all end points (Raimondi et al. 2010).

The GLOBE study (Liaw et al. 2009), which compared different responses to lamivudine and telbivudine, found the latter to be more effective in establishing HBeAg seroconversion for carriers of genotype C, whilst no difference was observed in other genotypes. Several studies aimed at evaluating genotype related responses to nucleot(s)ide analogues have however concluded that there is no significant difference attributed to genotype (Moskovitz et al. 2005; Raimondi et al. 2010) yet some have noted genotype-specific mutations in the polymerase gene, possibly associated with resistance (Mirandola et al. 2012).

Whether or not these genotype specific responses relate to actual differences at the DNA level or are related to differing baseline characteristics/tendencies remains to be clarified. Clearly, further studies on the different genotypes and subgenotypes of HBV and how this relates to clinical consequences in Africa is warranted. The reason for this is three-fold. First, HBV is hyperendemic in sub-Saharan Africa. Second, genotypes previously considered to be limited to Africa are emerging worldwide as far apart as Spain and the United States of America where vaccination may not be strictly enforced and given the delayed response in developing immunity based on the *a*-determinant (Legler et al. 1983) along with the possibility of amino acid substitutions due to immune pressure or therapy causing changes in the overlapping *Pol* gene, may seriously hamper global efforts to eradicate the virus. This is particularly noteworthy as the majority of recombinant vaccines are derived from HBV/A2 and breakthrough infections with non-A2 strains have been identified in previously vaccinated individuals (Stramer et al. 2011). Third, in depth analysis of isolates from Africa could provide valuable insights regarding the origin and evolutionary patterns of HBV both in global populations and closed populations of HBV where there is a 'founder effect' complete with its own unique genotype, subgenotype, recombinants and mutants (Kramvis and Kew 2007).

1.8 NEXT GENERATION SEQUENCING

Studies on the sequence of the approximately 3.2kb DNA genome of HBV have enabled virologists to classify the virus into one of eight (A-H) genotypes (Seeger et al. 2007; Gerlich 2013; Mason et al. 2012). These genotypes, with the exception of HBV/E, HBV/G and HBV/H, are further subdivided into

subgenotypes (Gerlich 2013) and several relevant substitutions or mutations have been noted. This was all possible due to the, at their time, ground-breaking methodologies of DNA amplification (Mullis et al. 1986) by the polymerase chain reaction and DNA sequencing (Sanger et al. 1977) by Sanger/first generation chain terminating sequencing, which was improved upon by the addition of base-specific fluorescent dye molecules and capillary electrophoresis (Radford et al. 2012).

The main restrictions in the use of these technologies are low throughput, cost and labour when applied to larger fragments/genomes and the frequent reliance upon prior sequence knowledge for template specific amplification by PCR or clonally derived in bacteria (Radford et al. 2012). A further limitation is bias introduced in cloning and the problem that only major viral populations/variants are detected (Chevaliez et al. 2012).

Novel DNA sequencing technologies, collectively termed “next-generation” sequencing (NGS), have emerged since 2005 which enable high speed as well as high sample throughput and can generate a vast amount of sequence data from a single specimen. Perhaps the largest advantage of NGS is the determination of sequence data from one sample without cloning (Barzon et al. 2011) or the need to design template specific sequencing primers.

The first NGS platform was the 454 FLX (Roche) which became commercially available as of 2005. This was followed by the Illumina platforms (MiSeq, HiSeq etc.), SOLiD (Applied Biosystems), Heliscope (Helicos), Ion Torrent PGM (Life Technologies) and PacBio RS (Pacific Bioscience). Differing platforms have

different sequencing chemistries and differing protocols. As a result each platform comes with its own pros and cons with regards to suitability to specific applications based on read lengths, error rates etc. (Barzon et al. 2011).

All of the above technologies include the steps of template preparation, sequencing and imaging. A comprehensive overview of each of the sequencing technologies was recently published in a review by Radford et al. (2012). The Illlumina MiSeq, Hiseq and Genome Analyser systems are currently dominating the NGS market (Beerenwinkel et al. 2012). Illumina technology, contrary to the emulsion technology of 454 (Roche), relies upon solid phase amplification on a cartridge with a lawn of primers to which the template anneals after pre-processing. The latter includes adding adaptors to the library of fragments generated by enzymatic cleavage which are complimentary to the sequencing primers on the solid phase. This pre-processing step is significantly less damaging and cheaper than the nebulizing process used in 454 (Radford et al. 2012).

Furthermore, although all NGS platforms may introduce sequencing errors, Illumina platforms deliver reads with a comparably lower error rate (10^{-2} to 10^{-3}) and are less susceptible to indels in homopolymeric regions while indels outside these regions have similar frequencies of artificial indels and substitutions (Barzon et al. 2011; Beerenwinkel et al. 2012). The main sources of error in Illumina reads are; signal interference from neighbouring clusters, homopolymers, phasing and low coverage of AT rich regions (Barzon et al. 2011).

A recent study that compared three sequencing platforms (Quail et al. 2012) for the quality of reads generated for chromosome 11 of *P. falciparum* found that

most reads on Illumina had a phred score of greater than 30 (Beerenwinkel et al. 2012; Chevaliez et al. 2012) with an observed error rate of 0.80%, less than half of that observed on other platforms.

Next generation sequencing promises to be particularly useful to both basic and clinical research. To clinical research fellows NGS enables the detection of pathogens as well as initial drug sensitivity screening and therapeutic monitoring (Chevaliez et al. 2012) as has been done for several viruses including HBV (Nishijima et al. 2012). For the basic virology researcher this technology enables the *de novo* detection and/or re-sequencing of the entire viral quasispecies, inclusive of major (>20%), intermediate/low frrquency (5-20%) and minor (<1%) variant populations (Chevaliez et al. 2012; Beerenwinkel et al. 2012). Ultra-deep mapping has been used to reconstruct the full genome for HIV (Vrancken et al. 2010), influenza A (Kampmann et al. 2011), human rhinovirus (Tapparel et al. 2011), herpes simplex virus 1 (Szpara et al. 2010) and several enteric viruses.

Due to the novelty of the application of NGS technologies in studying the highly variable nature of viruses, very few programs exist that specifically apply algorithms for variant reconstruction. Of the handful of programs publicly available, the most notable ones are ShoRAH, ViSpa, QColors and QuRe (Prosperi and Salemi 2012) – all of which rely upon read graphs which condense and store the sequence data and eventually act as the main source of data for reconstruction (Beerenwinkel et al. 2012).

All the above mentioned programs are run from the command prompt and are mostly implemented in programming packages such as BioPython and Perl

originally written for machines running on the Linux operating system and need to be compiled from source codes which requires and assumes extensive bioinformatics knowledge which most virologists have not yet obtained. The single exception is QuRe (Prosperi and Salemi 2012), identified as part of an extensive literature search, which can be implemented in Windows using the Java Development Kit Standard Edition (http://jaligner.sourceforge.net/) and run from the DOS command prompt. Interestingly, QuRe also has improved optimization procedures for finding the quasispecies that minimize the number of *in silico* recombinants which has been proven using simulated and real NGS experiments (Prosperi and Salemi 2012; Beerenwinkel et al. 2012). Also, QuRe's algorithm applies a correction for homo-polymeric as well as hetero-polymeric sequencing errors which is one of the anticipated errors from Illumina platforms (Barzon et al. 2011).

CHAPTER 2

RESEARCH METHODS

2.1 INTRODUCTION AND PROBLEM STATEMENT

A recent study in the research laboratory of the Department of Medical Virology, University of Pretoria, which observed the impact of HIV on HBV infection in South Africa, reported a threefold greater prevalence of HBV in the HIV infected cohort compared to the HIV negative controls (Mayaphi et al. 2012). The subsequent genotyping of all HBV positive specimens by sequencing and phylogenetic analysis of the PreC/core gene sequence (using neighbor joining inference) revealed that, as expected, the majority of the specimens partitioned with well characterized Genbank reference sequences in known clades-including several South African sequences which clustered with Asian references. However, a number of the specimens partitioned away from the typical subgenotype A1 clades with high bootstrap values and long branch lengths. Further analysis using a Bayesian inference approach identified a significant clade (with a posterior probability of 1); partitioning away from all other African subgenotype A1 specimens and references on Genbank. Another, previously uncharacterized, separate clade (also with a posterior probability of 1) was also found within subgenotype A1 however a Genbank BLAST search identified a few relevant reference sequences for comparison in the study (Mayaphi et al. 2013).

All of the 25 HBV DNA-positive specimens were amplified successfully using the core primers whereas only 11 (44%) of the samples could be amplified using S gene primers. Since the core primers were used for specimens that did not amplify with the surface primers most have low to very low viral loads and data is only available in both regions for three of the seven most interesting specimens. Protein analysis showed interesting change in the core specimens at the nucleotide level and a frequency plot of the genetic distance of these specimens which was backed up by a BLAST search showed that there was only a 95% sequence similarity – which is less than the >96% intra-subtype difference expected within a subtype – between the best match and our study variants. The surface variants did not share common variation.

Additionally, RNA viruses such as the hepatitis C virus and influenza virus, and reverse transcriptase dependent viruses such as HBV and HIV, show high intra-host variations. This is likely due to the high replication capacity yet low fidelity (lack of proofreading activity) of the viral polymerase which results in between 10^{-5} and 10^{-3} substitutions per site per cycle (Capobianchi et al. 2013). This variability within the host is variably referred to as the mutant cloud, mutant swarm or viral quasispecies. As the population dynamics can't be understood from the fittest strain alone–because selection acts on the entire population (Beerenwinkel et al. 2012)–and low to minor frequency variants (<5%) can only be detected by next generation sequencing technologies (Chevaliez et al. 2012), ultra deep sequencing must be used to characterize virus populations within selected strains of interest.

This study will examine the full genome of selected specimens to better clarify the diversity observed in the core and surface sub-genomic fragments and broaden our knowledge of HBV genotypes, subgenotypes, and quasispecies of subgenotype A1, subgenotype D4 (Mayaphi et al., 2013) and a genotype E specimen which caused a recent outbreak in Pretoria. Results will be assessed and related to their possible use in the prognosis/diagnosis and HBV disease in the area.

2.2 AIM AND OBJECTIVES

The aim of the present study is to characterize the full genome of unique, atypical laboratory specimens as well as rare or unusual genotypes of hepatitis B identified in an urban cohort from a secondary referral hospital in Pretoria, South Africa. Specific objectives are:

a. To establish a full genome PCR and sequencing assay using optimised methods and primers on a typical African subgenotype A1 specimen from a patient with a high viral load.
b. Use this assay to perform full genome sequencing on unusual specimens identified in previous studies by PCR and sequencing of the core and surface regions.
c. Mapping known and unknown change (from this study) onto a linear template reference genome mapped on to Genbank specimen X02763 (appendix A) with numbering commencing from the EcoR1 site.

d. Phylogenetic analysis of sequences to determine the prevalence of genotypes in the cohort and detailed characterization of these specimens.

e. Further analysis of the variation and the possible role of recombination in different parts of the genome including epitope regions within this this sequence data and assessment of known and unknown variation for its clinical relevance.

2.3 MATERIALS AND METHODS

2.3.1 Samples

Samples to be used for the purposes of this study were selected from cohorts used in previous study (UP 35/2007), which observed the impact of HIV on HBV infection in South Africa (Mayaphi et al. 2012; Mayaphi et al. 2013). These samples, both plasma and serum, were collected from participants recruited at the Tshwane District Hospital HIV Clinic and serological screening was performed to establish if HBV co-infection was present (Mayaphi et al. 2012). A total of 20 samples (Table 2.1) were selected for purposes of the present study. These samples presented as outliers from reference clades in standard phylogenetic analyses, yet clustered together with significant bootstrap values (Mayaphi et al. 2013). Approval for the use of these samples in the present study was obtained from the student ethics committee of the faculty of health sciences (S 137/2012).

Table 2.1: Samples selected for the present study with primary and secondary serology marker results.

	ID	HIV	HBsAg	Anti-HBs	Anti-HBc	HBeAg	Anti-HBe	ALT (U/L)	Viral load (IU/mL)
1	3791	+	+	-	+	+	-	58	>110 x 10^6
2	N199	-	-	-	+	-	-	30	99
3	3269	+	+	-	+	-	-	21	95
4	N005	-	+	-	+	-	+	21	2 x 10^3
5	3319	+	-	-	+	-	-	35	127
6	4070	+	+	-	+	+	-	46	>110 x 10^6
7	4312	+	+	-	+	+	-	61	7 x 10^6
8	3274	+	+	-	+	+	-	60	10 x 10^6
9	N011	-	+	-	+	+	-	28	>110 x 10^6
10	3658	+	+	-	+	-	+	29	17 x 10^6
11	3678	+	-	+	+	-	-	16	290
12	3358	+	+	-	+	-	-	ND	ND
13	3768	+	-	-	+	-	-	ND	ND
14	3354	+	+	-	+	+	-	31	33 x 10^6
15	N060	-	+	-	+	-	+	44	157
16	LA03	+	-	-	+	-	-	ND	ND
17	LA05	+	+	-	+	-	+	ND	ND
18	LA06	+	+	-	-	+	-	ND	ND
19	LA09	+	-	-	+	-	+	ND	ND
20	PO04	-	-	-	+	-	-	ND	ND

ND = not determined/available

2.3.2 DNA Extraction

HBV DNA was extracted from plasma samples on the MagNA Pure LC™ (Roche Diagnostics, Mannheim, Germany) with the MagNA Pure total NA™ extraction kit (Roche Diagnostics, Mannheim, Germany) according to manufacturer's instructions.

A sample volume of 1000 µL was used and a final elution volume of 100 µL. Extracted samples were aliquoted into 4 Eppendorf tubes (25 µL each) of which one was stored at -4°C and the remaining three at -20°C to minimize freeze-thaw damage to DNA in the working extract.

Where only small sample volumes, with a relatively low viral load, were available, the QIAamp MinElute™ Virus Spin kit (Qiagen GumbH, Hilden, Germany) was used according to manufacturer's instructions. A final elution volume of 20 µL was used.

2.3.3 PCR Amplification & Agarose Electrophoresis

Amplification (Mullis et al. 1986) for full-length HBV genomes was performed using the method first described by Günther et al. in 1995, with some modification, by means of the Expand high-fidelity™ PCR assay (Roche Applied Science, Mannheim, Germany). A "hot start" method was used by preparing two separate mixes of which the first consisted of 200µM dNTP's (Thermo Scientific, Waltman, USA), 300nM of forward and reverse primer (IDT, Coralville, USA) respectively, 1x Expand Hi-Fi buffer (with 1.5mM $MgCl_2$), 5µL of extracted NA and molecular grade ddH_2O to a final volume of 20µL. The second mix contained 1x Expand Hi-Fi buffer, 2.6 Units of Expand Hi-Fi enzyme and molecular grade ddH_2O to a final volume of 5µL.

Based on multiple pairwise alignments of references, the primer regions were checked to verify conservation and one forward primer (P1) as well as three reverse primers (P2) designed (Table 2.2) to accommodate differences between subgenotype A1 and other genotypes. These primers were also checked with OligoAnalyzer 3.1

(IDT; http://eu.idtdna.com/analyzer/applications/oligoanalyzer/) for the potential to form homo- or hetero-dimers as well as hairpin/loop structures that may interfere with efficacy. This is screened by assessing the ΔG where a value of > -5 kcal.mole^{-1} is ideal.

Thermal cycling was done on the Px2 (Thermo Scientific, Waltman, USA) thermal cycler with an initial denaturation step at 94ºC for 2 minutes (mix 1) after which samples were cooled to 58ºC before adding the enzyme containing mix. This was followed by 40 cycles of denaturation at 94ºC for 40 seconds, annealing at 55ºC for 90 seconds (increasing by 1.25ºC every ten cycles) with a 1½ min elongation at 68ºC (increasing by 2 minutes every ten cycles).

Samples were initially subjected to a PCR reaction with primers that selectively amplify all known genotypes excluding subgenotype A1 to screen for infections with other genotypes that may occur less frequently or as a co-infection with the typical HBV A1. Known positive samples that did not amplify with the P1/P2 primer set were amplified with a PCR primer set specific for A1 namely the P1/P2_A1 primer set. Difficult to amplify samples were further tested with the degenerate P2_RM primer and/or a different Taq DNA polymerase such as Q5 high-fidelity (New England Biolabs, Ipswich, MA, USA).

Five micro liters of each of the PCR products were separated with a 0.9% Seakem® LE Agarose (Cambrex Bio Science, Rockland, ME, USA) – TBE (Sigma-Aldrich, St. Louis, MO, USA) gel by electrophoresis (170 V; 50 min) together with a 1 kb Gene

Ruler™ DNA ladder (Thermo Scientific, Waltman, WY, USA) in order to confirm the success of the amplification.

Table 2.2: Primers used for PCR

Primer		Positions	Sequence	Length	Tm*
P1	(forward)	1821--1841	5'- CTT TTT CAC CTC TGC CTA ATC A -3'	22	60.2
P2	(reverse)	1825--1806	5'- AAA AAG TTG CAT G**G**T G**C**T GG -3'	20	61.6
P2_A1	(reverse)	1825--1806	5'- AAA AAG TTG CAT G**A**T G**A**T GG -3'	20	56.9
P2_RM	(reverse)	1825--1806	5'- AAA AAG TTG CAT G**R**T G**M**T GG -3'	20	69.3

Table indicating the single forward (1) and three reverse (2) primers used in the study along with their genomic coordinates and melting temperatures for the study PCR chemistry (*)

Gels were pre-stained with ethidium bromide (Sigma-Aldrich, St. Louis, MO, USA) at a final concentration of 5 µg/mL. Hereafter, gels were visualized on the Gel Doc XR™ imaging system (BioRad, Hercules, CA, USA) and data captured using the provided Quantity One™ (BioRad, Hercules, CA, USA) software. A linear regression analysis was performed using the Gel Analyzer 2010 software package to confirm that observed amplicons fell within the expected size range and to reveal the presence non-specific amplification.

2.3.4 PCR Clean-Up

Prior to further experiments, the PCR amplicons were purified with the DNA Clean & Concentrator™-25 (Zymo Research, Irvine, CA, USA) kit according to manufacturers' instructions, eluting to a final volume of 20 µL in TE elution buffer.

2.3.5 Next Generation Sequencing

After PCR, the amplified and cleaned samples along with relevant controls were sent to Inqaba Biotechnical Industries (Pty) Ltd (Sunnyside, Pretoria, RSA), for next generation sequencing on the MiSeq™ sequencer (Illumina, San Diego, CA, USA). This would enable the detection of up to 10 full genome variants per sample.

Briefly, a fragment library was created for each specimen using the Illumina Nextera®XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) according to manufacturer's instructions. This involved the tagmentation of input DNA with the Nextera® XT transposome which both fragments the sample into 150-250 bp length segments and then ligates adapters to either end of each fragment for subsequent library amplification. After amplification the libraries were normalized and loaded to a MiSeq™ cartridge for sequencing.

The MiSeq™ cartridge (Illumina, San Diego, CA, USA) is coated by a lawn of primers complimentary to the adaptor sequences and this enables hybridization between the sample and the surface in a similar way as used in micro-array technologies. The hybridized adaptors and primers cause the sample library to attach in a loop manner which is followed by several rounds of 'bridge amplification', allowing many clusters or tiny islets of amplified template to form which serve as 'clones' for subsequent sequencing. The sequencing chemistry uses fluorescently labeled chain terminating NTPs, similar to that used in traditional Sanger sequencing, with the important difference that termination is reversible (Radford et al. 2012).

After one nucleotide wash, a high resolution digital imager is used to capture the fluorescence of the incorporated NTP after which the fluorochrome is cleaved and washed away, reversing the termination by the addition of a free 3'-OH group, and clusters subjected to another sequencing round to capture the identity of the next nucleotide in the sequence of each cluster (Radford et al. 2012; Beerenwinkel et al. 2012).

Reads are then automatically computed and output as a fastq file (two per sample) with phred scale quality scores associated with each read. The resulting sequence data was reported back to the Department of Medical Virology in .fastq.gz format for analysis.

2.3.6 Data Analysis

2.3.6.1 NGS raw data analysis and processing

As sequence data follows the Poisson distribution, the Coverage for each specimen was calculated with the Lander-Waterman equation; C = (L x N)/G (Lander and Waterman 1988). Here the C stands for coverage, the L stands for the read length for the platform (250bp) which is multiplied with N, the total number of reads generated for the specimen, and divided by the full haploid genome length (G; 3221bp).

Sequence data obtained from NGS was uploaded and saved on the local GALAXY server (http://galaxy.bi.up.ac.za/), forward and reverse reads concatenated in a single velvet input file with FastqShuffleseq, converted to fastqsanger format and subjected to quality control analysis (Blankenberg et al. 2010). From the fastq toolbox in

GALAXY, FastQC was used to visualize the quality score data and quality control (QC), which included trimming the sequences, removal of small sequences and of reads with a phred-scale score of below 20, was performed with Fastqfilter (Blankenberg et al. 2010). The phred-scale score reflects the logarithmic relationship between error probability and quality; Q = -10.LogP, where Q is the phred score and P is the error probability (Ewing and Green 1998; Richterich 1998). The probability is determined by comparison of the specimen read signal to calibrated reference tables (Ewing and Green 1998). The full read library of each specimen was then aligned to an appropriate reference genome with BWA for Illumina (Li and Durbin 2010) and saved in the .sam and .bam format. These files were downloaded from the server and visualized locally with IGV 2.3 (Thorvaldsdóttir et al. 2013; Robinson et al. 2011)http://www.broadinstitute.org/igv/) to establish whether sufficient reads were available for full genome coverage.

As QuRe can only run on smaller read files it necessitates that the files generated for our samples be systematically filtered in GALAXY to hone in on the unique sequences that cover the full genome range, until < 100,000 reads were retained from the raw data. This is due to the fact that the program can only handle approximately 30,000 reads per GB of RAM at its disposal. Thus, a 4GB RAM computer can functionally only allow for up to 1.5GB to be allocated to QuRe without disrupting other computer functions. Ideally one would need a computer with 8GB of RAM and be able to allocate all of the memory to QuRe in order to run larger files.

The filtering of files was done using FastqFilter (Blankenberg et al. 2010) by filtering reads based on minimum QC (phred scale) scores. The excellent quality observed in raw read files diminishes any real value to filter by quality but none the less we could decrease the number of reads to a data size more optimal for down-stream analyses. Generally filtering was started by excluding any reads with QC<20. Hereafter, based on the number or reads retained, subsequent filtering had to be done by increasing increments (sometimes as high as QC<30). Files were also imported to Geneious (Kearse et al. 2012) and further filtered by removing duplicated that are identical and only retaining one copy of each read. Files could then be exported in the .fasta format for variant reconstruction using QuRe_v0.99971 (Prosperi and Salemi 2012).

2.3.6.2 Variant reconstruction

The QuRe program is implemented from the command line and built via the Java Development Toolkit 1.7. The first steps involve specifying path locations to the directory in which QuRe.class and .jar files are stored, as well as specifying the input read file, input reference file and parameters such as the homopolymeric and non-homopolymeric rates as well as the number of iterations. These parameters are set to 0.01, 0.005 and 1000 by default unless specified.

The preliminary processes parse the input files, build a dictionary index for the reference and calculate the quasi-random alignment score distribution before commencing alignment/mapping of reads to the reference. Once the reads have been mapped, reads with a p-value > 0.01 are removed and the remaining reads are used to

run QuRe. Steps performed in QuRe include: Phase 0 - fixed-size sliding window overlaps, Phase 1 - random overlaps, Phase 2 – multinomial distribution matching (based on a maximum-likelihood guide distribution) and Phase 3 - assessing best *a posteriori* overlap set, executing core reconstruction algorithms and final clustering (which includes Bayesian Information Criterion selection) of variants. This was done on a computer with seven parallel core processing enabled (i7 Intel core) and default parameters adjusted to use 4 GB RAM.

Each variant in the output files has its' relative frequency appended to the name and are ordered in a descending manner. The output files, in fasta format, containing clustered variants could then be checked in BioEdit 7.2.1 for errors and/or gaps.

Reconstructed variants could then be saved in the fasta format and used in the same way Sanger sequencing data is used to perform phylogenetic analysis of a sequence file, including reference sequences selected from Genbank using BLAST searches. Relevant reference sequences (appendix C) were downloaded from Genbank (NCBI) in the .fas format and imported to the same file as the sequence data.

2.3.6.3 Phylogenetic analyses

Sequence files were used to perform a pair-wise and multiple sequence alignment using MAFFT 7 (Katoh et al. 2009). The aligned sequences were exported in the .fas format and imported to MEGA 5.2 (Tamura et al. 2011), where they were converted to the .meg format and used to perform phylogenetic analysis.

In MEGA 5.2, phylogenetic trees were constructed (Maximum Likelihood, Neighbour Joining and Parsimony) with 1000 bootstrap repeats using the Kimora 2-parameter and pair-wise deletion model, inclusive of both transitions and transversions. From these trees, and by means of reference sequences, sequences were assigned to specific genotypes, subgenotypes and clades.

2.3.6.4 Recombination analyses

When recombination was suspected, samples were analyzed using Boot Scan with a 1000 bootstrap repeats and Grouping Scan, both with the Kimura-2-Parameter model, a part of the simple sequence editor (SSE) analysis package (Simmonds 2012) testing for both intra-genotype recombination. For inter-genotype recombination detection the jpHMM (Schultz et al. 2012) online tool (http://jphmm.gobics.de/jphmm.html) was used.

The first analysis was performed with the online program jpHMM (Schultz et al. 2012) for the detection of inter-genotype recombination in the circular HBV genome. This program employs a bootstrapping algorithm based on a standard set of genotype references to infer genetic likeness. The second part of the analyses used two algorithms implemented in SSE v1.1; Bootscan and Groupscan (Simmonds 2012). The Bootscan algorithm is a bootstrapping based method, similar to SimPlot, which performs and compares bootstraps between the query sequences

Groupscan on the other hand uses a probability scoring matrix to scan the query sequence and graphically plot similarities between the query and tagged group

sequences. The latter of these two algorithms is said to be more stringent and reliable than ordinary bootstrapping methods.

2.3.6.5 Site specific nucleotide and amino acid changes

Furthermore, sequence data was analyzed for site specific known and unknown variation that may be related to the serology of the patients from which samples originate. Analysis of site specific change was done in BioEdit 7.2.1 (http://www.mbio.ncsu.edu/bioedit/page2.html) and MEGA 5.2 (Tamura et al. 2011). Both unknown as well as known variations that occur infrequently in comparison to the references (appendix C) were noted and compared with change at the protein level. Samples were also screened for drug-resistance (Gnaneshan et al. 2007) and vaccine escape mutations.

2.3.6.6 Appropriation of serology data

The serology for study specimens as previously determined (Mayaphi et al. 2012) could be analysed according to diagnostic approaches previously described (Center for Disease Control and Prevention 2012; Previsani and Lavanchy 2002; Bowyer et al. 2011) to determine disease state/progression.

2.4 ETHICAL CONSIDERATIONS

Ethics approval was obtained for samples used in previous studies (UP 35/2007) and blanket consent for prospective sample collection and their use in this project was obtained from the UP student ethics committee (S137/2012, appendix H).

CHAPTER 3

RESULTS

3.1 FULL GENOME EXTRACTION

Of the 20 samples included in this study seven (3791; 3274; N011; 3658; 3768; LA06 and PO04) were successfully amplified from template generated by automated extraction (MagNA Pure) while the remaining eight (N199; 3269; N005; 3319; 4070; 4312; 3358 and N060) could only be amplified after manual extraction using the QIAamp MinElute™ Virus Spin kit. Manual extractions are more compatible with down-stream processes because, for example, they allow the user to manually select both sample and elution volume and, as expected, many (but not all) of the specimens that could not be amplified from automated extractions often had an extremely low viral load (e.g. N60, 3269, 3319, N199).

3.2 PCR AMPLIFICATION

Samples 3791, 3274, N011 and PO04 were amplified with the P1/P2 primer (figure 3.1, blue dot) combination whilst samples 4070, 4312, 3658, 3354 and LA05 successfully amplified with the customised P1/P2_A1 (purple dot) primer set (Table 2.2). The remaining samples were successfully amplified using the degenerate P1/P2RM primer (yellow dot) combination. Interestingly, all low viral load, manually extracted specimens, most of which were amplified using the degenerate P2RM primer generated two bands (Figure 3.1) instead of a single band of ~3kb. Figure 3.1 shows all 19 specimens re-run together on the 0.9% TBE-Agarose gel to compare quality, concentration and size of the cleaned

amplicons prior to sequencing. The larger band was approximately 1750bp and the smaller band approximately 1250bp. Taken together the total size of the amplified region would be approximately 3000 ± 100bp (linear regression analysis) which could only constitute the full genome if the two fragments do not overlap.

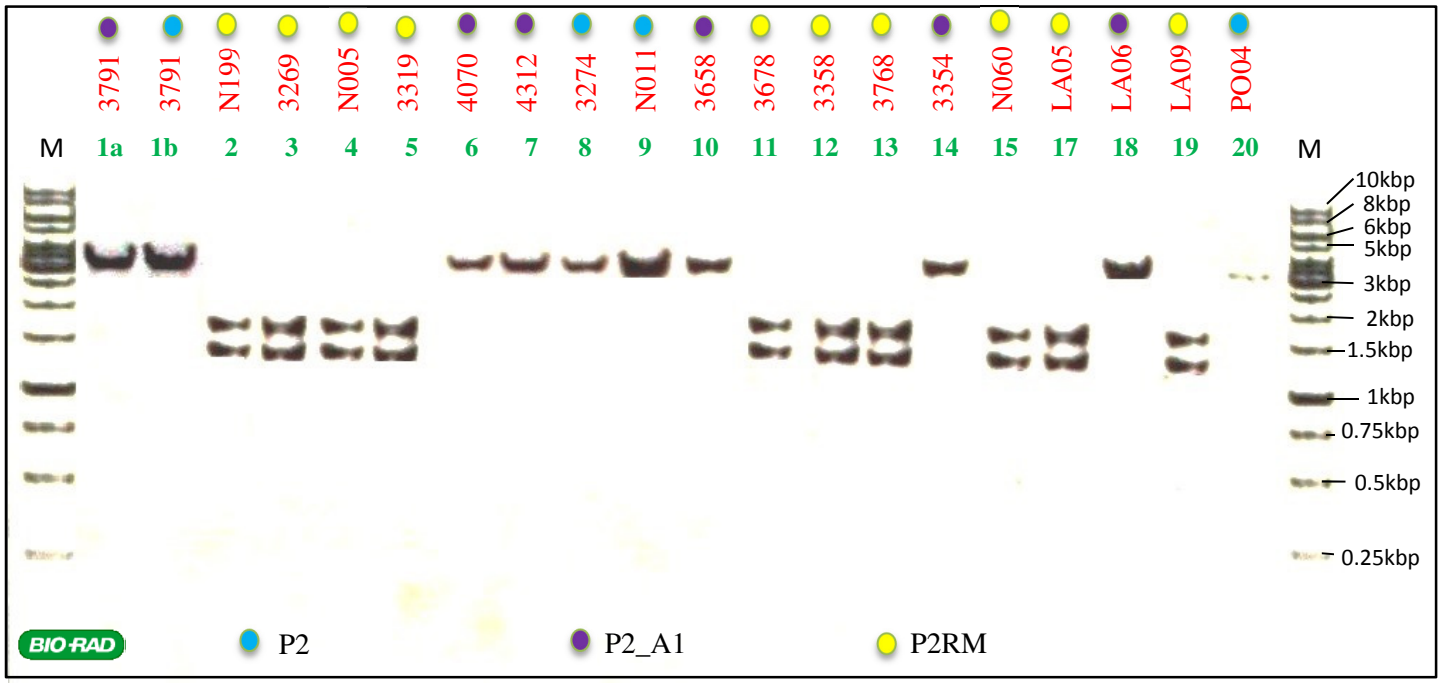


Figure 3.1: SeaKem® LE Agarose - TBE gel (0.9%) image of 20 PCR products generated flanked by two molecular marker lanes (M) (GeneRuler™ 1kb DNA ladder, Thermo Scientific), with the positive control in the second lane from left.

3.3 QUALITY OF DNA SEQUENCE DATA

Only fifteen samples (Table 3.1) were selected for sequencing. Of the remaining five samples, two (3678 and 3354) were previously identified as typical African A1 specimens and another two, LA06 and LA09, were typical Asian A1 specimens. The remaining sample shown as 1a (generated with P1/P2_A1) in

Figure 3.1 was the positive control and a duplicate of 1b (generated with P1/P2) and was thus not sequenced.

Table 3.1: Summary of reads and coverage data for the fifteen sequenced samples

Sample	Raw reads	Coverage (per base)	Coverage (percentage)	Filtered reads	Coverage (per base)	Coverage (percentage)
3791	1,038,628	80,614	7.76%	166,126	12,894	7.76%
N199	602,168	46,738	"	96,239	7,470	"
3269	414,819	32,196	"	Incomplete coverage		
N005	896,104	69,552	"	87,156	6,765	"
3319	1,007,396	78,190	"	75,808	5,884	"
4070	870,998	67,603	"	67,347	5,227	"
4312	987,372	76,646	"	57,390	4,454	"
3274	1,051,200	81,590	"	70,503	5,472	"
N011	387,632	30,455	7.86%	68,542	5,385	7.86%
3658	1,168,760	90,714	7.76%	61,106	4,743	7.76%
3358	148,154	11,499	"	148,154	11,499	"
3768	148,986	11,564	"	Incomplete coverage		
N060	60,922	4,728.5	"	60,922	4,728.5	"
LA05	84,684	6,573	"	Incomplete coverage		
PO04	45,850	3,572	7.79%	45,850	3,572	7.79%

The table summarizes the number of reads contained in the two read files generated for each sample along with the calculated per base sequence coverage and the percentage of coverage for the sequencing run as per the Lander-Waterman equation (Lander and Waterman 1988)C=L.N/G)

Two read files were generated containing the forward and reverse raw reads, respectively. The total number of raw reads as well the size of the final filtered files for downstream analysis are summarized in table 3.1 along with the per base coverage for each as calculated by the Lander-Waterman equation (Lander and Waterman 1988). Coverage of above 2,500 times was observed for all read files which enables the detection of point mutations that occur in <1% of the total reads. For each specimen the two read files, as well as the concatenated reads, showed excellent quality; per base quality scores (± standard deviation) well

above the minimum quality score (Chevaliez et al. 2012) and cut off value of 20 (Figure 3.2).

Furthermore, a quality score of above 20 indicates a probability of <1% for false mutation calling (Ewing and Green 1998). Per base sequence content was relatively parallel as of the 15th base in each read and sequence duplication levels were low, dropping to less than two. The per sequence GC content was normally distributed with a single peak (mean 43%) that overlapped with the theoretical distribution, validating high read quality.

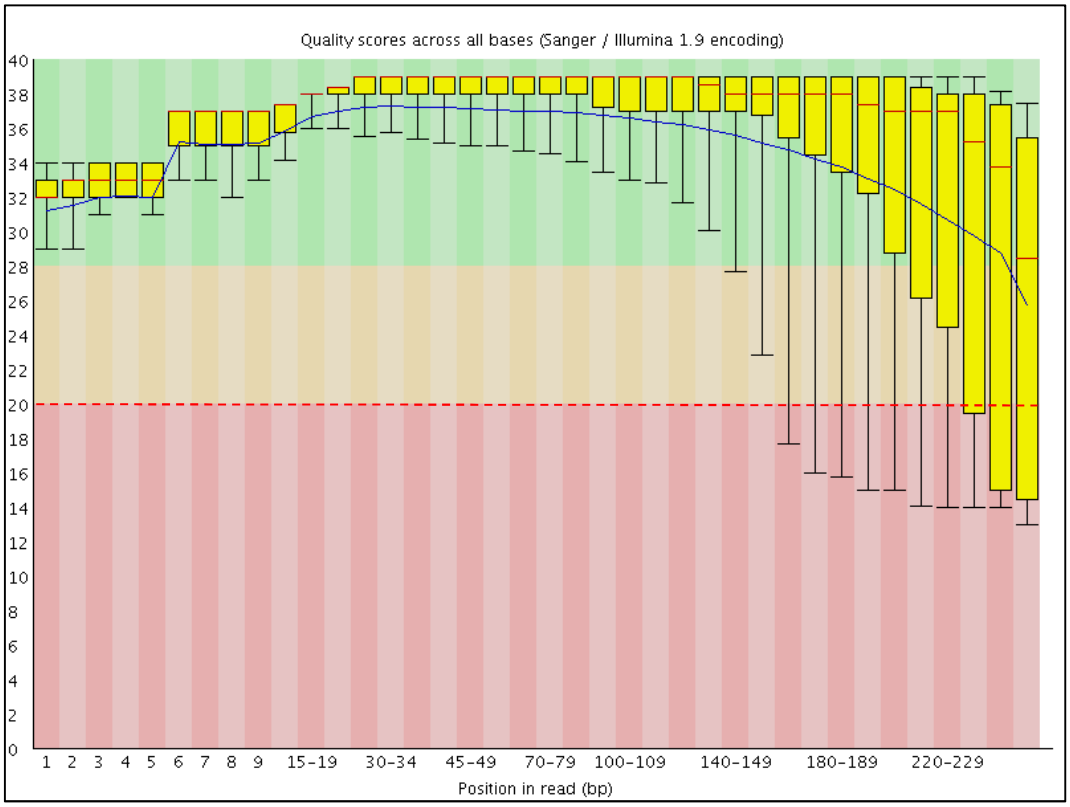


Figure 3.2: Per base quality scores for sample 3791 determined using FastQC in GALAXY. The blue line plots the mean score in a positional manner while the yellow boxes plot the interquartile range around the median (red line) with black whiskers extending to the outer range limits.

Relevant reference genomes were identified based on the core data available from previous studies (Mayaphi et al. 2013) by performing a BLAST search and selecting the best match of which a full genome was available on Genbank. Based on this search the positive control (3791), which partitions with typical African A1 specimen, was assigned reference AY233277. For the outliers (N199, 3269, N005, 3319, 3274, N060 and LA05) the reference genome AY233290 was most appropriate. The genotype D specimen, N011, was most similar to FJ692536 and the genotype E, PO04, matched HE974384.

Each Velvet input file (containing the concatenated forward and reverse reads output from FastQ shuffleseq (Blankenberg et al. 2010) running online from the GALAXY toolbox) was then used with the appropriate reference from the BLAST search results and mapping performed. These files were downloaded, indexed and visualized locally with Integrated Genome Viewer (IGV 2.3; (Thorvaldsdóttir et al. 2013) to confirm that the read files mapped to the chosen reference and to confirm the extent of full genome coverage.

Full genome coverage of mapped reads was achieved in all except three specimens, accounting for most samples that amplified in two bands; 3269, 3768 and LA05 did not cover the full genome. Two of the visualized panels are shown in figure 3.3. The top bar indicates the relative degree of coverage at each position while the bottom bar shows each mapped read (1-3221bp from the HBV EcoR1 site) across the genome with red, blue, green and black lines indicating SNP sites.

In figure 3.3A, complete coverage is observed for sample 3791 with a large amount of reads for each position (80,614 reads per base) and adequate overlapping. Figure 3.3B depicts the mapped reads for sample 3269, one of three

that did not give full genome coverage. Here there are clearly far less reads (32,196 reads per base) for mapping as was reflected by the low read count in the raw data files and these aligned unevenly and gave sparse coverage. Interestingly all samples (3319, N005, N060, N199) amplified in two fragments with the P2RM primer (Figure 3.1) generated reads across the genome although the individual fragments were both less than full genome size.

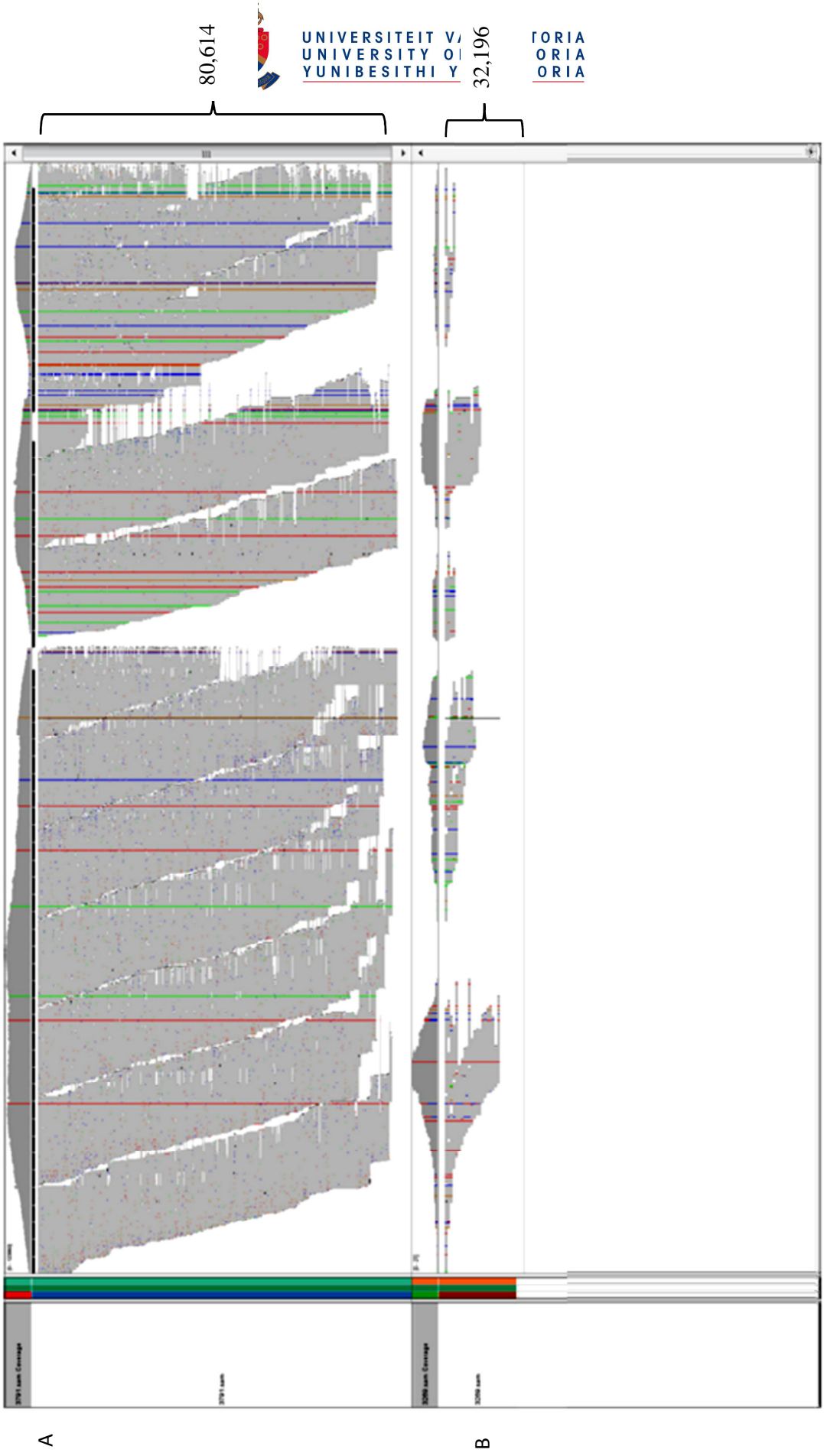


Figure 3.3: IGV visualizations showing the large number of reads (80,614 reads per base coverage) spanning the entire length of the genome for 3791 (A) as well as the lower coverage (32,196 reads per base) as well as incomplete coverage observed for 3269 (B). SNP sites are indicated by coloured lines.

3.4 VARIANT RECONSTRUCTION

The filtered files containing the best 100,000 reads were downloaded from the local GALAXY server and imported into the Geneious (version 6) software package (Kearse et al. 2012) where data required further filtering by removing duplicates and extracting only one copy of each read, for which two identical reads existed, while retaining the proportions of all unique sequences within the quasispecies.

The filtered reads exported from Geneious (Kearse et al. 2012) in fasta format formed the input files to the viral quasispecies reconstruction program, QuRe. In QuRe (v0.99971), the preliminary processes that ran parsed the input files, built a dictionary index for the reference and calculated the quasi-random alignment score distribution before commencing alignment/mapping of reads to the reference (see appendix B). The distribution of quasi-random alignment scores, generated as the reads are mapped to the reference, is compared to the random score distribution by means of a z-test and readings which do not reach significance are discarded. In general a quasi-random alignment score of 65-67 was obtained across the study samples and no less than 2,000 reads retained.

QuRe outputs the clustered variants, for which frequency estimations are recalibrated and refined from the preliminary un-clustered variants by means of a random search and a Bayesian Information Criterion (BIC) selection algorithm, in fasta format along with the calculated relative frequency of each variant. Full genome variants were reconstructed for twelve of the fifteen samples; the remaining three were rejected as they had insufficient reads for complete

coverage. The variants along with their relative frequency and DNA changes that separate them are summarized in table 3.2.

Four variants were reconstructed for specimen 3791 of which one represents the major variant (98.05%) and the remaining three represent minor variants (≤1%). The second specimen, N199, also has four variants which represent two groups (2 x 34.09% and 2 x 15.91%) of which each group comprised two variants of which each is approximately equally prevalent. Sample N005 was most diverse with eight variants in total of which half represent an intermediate (<10%) viral population variant. Both samples 3319 and 4070 have two variants. For 3319 this includes a major variant (88.34%) and a low frequency variant (11.66%), while 4070 has a major variant (98.8%) as well as a minor (1.2%) population variant.

Samples 4312, 3274, 3358 and N060 each has 3 variants. The first, sample 4312, has one major variant (94.61%) and two minor variants with a frequency of 4.38% and 1.01%, respectively. Sample 3274 has two equally distributed (49.97%) major variants along with a very low frequency minor variant (0.07%). The viral quasispecies for N060 comprise two major variants with a relative frequency of 67.31% and 22.68%, respectively, along with a low frequency variant (10.01%). Only one variant was reconstructed for N011. The two remaining specimens, 3658 and PO04, had two variants each. In the case of 3658, one major variant (90.58%) and one minor (9.42%) population variant was detected. For PO04, one major variant with a frequency of 98.19% was detected along with a minor variant (1.81%).

These variants (Table 3.2) were each screened for genotype and drug resistance mutations with the online HBVdb tools (Hayer et al. 2013; Gnaneshan et al. 2007) as well as BLAST searched to assess divergence from known GenBank sequences.

All of the specimens and their respective variants were classified as sensitive and did not contain evidence of treatment associated mutations. Of the fifteen samples 13 were classified as genotype A, whilst N011 and PO04 were classified as genotypes D and E respectively.

Of the genotype A specimens, 3791, 3319, 4070 and 4312 all had a 99% match with several subgenotype A1 references in a BLAST search. Samples 3658 and N060 (variant 3) had a 98% match and N005, 3274 (variants 1 and 2) and N060 (variants 1 and 2) had a 97% match. A 96% match was found for sample 3274 (variant 3) while a 95% match and thus >4% difference was observed across all 4 variants of N199.

Furthermore, the reconstructed variants were compared in a specimen specific manner at the molecular level to evaluate the intra-host variability of the virus. In some cases, such as 3319, 4070 and PO04 the distinction between variants was limited to no more than two position specific changes whilst variants for other samples differed at between 4 and 20 positions (<1% of the full genome). The only exception was the third variant of 3274, which differed at 49 sites (±1.5%) and resulted in the observed larger divergence from Genbank entries, as compared to the first two variants.

Table 3.2: General data profile of the reconstructed variants

Sample number	Sample ID	Variants*	Frequency*	Observed intra-specimen variation	BLAST match	Resistance†	Genotype‡	Accession nr
1	3791	1	98.05	-	99%	Sensitive	A	KF922406
		2	1.02	A201G; A2236G; A2407G; A2572G	99%	Sensitive	A	KF922407
		3	0.24	T192C; T222C; A2568G; G2667A; T2683A	99%	Sensitive	A	KF922408
		4	0.69	A221G; A2193G; A2425G; T2915C; A2921G	99%	Sensitive	A	KF922409
2	N199	1	34.09	A199G; A356G; G925A; N969G; G1801A; G1803C	95%	Sensitive	A	KF922410
		2	34.09	A199G; A356G; G925A; N969A; C1799T; T1800C; G1801T; G1803T; C1804T; C1806G	95%	Sensitive	A	KF922411
		3	15.91	A245G; G587A; N969G; C1799T; T1800C; G1801T; G1803T; C1804T; C1806G; A2577C; T2612A	95%	Sensitive	A	KF922412
		4	15.91	A245G; G587A; N969G; C1799T; G1801T; C1802G; C1804T; A1805T; A2577C; T2612A	95%	Sensitive	A	KF922413
3	3269			Incomplete				
4	N005	1	21.51	C26A; G381A; T770C; T1809G; C1988T; G2364A	97%	Sensitive	A	KF922414
		2	21.51	C26A; G381A; T770C; T1809G; C1988T	97%	Sensitive	A	KF922415
		3	15.7	C26A; G381A; T770C; T1809G; C2260T; A2569G	97%	Sensitive	A	KF922416
		4	15.7	C13T; G381A; T770C; C1988T; C2260T; C2354T; A2569G; A2687G; A2871C	97%	Sensitive	A	KF922417
		5	9.17	C13T; G381A; T770C; C2222A; A2375G; A2687G; A2871C	97%	Sensitive	A	KF922418
		6	9.17	C13T; G381A; T770C; C1988T; C222A; A2687G; A2871C	97%	Sensitive	A	KF922419
		7	3.62	C26A; G381T; C2260T; A2687G; A2871C	97%	Sensitive	A	KF922420
		8	3.62	C26A; G381T; C1988T; C2222A; A2297G; A2687G	97%	Sensitive	A	KF922421
5	3319	1	88.34	-	99%	Sensitive	A	KF922422
		2	11.66	T779C	99%	Sensitive	A	KF922423
6	4070	1	98.8	-	99%	Sensitive	A	KF922424
		2	1.2	deletion G2085	99%	Sensitive	A	KF922425
7	4312	1	94.61	C286A; T1218C; T2613C; G2672T	99%	Sensitive	A	KF922426
		2	4.38	C286A; A2460C; T2516C; T2543C; G2672C	99%	Sensitive	A	KF922427
		3	1.01	C286G; A2460C; T2516C; T2543C; T2613C; G2672T	99%	Sensitive	A	KF922428
8	3274	1	49.97	T2035G; T2869G	97%	Sensitive	A	KF922429
		2	49.97	T2035A; C2100T; C2102T; A2104T; T2869G	97%	Sensitive	A	KF922430
		3	0.07	C13T; C105T; T192C; G241A; T344C; G348T; C353T; A356G; T358C; T359C; G379C; G381A; T382C; C383T; C386A; G387T; G388A; T390A; C427T; T429G; T432C; C433T; T434G; T438C; G449A; T451G; T452C; A453C; C455T; A457G; G458A; A461C; G463C; T464C; T465A; C467A; C468G; G470A; T473G; G474C; T478C; A481C; T483C; T484C; C485T; T592C; T777C; A1368C; T2035A	96%	Sensitive	A	KF922431
9	N011	1	100	n.a.	98%	Sensitive	D	KF922432
10	3658	1	90.58	T54G; A541G; G542A; G543A; T705C; C717T; A1612C; C1766T; T1768A	98%	Sensitive	A	KF922433
		2	9.42	C52T; A97C; T729C; C732T; G765A; G1613A; A1635G; C1653T; A1764G	98%	Sensitive	A	KF922434
12	3358	1	65.69	-	99%	Sensitive	A	KJ010776
		2	25.34	G2668A; T2684A; C2738T; C2745A; G2792A; G2852T; C2871A; C2910A; T2916C; A2922G; C2979T; C3045G; T3104C; C3132T; A3133C; C3163A; G3216A	99%	Sensitive	A	KJ010777
		3	8.95	G2852T	99%	Sensitive	A	KJ010778
13	3768			Incomplete				
15	N060	1	67.31	A1234C; A1368G; C1449T; T2011G; C2245T; T2831C	97%	Sensitive	A	KF922435
		2	22.68	A1234C; A1368G; C1449T; G1993T; A2034T; A2059G; C2245T; T2552A; A2851C; T3154A	97%	Sensitive	A	KF922436
		3	10.01	T1636G; G1993T; T2008A; G2017A; C2022G; A2029G; A2034C; G2035T; C2047A; A2075G; G2129C; A2137T; C2176T; G2188A; T2191C; A2257G; C2266T; A2269G; A2278T	98%	Sensitive	A	KF922437
17	LA05			Incomplete				
20	PO04	1	98.19	-	99%	Sensitive	E	KF922438
		2	1.81	A1934T; A2445C	99%	Sensitive	E	KF922439

* As per QuRe analysis (% per specimen)
† As per HBVdb and HepSEQ analysis
‡ As per Phylogenetic analysis

3.5 PHYLOGENETIC ANALYSES

For phylogenetic analyses, two separate fasta files were created; the first contained the variants along with relevant representative references from the eight (A to H) genotypes, including representatives from the main subgenotypes of each. The second fasta file contained only the samples identified as HBV/A1 and included a large number of subgenotype specific references (appendix C) covering all major and minor clades available for deeper characterization of variation within the subgenotype–not apparent in the first, more general analysis. This analysis was rooted using GenBank accession number JN315779, which represents the oldest HBV full genome (Kahila Bar-Gal et al. 2012) sequenced to date.

Both alignments were checked in BioEdit for alignment errors prior to further analysis - no alignment errors were present - although some gaps were necessary to handle insertions (genotype G) and deletions (genotype D) between the specimens. The respective files were imported to MEGA 5.2 (Tamura et al. 2011) for phylogenetic analysis. Both Neighbour-Joining and the Maximum Likelihood methods were compared but no major differences were observed between the results generated by the two algorithms.

The evolutionary history for A to I, as inferred by the Neighbour-Joining method (Saitou and Nei 1987), is reported in figure 3.4, where the optimal tree with the sum of branch length = 1.48933730 is shown (circular; see appendix D for rectangular). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The evolutionary distances were computed using the Kimura 2-

parameter method (Kimura 1980) and are in the units of the number of base substitutions per site. The analysis involved 103 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 2942 positions in the final dataset.

All variants for samples 3791, N199, N005, 3319, 4070, 4312, 3274, 3658, 3358 and N060 clustered within genotype A with a 97% bootstrap support. More specifically they all clustered within subgenotype A1 (100% bootstrap support). Sample N199 stood out from the rest of the samples with a rather long branch length as compared to other specimens. The sample N011 clustered within genotype D with a 100% bootstrap support and partitioned with subgenotype D4. PO04 clustered with HBV/E references and most closely matched reference sequences from Namibia and Angola.

The second alignment file was used to for the genotype specific analysis of HBV/A1 specimens to infer clade variability and included closest match (95-97%) references to better characterise the specimens of interest. The evolutionary history was inferred using the Neighbour-Joining method (Saitou and Nei 1987) and the optimal tree with the sum of branch length = 0.51258681 is shown in figure 3.5 (circular; see appendix E for rectangular). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches.

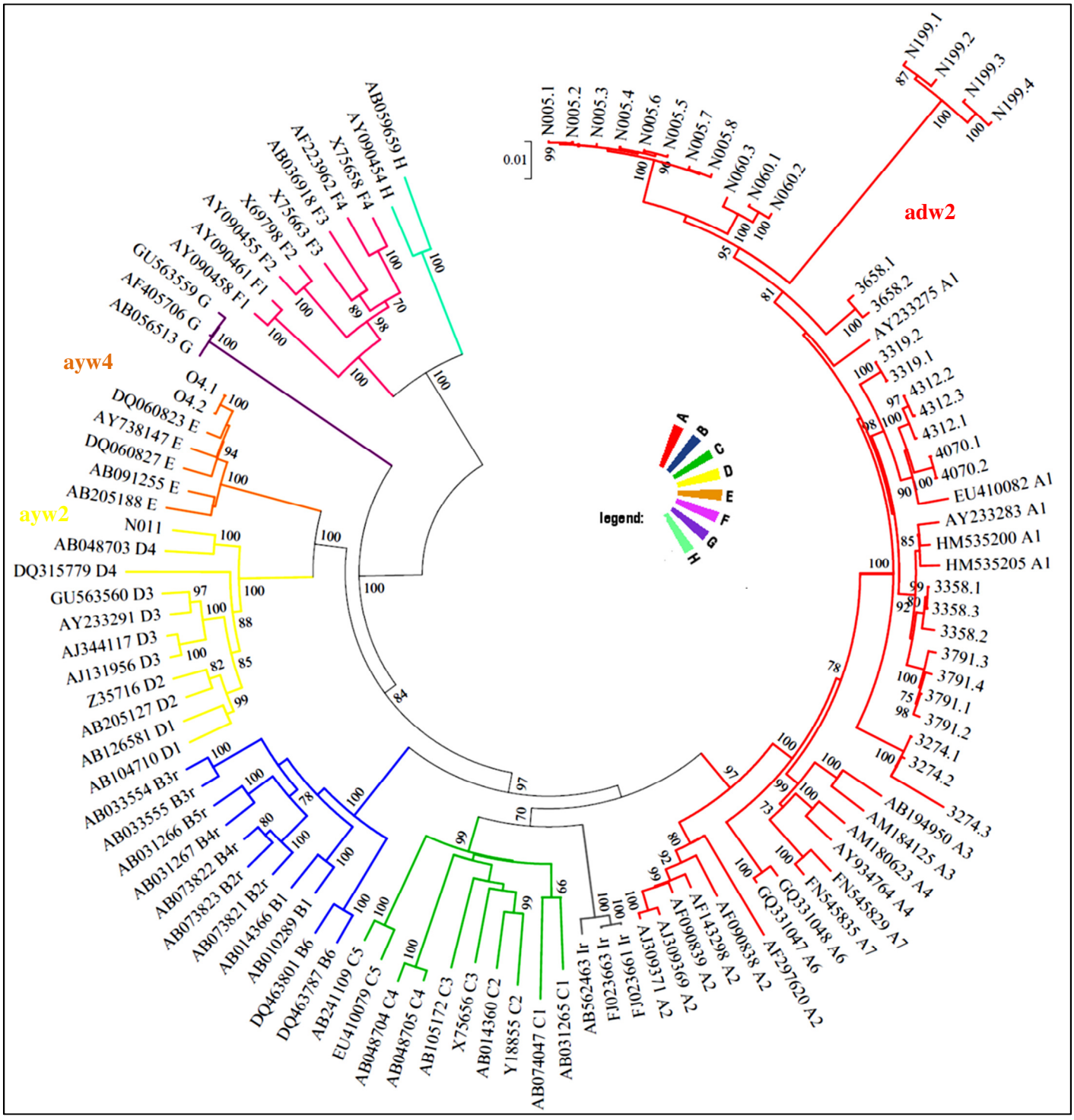


Figure 3.4: Circular display of the phylogenetic tree constructed in MEGA by applying the neighbour joining method. HBV genotypes A (red) to H (turquoise) were included and branches are colour coded according to the legend, the putative recombinant genotype I is shaded in grey. The serotypes of study samples are also indicated. The tree was routed with HBV/F (pink) and H.

The evolutionary distances (not pairwise) were computed using the Kimura 2-parameter (Kimura 1980) method as before. This analysis involved 69 nucleotide sequences. All positions containing gaps and missing data were eliminated resulting in a total of 3081 positions in the final dataset.

As expected, samples 3319, 4070 and 4312 grouped within the Asian A1 clade with a bootstrap support of 99% (figure 3.5, blue). Surprisingly, 3319 which grouped with the subgenotype A1 variants in the previous study (Mayaphi et al. 2013) partitioned with weak bootstrap support (64%) with sequences from Haïti along with sample 4070 (low bootstrap support of 38%). Sample 4312 fell within a different branch of Asian A1 and is most closely associated with AY233278 from South Africa.

Sample 3791, along with other sequences from South Africa and Zimbabwe, clustered within the typical African A1 clade (figure 3.5, pink) with a bootstrap value of 97%.

The outlier group of the study comprising samples N199, N005, 3658 and N060 all clustered together with three references (AF297621, AY233290 and U87742) from South Africa with good bootstrap support (75%). As was observed in the previous preC/C study (Mayaphi et al. 2013), 3274 presented as an outlier to this clade, forming a separate sub-group away from its nearest references. Within the clade, N005, N060 and N199 all clustered together (94% bootstrap) with all three references. Additionally N005 clustered separately with AY233290 as previously but N199 which previously clustered with 3274 uncharacteristically clustered away from the variant group and had much longer branch lengths.

In each instance the sample specific variants all clustered together as a single group within their relevantclades/branches.

3.6 RECOMBINATION ANALYSES

Since specimen, N199, was found to differ from all known references on Genbank by >4% across the full genome by 5 % (95% match) and partitioned abnormally away from all other specimens, the main variant of this sample was subjected to analyses for the detection of recombination.

The first analysis was performed with the online program jpHMM (Schultz et al. 2012) for all variants of N199 to detect inter-genotype recombination in the circular HBV genome. This program employs a bootstrapping algorithm based on a standard set of genotype references to infer genetic likeness and the output for the major variant of N199 is depicted in figure 3.6A.

In this figure (3.6A) the entire mapped genome for the query sequence, N199, is displayed in its circular form and shaded in the colour of the genotype it matches. No evidence of inter-genotype recombination was detected as the entire mapped sequence is of one colour (red), allocating the query sequence to genotype A.

The second part of the analyses used two algorithms implemented in SSE v1.1; Bootscan and Groupscan (Simmonds 2012). In this case, sample N199 and several tagged group representing sequences from the different established subgenotypes within genotype A (1-7) was used to screen for and detect intra-genotype recombination and the resulting graphs are shown in figure 3.7.

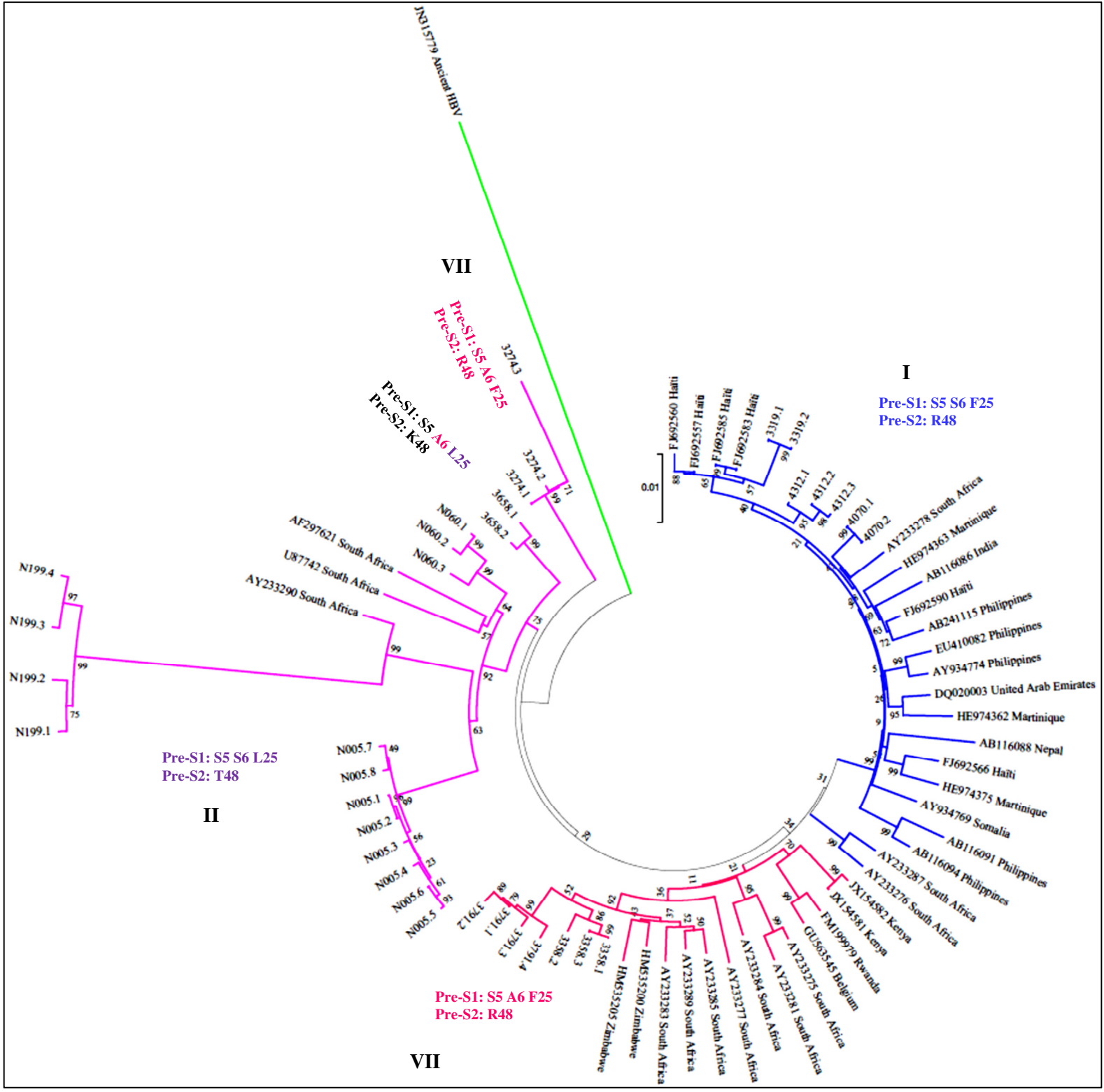


Figure 3.5: Circular display of the phylogenetic tree constructed in MEGA by applying the neighbour joining method. HBV genotypes A1 references were included and branches are colour coded according to show the African A1 (clade VII; pink), Asian A1 (clade I; blue) and outliers (clade II; purple). The amino acids of the Pre-S1 and Pre-S2 regions shared within clades (Makondo et al. 2012; appendix F) are also indicated. The tree was rooted with JN315779 (green) which represents the oldest HBV full genome to date (HBV/C2).

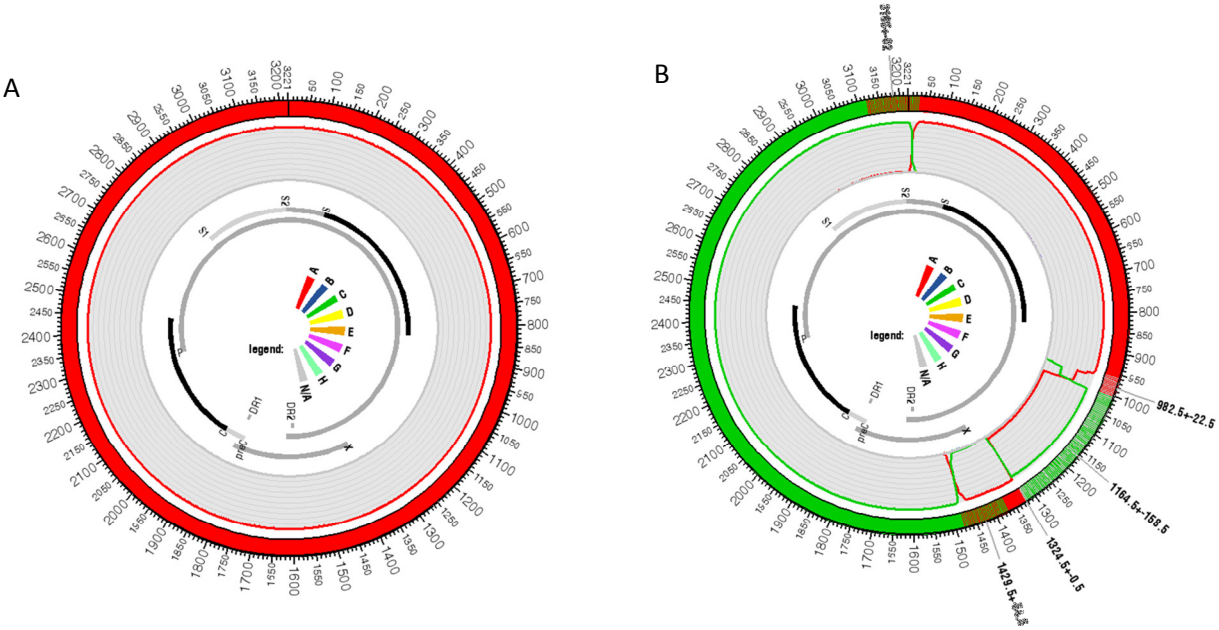


Figure 3.6: jpHMM recombination analysis output for N199 (A) showing no evidence of recombination between established genotypes as well as EU835242 (B) a known A/C recombinant.

For Bootscan results (figure 3.7A) the general consensus, as for phylogenetic analysis, is that a bootstrap percentage is considered significant if ≥70% (Hillis and Bull 1993). In figure 3.7 (A) there was some evidence of recombination between subgenotypes A1 and A4 from position 240 to 400 however the observed peak only had 60% bootstrap support. Another recombinatory event was observed from position 1636 to 2029 which did reach 70% bootstrap support.

Grouping Scan analysis (figure 3.7 B) showed a much more conserved pattern and did not show evidence of recombination with A4. A similar pattern for recombination between A1 and A2 was once again observed between positions 1636 and 2029, with a grouping score peaking at 0.4. For the full genome the global score for A1 was 0.9624 and 0.0117 for A2.

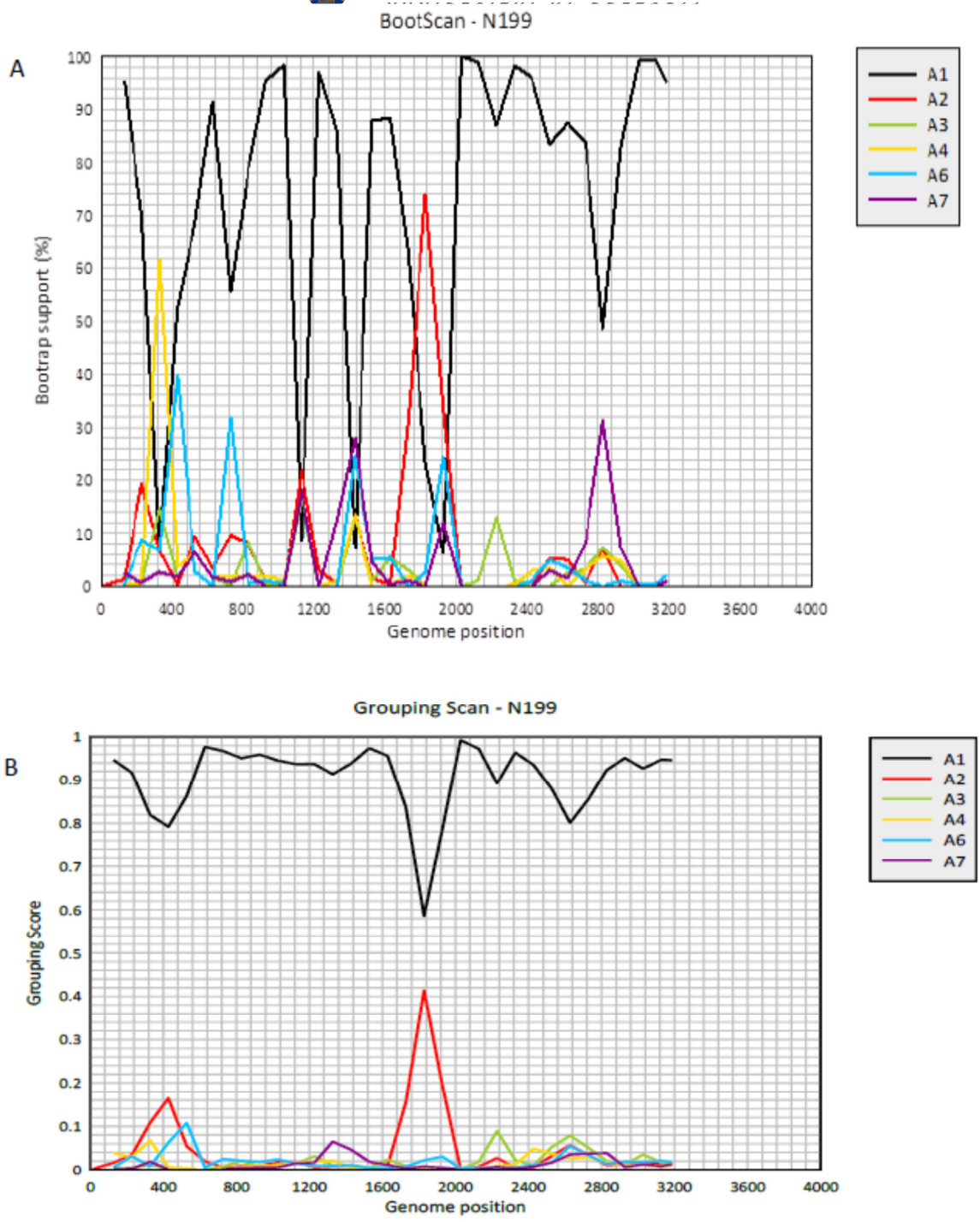


Figure 3.7: Results of Bootscan (A) and Groupscan (B) analysis of N199 where the x-axis shows the genome position and the y-axis the bootstrap support and grouping scores respectively. Each subgenotype is represented by a different colour graph and recombination is depicted by the crossing of different graphs.

3.7 SITE SPECIFIC UNIQUE CHANGE IN HBV/A1 SAMPLES

To assess position specific change across the full genome of reconstructed variants, the multiple alignment files generated by MAFFT were imported into MEGA 5.2 and each sample sequence, along with closely matching references, as per phylogeny, highlighted for variable (200 ± 90) sites. The sparse (matching but only one or two references) and unique variation observed for each specimen was recorded and tabulated at each nucleotide position across the full genome (appendix G). The results for all specimens and their respective variants was compared (table 3.3) along with relevant references in the event of sparse change in addition to comparison with results from previous studies (Mayaphi et al. 2013).

At the nucleotide level, eighty-six variant positions were observed across the full genome in more than one specimen (table 3.2). The most variable of these were the outliers N199, 3274, N005, N060 and 3658. The changes and or wild type observed at C96A, G132A/C, C147T, C290A, C373T and T442A highlighted by Mayaphi et al. 2013 in the S gene of specimens 3274 and N60 were all confirmed. Similarly, we could confirm the pattern of mutations observed previously in the Pre-C/C region between C1981T and A2335G in samples N005, N060 and 3274 but it was equally apparent that all variants of N199 and 3319 as well as variant 3 of N60 were not the same as sequenced previously (Mayaphi et al. 2013) These specimens have very low viral loads of 99, 127 and 157, respectively, so sampling differences would be expected between PCR experiments.

Table 3.3: Shared site specific nucleotide change shared in the study and close references

	Study samples									References									
	3791	**3274**	**N199**	**N005**	**N060**	**3658**	**3319**	**4070**	**4312**	**AY233281**	**AY233288**	**AY233290**	**AF297621**	**AF297622**	**AY233289**	**JN315779**	**U87742**	**Mayaphi et al. 2013**
C81T	X		X									X			X			
C96A		X					X											X
C117T								X	X									
G132A						X				X								X
G132C			X	X	X							X	X					X
G134C			X	X	X													
C147T			X	X	X							X	X					X
C150T	X					X	X	X	X									
T192C	X	X																
C290A		X																X
C353T		X	X															
A356G		X	X															
T358C		X	X															
C373T		X																X
G381A		X		X														
T442A			X	X	X							X	X				X	X
A457G		X			X										X			
C/A493T	X						X			X								
T705C				X		X												
C732T	X					X	X											
T777C		X^{v3}	X			X						X						
A849T				X	X							X					X	
G925A			$X^{v1\&2}$						X									
T975C					X		X					X	X				X	
A1368G					$X^{v1\&2}$								X	X				
A1368C		X^{v3}					X		X			X						
T1425C				X	X			X				X						
C1470T		X	X				X	X	X				X					
A1479G		X				X				X		X						
T1527A			X									X						
T1544A	X					X							X					
T1574A	X					X							X	X				
A1612C			X	X	X	X^{v1}						X					X	

	3791	3274	N199	N005	N060	3658	3319	4070	4312	AY233281	AY233288	AY233290	AF297621	AF297622	AY233289	JN315779	U87742	Mayaphi et al. 2013
T1631C	X												X	X				
G1634A	X												X	X				
C1637A	X												X					
C1638T	X					X							X					
A1727G	X	X	X	X	X	X						X	X				X	
T1740G	X				X								X	X				
T1753C		X	X									X						
A1762T		X	X	X								X						
G1764A		X	X	X		X^{v1}				X		X						
G1809T				$X^{v7,8}$			X	X	X			X						
C1810T							X					X						
C1812T			X^{v3-4}	$X^{v1-3,7-8}$			X	X	X		X	X						
C1812G										X								
T1815G																		
A1850T	X				X	X	X									X		
C1858T	X				X	X										X		
G1862T								X	X									
G1896A					X													
T1909C					X													
G1931T					X													
A1934T					X													
C1981A			X	X	X							X	X	X			X	X
C2002T			X			X												
C2002A				X										X			X	X
C2004T		X																X
G2029A			X	X	$X^{v1,2}$							X	X				X	X
T2035A		X																X
T2035G			X	X	$X^{v1,2}$							X	X	X			X	X
A2047C		X	X	X	$X^{v1,2}$	X						X	X	X			X	X
C2063A		X																X
C2080A		X	X	X		X						X	X	X				X
A2095T			X									X						
C2100A			X	X								X						X
C2100T		X																
A2108T			X									X						

A2131G		X															X
A2131C			X											X			
A2137T	X				X^{v3}		X				X						
	3791	**3274**	**N199**	**N005**	**N060**	**3658**	**3319**	**4070**	**4312**	**AY233281**	**AY233288**	**AY233290**	**AF297621**	**AF297622**	**AY233289**	**JN315779**	**U87742**	**Mayaphi et al. 2013**
T2151A			X	X								X						
T2167C		X	-			X							X					X
G2188A	X	X			X^{v3}		X			X	X							
C2191A		X	-															X
C2191T			X	X	$X^{v1,2}$	X							X	X			X	X
G2237C			X									X						
C2245T			X		$X^{v1,2}$							X						
G2257A			X	X	$X^{v1,2}$	X						X	X	X			X	
T2278A			X		$X^{v1,2}$	X						X	X	X				X
C2293T		X																X
A2302		X																X
C2304A		X																X
A2326T		X																X
A2335G		X																X
A2358G		X		X														
A2358G		X		X														
C2504T	X		X					X	X									
T2518A			X	X	X	X						X	X				X	
C2519G			X									X						
C2519T				X	X	X							X				X	
T2926C			X	X	X	X						X	X				X	
A2995G							X	X	X									
C3021T			X	X	X	X		X				X	X					
T3111C			X			X	X	X	X			X	X				X	
C3115T			X	X	X	X						X	X				X	

Summarizing the observed changes in appendix G, highlighting pivotal (bold) and shared changes between samples for the unique changes, in comparison to references. The vertical axis gives the genomic coordinates of variation observed in HBV/A1 specimens with numbering from the EcoR1 site. The specimens are labelled on the horizontal axis along with the appropriate references. A full list of the corresponding amino acid changes can be seen in appendix G.

Table 3.4: Common amino acid changes

		3791	N199	N005	3319	4070	4312	3274	3658	N060	FJ692585	AY233290	AF297621	AB116088	AY233289	JN315779	U87742
CORE	T13S								X	X			X				
	V17F					X	X				X						
	E93D		X	X								X					X
	L94V		X	X								X					X
	T96N		X	X								X				X	X
	L113Q*		X	X								X					X
	D182G*			X				X						X			
X-GENE	G22S*			X	X		X				X						
	L30F				X		X				X		X				
	R32G*	X	X							X		X	X		X		
	P33S*				X	X	X	X			X		X				
	E80A*		X	X					X	X		X					
	S146A*	X	X	X									X			X	
	S147P*	X	X						X	X			X				
	S6A*	X						X	X				X		X		
SURFACE	F25L		X	X					X	X		X	X				X
	I48V				X	X	X				X						
	V88L		X	X					X	X		X	X				X
	A90T*		X							X		X					X
	R167T*		X	X						X		X	X				X
	A172V		X	X						X		X	X				X
	L173P					X	X					X	X	X			X
	S367L*	X			X												
	V368A					X	X						X	X			X
	S378N		X						X			X		X			
	I382T*							X	X			X					

Change at the protein level across three of the ORFs. The horizontal axis lists the sample ID's whilst the vertical axis indicates the amino acid positions along with the respective gene. The amino acids of the core gene are numbered from the start of the Pre-C region with those of the C region starting at the 30th aa. Similarly the surface gene is numbered

from the Pre-S1 with Pre-S2 starting at the 120th aa and the S region at the 175th. *non-conservative change

Some of the change appeared to be clade or cluster specific such as the T1544A, T1574A and C1636A changes which were shared by both 3791 (typical African A1) and 3658 while samples 3319, 4070 and 4321 shared common variation at the nucleotide level. Furthermore, when compared to reference sequences, six changes (C353T, A356G, T358C, G381A, T705C and C732T) were not shared with any published genotype A1 sequence and unique to this study.

In the kozac sequence, which overlaps with the primer region (1806-1825), the almost characteristic transversions at position G1809T and C1812T to thymine were absent in samples 3791, 3274, 3658 and N060. Changes such as A1762T and G1764A which lead to HBeAg negativity were observed for samples 3274, N199 and N005 whilst the stop codon mutation G1896A (with C1858T) was observed for sample N060. Several stop codons (core: Q18stop, Q208stop; surface: Q190stop, Q204stop, L268stop; x-gene: G27stop, P33stop; DNA pol: W153stop) were observed in all four reading frames of N199.

The four ORF regions (pre-C/C, S-, *Pol* and X-regions) were extracted for the variant sequences and translated to obtain the amino acid composition *in silico*. These regions were analysed in MEGA, as for nucleotide change, to establish changes that have occurred at the protein level (appendix G) and common variations were included in table 3.4. Twenty-six amino acid substitutions were observed. In the core gene, for which seven shared changes were observed, samples N199 and N005 shared four changes at E93D, L94V, T96N and L113Q

and differed in only one site. Samples 4070 and 4312 shared a V17F change while T13S change was observed for 3658 and N060.

In the region of the x-gene (table 3.4), the outlier clade shared change at E80A and S146A (N199 and N005), while the Asian A1 clade specimens shared common change at position P33S along with sample 3274 and change at L30F (3319 and 4312).

The surface gene appeared to be more variable and 14 sites of unique or sparse change (table 3.4) were shared between the samples. Two changes, F25L and V88L, were observed in four (N199, N005, 3658 and N060) of the nine A1 specimens. These four specimens are all representative of the outlier clade. The S6A change, characteristic of African A1 was observed for samples 3791, 3274 and 3658. Two sample specific changes were observed within the a-determinate (amino acid 124 to 147) of the HBsAg. In sample N199, the non-conservative change of G145E (variant 1 and 2) or G145K (variant 3 and 4) was observed in the minor loop (139-147 aa) while the change of G130N was observed for the first variant of sample 3658.

3.8 INTERPRETATION OF SEROLOGICAL DATA

The 15 sequenced study samples (table 3.1) can be diagnosed based on the results obtained for their primary and secondary HBV markers; HBsAg, anti-HBs, total anti-HBc, HBeAg and anti-HBe, respectively as described in section 1.5.1.

The primary marker screen ($HBsAg^+$, $anti\text{-}HBs^-$ and $anti\text{-}HBc^+$) which is interpreted as typical of both acute and chronic disease (grouping II; see table 1.1) and final diagnosis requires screening of secondary markers. Based on secondary

markers 3791, 4070, 4312, 3274 and N011 who were HBeAg^{+} and anti-HBe^{-} with high viral loads could either have an acute infection < 16 weeks (symptomatic phase) or chronic infection < 55 months (immuno-tolerant phase). ALT levels were also high (>45U/L).

3269 and 3358 are also part of the primary grouping II but were negative for both HBeAg as well as its antibody which, in conjunction with the low viral load and ALT levels, could indicate an acute infection of less than 24 weeks or a chronic infection between 60-100 months post-exposure. This period is characterized by HBeAg seroconversion where both the antigen and antibody titres are below the detection limit whilst surface antigen is still detectable.

Specimens N005, 3658, N060 and LA05 were also with primary group I but had a secondary marker combination of HBeAg^{-} and anti-HBe^{+} with low viral loads and ALT levels. This could be characteristic of an acute infection between 18 and 22 weeks post exposure which has become self-limiting prior to an anticipated healthy inactive carrier state. However, this pattern may also be present in chronic infections between 62 and 100 months post exposure which, as with acute infections, marks a period of lower viral replication which may lead to convalescence upon HBsAg seroconversion. The remaining samples (N199, 3319, 3768 and PO04) fall within grouping V which is positive for anti-HBc only and may represent resolving acute, resolved, passive transfer or occult infection.

Since acute infection is usually asymptomatic and occurs in childhood in Negroid South Africans, it is most likely that all of the patients are indeed chronic carriers and their diagnoses will be further discussed in terms of the effect of their ultra-deep sequence results also taking into consideration their HIV status.

CHAPTER 4

DISCUSSION

The most widely used protocol for amplifying the full genome of HBV (Gunther et al. 1995) was modified in this study to accommodate for changes within the published primers that have been identified within HBV/A1.This is due to the fact that genotype A commonly has 1809T and 1812T (Kramvis 2008). The use of the subgenotype A1 primers together with degenerate primers also enabled the screening of samples for dual infections. The optimized methods described in this study, inclusive of a degenerate primer mix and utilizing but a single thermal cycling protocol, successfully amplified all specimens–including occult samples with extremely low viral loads. The modified method also enabled the amplification of all variants irrespective of changes within the kozak sequence, which overlaps with the reverse primers, which means that the PCR method not only took variation into consideration but changes that occurred within the quasispecies could be detected when using the degenerate primer. This is evidenced by the fact that variation was seen between individual variants of N005 and N199 within the primer region where only a select group of variants had 1809T and 1812T while others had the wild type seen in other genotypes, both samples having been amplified with the degenerate primer. Samples from genotypes A, D and E were amplified with no apparent cases of dual infection.

The chosen NGS platform, Illumina MiSeq, generated high quality reads (quality scores > 20) and excellent per base coverage for all samples. Extensive quality filtering and trimming was not necessary, which may have been necessitated had a

different platform been used as other studies have reported lower quality and higher error rates (Quail et al. 2012). Mapping results revealed full genome coverage for 12 of 15 samples and based on these results it is evident that those samples which amplified in two bands with the P2_RM primer did in fact constitute the full length of the genome.

Variant reconstruction did however pose many problems. One important pitfall of QuRe is the use of a less robust alignment program that can only effectively process read files containing less than 100000 reads. This hampers the utility of generating large read files. None the less, filtering read files by already exceptionally good quality scores could be used to decrease the number of reads. Also, QuRe's algorithm applies a correction for homo-polymeric as well as hetero-polymeric sequencing errors which is one of the anticipated errors from Illumina platforms (Barzon et al. 2011) and was recently validated empirically on hepatitis C (Prosperi et al. 2013) making it particularly utile in our study.

By applying these methods we were able to successfully reconstruct the viral variants within a quasispecies for 12 of the 15 study samples and generated a total of 34 full genomes. The 12th sample, 3358, could only by reconstructed after numerous attempts on different read files, using a computer with 8GB RAM available and could thus not be fully characterized in this dissertation. The remaining three samples could not be reconstructed from a single run. The main reason for this appeared to be poorer mapping quality when fewer reads are included, a problem that has been reported in other studies as well (Barzon et al. 2011; Cheval et al. 2011).

As NGS has the ability to detect minor variants we were interested in assessing the presence of drug mutations in our cohort, as other studies have detected emergent resistance mutations within a quasispecies (Nishijima et al. 2012), but no evidence of drug resistance was observed. This was anticipated as all specimens came from treatment naïve patients. The fact that we were able to reconstruct the full quasispecies and in doing so confirm the absence of variants containing drug resistance mutations, which may be positively selected for upon treatment, is particularly advantageous for clinicians considering first line therapy. The ability to detect low level or minor variants prior to the initiation of therapy vastly improves first line treatment options with the most efficacious drug which is not compromised by the presence of resistance mutations within the quasispecies.

Generally the viral quasispecies showed some degree of intra-specimen variation but their genetic distance from their best BLAST match were similar with the exception of one sample (3274) for which a minority variant (0.07%) had an increased divergence from known genomes and other variants of the same specimen. Of the 34 variants, 19 represent low frequency and minor variants (<20%). For some samples (3319, 4070, 4312 and PO04) the difference between variants was restricted to one or two positional differences whilst others differed at multiple sites while remaining less than 1% different from other variants of the same sample. This illustrates the ability of NGS to detect minor variants that occur at a frequency less than 5% (Chevaliez et al. 2012) even when the variance may be limited to a small number of positional changes, which would not have been possible with conventional methods. Furthermore, reconstruction of the

quasispecies enabled the detection of changes at the nucleotide level that is present in one variant but not in the other. Examples of this include sample 3658 where the T705C and A1612C change was only present in the first variant while the C732T change was present in the second variant. Similarly, G381A was present in variants 7 and 8 of N005 but not in the other six variants.

Phylogenetic analysis performed in MEGA by applying neighbour joining methods with 1000 bootstrap repeats and equal transition to transvertion rates with the kimura-2-parameter model proved to be adequate at modelling the relationships within and between genotypes and subgenotypes. The first analysis included references spanning all eight genotypes-as well as putative genotype I- and verified the genotypic classification determined as part of the HepSEQ and HBVdb analysis of each reconstructed variant. All samples belonged to one of the three genotypes A, D and E, known to be circulating in South Africa (Mayaphi et al. 2013).

In each instance the variants grouped together within a specimen specific cluster and did not separate and/or interleave with other samples or references. This also illustrates, as expected, that even when including both major and minor population variants, only detectable by means of NGS, there is a lower degree of intra-specimen quasispecies variation than inter-specimen variation. When specifically analysing the HBV/A1 specimens a similar pattern of specimen clustering was observed but samples that did not group with references in the analysis which included all of the genotypes did cluster with the closest BLAST match references (AF297621, AY233290 and U87742 from South Africa) from the more extensive A1 reference list. Samples seemed to divide into specific

clusters with shared signature amino acids within the surface gene. A similar pattern of clade divisions within A1 was observed in a previous study of HIV infected individuals in southern Africa (Makondo et al. 2012).

Specifically, clades identified in this (figure 3.5) and the Makondo (2012; appendix E) study consisted of Asian A1 references (99% bootstrap support) that grouped with study samples 3319, 4070 and 4312 as cluster or putative genogroup I. In addition to the definitive changes in the preS1 region of the surface gene S:5S, S:25F and S:167R (equivalent to Makondo's preS2:48R) highlighted in the Makondo study (2012), samples 4070 and 4312 shared the surface gene mutations S:L173P in the preS2 region and S:P220T and S:V368A in the HBsAg coding region. Other clade specific changes were in the core gene (C:V17F) and x gene (X:P33S along with sample 3274 and change at L30F (3319 and 4312)). The one sample within this clade representing an occult infection (3319 both versions) grouped separately from the other two Asian A1 specimens.

A second clade (figure 3.5, pink) was observed for African A1 samples and grouped with samples 3791 and 3358. This clade shares the characteristic S5, A6, F25 (Pre-S1) and R48 (Pre-S2), along with sample 3274, and is labelled as genogroup VII. This clade shared change at both the DNA (T1544A, T1574A and C1636A) as well as protein level (such as the surface gene change of S6A). Sample 3658 grouped separate from the African A1 clade but shared some of the variations typically associated with the clade.

The samples of interest for the present study, dubbed "outliers" grouped together in phylogenetic analyses (figure 3.5, bootstrap of 75%) and were highly variable

across the full genome, sharing ten to twelve of the common variant sites. The clade shared S5, S6, L25 (Pre-S1) and T48 (Pre-S2) which is characteristic of a subclade of Asian A1 strains, labelled genogroup III. The single exception was sample 3658 which has S5, A6, L25 (Pre-S1) and K25 (Pre-S2). This isn't characteristic of any of the proposed groups however the change in the Pre-S1 is shared with GQ355557 and AY576430 as well as in the Pre-S2 for the latter. The change of C1981T (sample N199, N005 and N060), T2035G (3274 and N060) and C2191T (N199, N005, 3274 and N060) observed in previous studies (Mayaphi et al. 2013) were also observed in this study. However, much of the core gene differed from what was observed in previous studies which may be due to difficulties in reconstructing novel variants from NGS reads in the presence of high variability and absence of an appropriately similar reference for mapping (Prosperi et al. 2013).

Within the outlier clade, three samples that were HBeAg negative with detectable levels of antibody grouped together with the occult infection specimen N199, however given the amino acid variations observed for this sample in the surface gene it may very well also represent a chronic hepatitis B virus infection that is HBeAg negative due to stop codon mutations. From the Makondo (2012) study this "outlier" clade is clearly a sub-clade of Asian A1 specimens designated to genogroup III and does not represent a new subgenotype based on BLAST similarity data.

Of the changes observed in the core region, most were conserved with the exception of the L113Q and D182G change observed for samples N199, N005 and 3274, all of which fell within the genogroup III clade. One of the changes,

V17F (G1862T), observed in samples 4070 and 4312 of the Asian clade was previously postulated (Kramvis 2008) to be involved in HBeAg seroconversion due to interference with signal peptide cleavage however in both of these instances the samples tested HBeAg positive and anti-HBe negative. Other noteworthy amino acid changes includes the non-conservative changes L84Q (N005 and N199) and T91I (N005) as well as the conservative change of L95I (N005), which fall within a cluster of amino acids between 84 and 101 that have been implicated in more severe liver disease (Ehata et al. 1992; Ehata et al. 1993).

Other changes to the pre-core/core gene and associated regulatory regions include A1762T (x-gene - K130M; N199, N005 and 3274) and G1764A (V131I; N199, N005, 3274 and 3658 variant 2) to the BCP which diminishes HBeAg transcription and secretion (Kramvis 2008). Both samples N199 and N005 were HBeAg negative however 3274 still had detectable antigen levels, possibly because the patient was already undergoing seroconversion whilst circulating HBeAg remained at detectable levels.

In the kozak sequence the G1809T (N005 variants 7-8, 3319, 4070 and 4312) and G1812T (N199 variants 3 and 4, N005 variants 1-3 and 7-8, 3319, 4070 and 4312) mutations, previously associated with HBeAg negativity (Kramvis 2008), were also observed. Two of these specimens, N005 and 3319 were HBeAg negative. This variation was not present in 3791 or 3274 nor N060 or 3658 which explains why these specimens amplified with the non-A1 primers rather than the specific A1 primer.

Furthermore, three samples (N199, N005 and N060) also have several stop codons, including the extremely rare-for genotype A1-G1896A (W28stop, C1858T) mutation for N060, which in conjunction with the aforementioned changes virtually abrogated HBeAg expression in all three specimens. However, given the fact that G1896A and C1858T were both present for N060, there would not be interference with the Watson-Crick bond formation as A-T paring will still occur in the ε-signal. Interestingly, three of these samples (N005, 3685 and N060) had antibody to HBeAg and have thus seroconverted while two of them, N199 and 3319, had neither antigen nor detectable levels of antibody, indicative of an occult infection. This supports the notion (Kramvis 2008) that these changes are associated with the establishing of HBeAg negativity in both seroconversion and occult infections.

Additionally, samples N199 and 3658 also bare a C2002T mutation which, along with the C1762T (N199, N005 and 3274) and T1764A (N199, N005, 3274 and 3658) change, has been associated with/implicated in disease progression towards hepatocellular carcinoma (Zhu et al. 2010). This is particularly of note for N199 and 3658 where the patient is infected with a strain carrying two to three (C1762T, T1764A and C2002T) mutations which, along with non-conservative change in the x gene (Toh et al. 2013) and changes to a cluster of genes in the core gene (Ehata et al. 1992; Ehata et al. 1993), are associated with disease progression to HCC.

Of the shared change observed for the surface gene, six were non-conservative and eight were conservative however none fell within the major antigenic loop of the a-determinate. Two sample specific changes were observed within the a-

determinate (amino acid 124 to 147) of the HBsAg. In sample N199, the non-conservative change of G145E (variant 1 and 2) or G145K (variant 3 and 4) was observed in the minor loop (139-147 aa) while the change of G130N was observed for the first variant of sample 3658. Although loss of antigenicity is, more often than not, the result of cysteine to serine (Waters et al. 1992) changes within the a-determinant, other changes such as G145R (Seddigh-Tonekaboni et al. 2000) have been associated with decreased antigenicity (Weber 2005). As sample N199 was found to be negative for both the surface antigen and its associated antibody the variation at amino acid G145E/K may very well be the reason as variation such as G145R mutants are not detected by most HBsAg assays that use antibodies directed against the second loop (Weber 2005). Amino acid changes, at this position normally result from immunological pressures (Sheldon and Soriano 2008) which could hold true for sample N199 as it is one of three study samples for which the patient was not co-infected with HIV. The variation observed for sample 3658 did not translate to a change in serology as the sample still tested positive for HBsAg. None of the changes due to therapy or immune pressure (figure 1.8; Sheldon and Soriano 2008) were observed in this study. This is likely due to the fact that all patients were treatment naïve and most were immunocompromised due to dual infections with HIV.

Some differences were highlighted between the results of the present study and that observed by Mayaphi et al. (2013). These differences are likely due to the fact that PCR can selectively amplify a single variant which is later the major variant for conventional sequencing while NGS sequences the full quasispecies. Also, the

availability of data enabled the inclusion of more closely matching references in the present analyses.

Due to the longer branch lengths of N199 in phylogeny reconstructions and a 5% divergence from known sequences in a BLAST search the majority variants were screened for evidence of recombination using a bootstrap (bootscan) as well as probability based (groupscan) method. Evidence of recombination, albeit below optimal confidence levels, was observed for intra-genotype recombination between A1/A2 within the region between 1636 and 2029 bp. Upon molecular analysis only three point variations were observed at positions 1727, 1809 and 1812, respectively. Sample N199 has 1727G, 1809G and 1812C (variants 1 and 2) which are characteristic of subgenotype A2, whilst the subgenotype A1 references have 1727A, 1809T and 1812T, respectively. However, many other positions within this region harbour nucleotides characteristic of HBV/A1 and not of subgenotype A2 and even some that were common to both or present in neither.

Thus, the detected variation in N199 might very well be due to complex recombination events in the pre-core/core and overlapping x-gene. It is however more likely the result of poor reconstruction when fewer reads are included (Barzon et al. 2011) for this sample as, in spite of large spans of sequence similarities, the four reconstructed variants have stop codon mutations in each of the four ORF's, inclusive of the viral polymerase, which would make for a defective virus. Further analysis on different read volumes would need to be done to assess the accuracy of this specimens' reconstructed quasispecies. However, it is still possible that this defective virus represents a minor variant population

which has been mutated to the point of large defects and only remains competent due to the presence of other, replication competent, variants not amplified from the PCR.

CHAPTER 5

CONCLUSION

In the present study samples of both high and low viral loads were amplified for the three most abundant genotypes in Africa, using a modified primer set and amplification strategy capable of detecting dual infections. No dual infections were detected. For sequencing by NGS the Illumina platform was evaluated and generated ample data of high quality phred-scale scores for reconstruction, without the need to design genotype/variant specific sequencing primers. Multiple variants from the samples were reconstructed, inclusive of minor variants (as low as 0.07%), generating nearly 40 full genome sequences. The present study observed several unique as well as rare or unusual changes and confirmed observations from previous studies. Phylogenetic analyses revealed that all study specimens belong to the three established genotypes co-circulating in southern Africa. Several study samples presented as an outlier clade with significant support but differed by less than the minimum criterion for distinguishing subgenotypes. The single specimen which did differ significantly showed evidence of recombination. The results herewith reported clearly illustrate the utility of next generation sequencing technologies in characterizing the full spectrum of variation within the viral quasispecies within a host and how changes at the genomic level relate to serology and disease progression. The ability to detect and characterize minor variants as they emerge during the course of infection and treatment could revolutionize not only our understanding of the virus but also greatly aid global efforts at the eradication thereof.

Sequence data has been submitted to Genbank; accession numbers KF922406-39 and KF010776-8.

CHAPTER 6

REFERENCES

A

Allain, J., Belkhiri, D., Vermeulen, M., Crookes, R., Cable, R., Amiri, A., Reddy, R., Bird, A. and Candotti, D., 2009. Characterization of occult hepatitis B virus strains in south african blood donors. Hepatology, 49(6), pp. 1868-1876.

Andernach, I.E., Hubschen, J.M. and Muller, C.P., 2009a. Hepatitis B virus: the genotype E puzzle. Reviews in Medical Virology, 19(4), pp. 231-240.

Andernach, I.E., Nolte, C., Pape, J.W. and Muller, C.P., 2009b. Slave trade and hepatitis B virus genotypes and subgenotypes in Haiti and Africa. Emerging Infectious Diseases, 15(8), pp. 1222-1228.

Araujo, N.M., Waizbort, R. and Kay, A., 2011. Hepatitis B virus infection from an evolutionary point of view: How viral, host, and environmental factors shape genotypes and subgenotypes. Infection Genetics Evolution, 11(6), pp. 1199-1207.

Arauz-Ruiz, P., Norder, H., Robertson, B.H. and Magnius, L.O., 2002. Genotype H: a new Amerindian genotype of hepatitis B virus revealed in Central America. Journal of General Virology, 83(Pt 8), pp. 2059-2073.

Ayoub, W.S. and Keeffe, E.B., 2011. Review article: current antiviral therapy of chronic hepatitis B. Alimentary Pharmacological Therapy, 34(10), pp. 1145-1158.

B

Barzon, L., Lavezzo, E., Militello, V., Toppo, S. and Palù, G., 2011. Applications of next-generation sequencing technologies to diagnostic virology. International Journal of Molecular Sciences, 12(11), pp. 7861-7884.

Beck, J., Urnovitz, H.B., Riggert, J., Clerici, M. and Schutz, E., 2009. Profile of the circulating DNA in apparently healthy individuals. Clinical Chemistry, 55(4), pp. 730-738.

Beerenwinkel, N., Günthard, H.F., Roth, V. and Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Frontiers in Microbiology, 3, pp. 1-16.

Billioud, G., Pichoud, C., Puerstinger, G., Neyts, J. and Zoulim, F., 2011. The main hepatitis B virus (HBV) mutants resistant to nucleoside analogs are susceptible in vitro to non-nucleoside inhibitors of HBV replication. Antiviral Research, 92(2), pp. 271-276.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J. and Nekrutenko, A., 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics, 26(14), pp. 1783-1785.

Bowyer, S.M. and Sim, J.G.M., 2000. Relationships within and between genotypes of hepatitis B virus at points across the genome: footprints of recombination in certain isolates. Journal of General Virology, 81(2), pp. 379-392.

Bowyer, S.M., Sim, J.G.M. and Webber, L.M., 2011. Current laboratory diagnosis of hepatitis B virus infection including 8 years of retrospective laboratory data: hepatitis B is far more infectious than HIV. Continued Medical Education, 29(5), pp. 210-213.

Bowyer, S.M., van Staden, L., Kew, M.C. and Sim, J.G., 1997. A unique segment of the hepatitis B virus group A genotype identified in isolates from South Africa. Journal of General Virology, 78(7), pp. 1719-1729.

C

Candotti, D., Lin, C.K., Belkhiri, D., Sakuldamrongpanich, T., Biswas, S., Lin, S., Teo, D., Ayob, Y. and Allain, J., 2012. Occult hepatitis B infection in blood donors from South East Asia: molecular characterisation and potential mechanisms of occurrence. Gut, 61, pp.1744-1753.

Capobianchi, M.R., Giombini, E. and Rozera, G., 2013. Next-generation sequencing technology in clinical virology. Clinical Microbiology and Infection, 19(1), pp. 15-22.

Carman, W.F., Karayiannis, P., Waters, J., Thomas, H., Zanetti, A., Manzillo, G. and Zuckerman, A.J., 1990. Vaccine-induced escape mutant of hepatitis B virus. The Lancet, 336(8711), pp. 325-329.

Cartmill, M., 1998. The status of the race concept in physical anthropology. American Anthropologist, 100(3), pp. 651-660.

Center for Disease Control and Prevention, January 31, 2012-last update, Hepatitis B Information for Health Professionals [Homepage of CDC], [Online]. Available: http://www.cdc.gov/hepatitis/HBV/HBVfaq.htm#general [December 2, 2013].

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F. and Berthet, N., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. Journal of Clinical Microbiology, 49(9), pp. 3268-3275.

Chevaliez, S., Rodriguez, C. and Pawlotsky, J., 2012. New virologic tools for management of chronic hepatitis B and C. Gastroenterology, 142(6), pp. 1303-1313. e1.

D

Deres, K., Paessens, A., Goldmann, S., Hacker, H., Weber, O., Kramer, T., Niewohner, U., Pleiss, U., Stoltefuss, J., Graef, E., Koletzki, D., Masantschek, R., Reimann, A., Jaeger, R., Gross, R., Beckermann, B., Schlemmer, K., Haebich, D. and Rabsamen-Waigmann, H., 2003. Inhibition of hepatitis B virus replication by drug-induced depletion of nucleocapsids. Science, 299(5608), pp. 893-896.

Dienstag, J.L., 2010. Chapter 37: Acute Viral Hepatitis. In: D.L. Longo and A.S. Fauci, eds, Harrison's Gastroenterology and Hepatology. 1st edition. Bethesda, MD: McGraw-Hill Professional, pp. 349-377.

Dienstag, J.L., 2011. Chapter 306: Chronic Viral Hepatitis. In: D.L. Longo, A.S. Fauci, D. Kasper, S. Hauser, J. Jameson and J. Loscalzo, eds, Harrison's Principles of Internal Medicine. 18th edn. New York: McGraw-Hill, pp. 2567-2588.

DiMattia, M., Watts, N., Stahl, S., Grimes, J., Steven, A., Stuart, D. and Wingfield, P., 2013. Antigenic Switching of Hepatitis B Virus by Alternative Dimerization of the Capsid Protein. Structure, 21(1), pp. 133-142.

E

Ehata, T., Omata, M., Yokosuka, O., Hosoda, K. and Ohto, M., 1992. Variations in codons 84-101 in the core nucleotide sequence correlate with hepatocellular injury in chronic hepatitis B virus infection. Journal of Clinical Investigation, 89(1), pp. 332.

Ehata, T., Omata, M., Chuang, W.L., Yokosuka, O., Ito, Y., Hosoda, K. and Ohto, M., 1993. Mutations in core nucleotide sequence of hepatitis B virus correlate with fulminant and severe hepatitis. The Journal of clinical investigation, 91(3), pp. 1206-1213.

Ewing, B. and Green, P., 1998. Base-calling of automated sequencer traces using Phred. II. error probabilities. Genome Research, 8(3), pp. 186-194.

F

Fang, Z., Hué, S., Sabin, C.A., Li, G., Yang, J., Chen, Q., Fang, K., Huang, J., Wang, X. and Harrison, T.J., 2011. A complex hepatitis B virus (X/C) recombinant is common in Long An county, Guangxi and may have originated in southern China. Journal of General Virology, 92(2), pp. 402-411.

Feld, J.J., Colledge, D., Sozzi, V., Edwards, R., Littlejohn, M. and Locarnini, S.A., 2007. The phenylpropenamide derivative AT-130 blocks HBV replication at the level of viral RNA packaging. Antiviral Research, 76(2), pp. 168-177.

Flink, H.J., Van Zonneveld, M., Hansen, B.E., de Man, R.A., Schalm, S.W. and Janssen, H.L.A., 2006. Treatment with peg-interferon α-2b for HBeAg-positive chronic hepatitis B: HBsAg loss is associated with HBV genotype. American Journal of Gastroenterology, 101(2), pp. 297-303.

G

Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M. and Green, J., 2007. HepSEQ: international public health repository for hepatitis B. Nucleic Acids Research, 35(suppl 1), pp. D367-D370.

Gerlich, W.H., 2013. Medical Virology of Hepatitis B: how it began and where we are now?. Virology Journal, 10(1), pp. 239.

Giladi, H., Ketzinel-Gilad, M., Rivkin, L., Felig, Y., Nussbaum, O. and Galun, E., 2003. Small interfering RNA inhibits hepatitis B virus replication in mice. Molecular therapy, 8(5), pp. 769-776.

Gish, R. and Adams, P.C., 2009. Therapy for hepatitis B:'La nouvelle vague'. Canadian Journal of Gastroenterology, 23(6), pp. 407.

Gopalakrishnan, D., Keyter, M., Shenoy, K.T., Leena, K.B., Thayumanavan, L., Thomas, V., Vinayakumar, K., Panackel, C., Korah, A.T. and Nair, R., 2013. Hepatitis B virus subgenotype A1 predominates in liver disease patients from Kerala, India. **19**(48), pp. 9294.

Gulube, Z., Chirara, M., Kew, M., Tanaka, Y., Mizokami, M. and Kramvis, A., 2011. Molecular characterization of hepatitis B virus isolates from Zimbabwean blood donors. Journal of Medical Virology, 83(2), pp. 235-244.

Günther, S., Li, B., Miska, S., Kruger, D.H., Meisel, H. and Will, H., 1995. A Novel Method for Efficient Amplification of Whole Hepatitis B Virus Genomes Permits Rapid Functional Analysis and Reveals Deletion Mutants in Immunosuppressed Patients. Journal of Virology, 69(9), pp. 5437-5444.

H

Hadziyannis, S.J., 2011. Natural history of chronic hepatitis B in Euro-Mediterranean and African countries. Journal of Hepatology, 55(1), pp. 183-191.

Hadziyannis, S.J. and Papatheodoridis, G.V., 2006. Hepatitis B e antigen-negative chronic hepatitis B: natural history and treatment, Seminars in liver disease. 2006,

Copyright© 2006 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA, pp. 130-141.

Hadziyannis, S.J. and Vassilopoulos, D., 2001. Hepatitis B e antigen - negative chronic hepatitis B. Hepatology, 34(4), pp. 617-624.

Hannoun, C., Norder, H. & Lindh, M. 2000, "An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam", The Journal of general virology, 81(9), pp. 2267-2272.

Hayer, J., Jadeau, F., Deléage, G., Kay, A., Zoulim, F. and Combet, C., 2013. HBVdb: a knowledge database for Hepatitis B Virus. Nucleic Acids Research, 41(D1), pp. D566-D570.

Herrmann, F., Romero, M.R., Blazquez, A.G., Kaufmann, D., Ashour, M.L., Kahl, S., Marin, J.J.G., Efferth, T. and Wink, M., 2011. Diversity of Pharmacological Properties in Chinese and European Medicinal Plants: Cytotoxicity, Antiviral and Antitrypanosomal Screening of 82 Herbal Drugs. Diversity, 3(4), pp. 547-580.

Hinkes, M.J., 2009. Race, Ethnicity, and Forensic Anthropology. , pp. 1348.

Hillis, D.M. and Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst.Biol., 42(2), pp. 182-192.

Hubschen, J.M., Mbah, P.O., Forbi, J.C., Otegbayo, J.A., Olinger, C.M., Charpentier, E. and Muller, C.P., 2011. Detection of a new subgenotype of

hepatitis B virus genotype A in Cameroon but not in neighbouring Nigeria. Clinical Microbiology and Infections, 17(1), pp. 88-94.

Hunt, C.M., McGill, J.M., Allen, M.I. and Condreay, L.D., 2000. Clinical relevance of hepatitis B viral mutations. Hepatology, 31(5), pp. 1037-1044.

Hussain, Z., 2013. Genomic Heterogeneity of Hepatitis Viruses (A-E): Role in Clinical Implications and Treatment. In: G. Serviddio, ed, Practical Management of Chronic Viral Hepatitis. InTech, .

J

Janssen, H.L.A., van Zonneveld, M., Senturk, H., Zeuzem, S., Akarca, U.S., Cakaloglu, Y., Simon, C., So, T.M.K., Gerken, G. and de Man, R.A., 2005. Pegylated interferon alfa-2b alone or in combination with lamivudine for HBeAg-positive chronic hepatitis B: a randomised trial. The Lancet, 365(9454), pp. 123-129.

K

Kahila Bar-Gal, G., Kim, M.J., Klein, A., Shin, D.H., Oh, C.S., Kim, J.W., Kim, T., Kim, S.B., Grant, P.R. and Pappo, O., 2012. Tracing hepatitis B virus to the 16th century in a Korean mummy. Hepatology, 56(5), pp. 1671-1680.

Kampmann, M., Fordyce, S.L., Ávila-Arcos, M.C., Rasmussen, M., Willerslev, E., Nielsen, L.P. and Gilbert, M.T.P., 2011. A simple method for the parallel deep

sequencing of full influenza A genomes. Journal of Virology Methods, 178(1), pp. 243-248.

Kao, J., Wu, N., Chen, P., Lai, M. and Chen, D., 2000. Hepatitis B genotypes and the response to interferon therapy. Journal Hepatology, 33(6), pp. 998-1002.

Kar, P., Asim, M., Sarma, M.P. and Patki, P.S., 2009. HD-03/ES: A promising herbal drug for HBV antiviral therapy. Antiviral Research, 84(3), pp. 249-253.

Kato, H., Orito, E., Gish, R.G., Sugauchi, F., Suzuki, S., Ueda, R., Miyakawa, Y. and Mizokami, M., 2002. Characteristics of hepatitis B virus isolates of genotype G and their phylogenetic differences from the other six genotypes (A through F). Journal of Virology, 76(12), pp. 6131-6137.

Katoh, K., Asimenos, G. and Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT. Methods in Molecular Biology, 537, pp. 39-64.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S. and Duran, C., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), pp. 1647-1649.

Keating, G.M. and Noble, S., 2003. Recombinant hepatitis B vaccine (Engerix-B): a review of its immunogenicity and protective efficacy against hepatitis B. Drugs, 63(10), pp. 1021-1051.

Kimbi, G.C., Kramvis, A. and Kew, M.C., 2004. Distinctive sequence characteristics of subgenotype A1 isolates of hepatitis B virus from South Africa. Journal of General Virology, 85(Pt 5), pp. 1211-1220.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), pp. 111-120.

Kramvis, A., 2008. Molecular characterisation of the genotypes and mutants of hepatitis B virus from South Africa. South African Journal on the Epidemiology of Infections, 23(1), pp. 29-32.

Kramvis, A., Bukofzer, S., Kew, M.C. and Song, E., 1997. Nucleic acid sequence analysis of the precore region of hepatitis B virus from sera of southern African black adult carriers of the virus. Hepatology, 25(1), pp. 235-240.

Kramvis, A. and Kew, M.C., 2002. Structure and function of the encapsidation signal of hepadnaviridae. Journal of Viral Hepatitis, 5(6), pp. 357-367.

Kramvis, A. and Kew, M.C., 2005. Relationship of genotypes of hepatitis B virus to mutations, disease progression and response to antiviral therapy. Journal of Viral Hepatitis, 12(5), pp. 456-464.

Kramvis, A. and Kew, M.C., 2007. Epidemiology of hepatitis B virus in Africa, its genotypes and clinical associations of genotypes. Hepatology Research, 37, pp. S9-S19.

Kramvis, A., Kew, M.C. and Bukofzer, S., 1998. Hepatitis B virus precore mutants in serum and liver of Southern African Blacks with hepatocellular carcinoma. J.Hepatol., 28(1), pp. 132-141.

Kramvis, A., Kew, M.C. and François, G., 2005. Hepatitis B virus genotypes. Vaccine, 23(19), pp. 2409-2423.

Kramvis, A., Weitzmann, L., Owiredu, W. K. B. A. and Kew, M.C., 2002. Analysis of the complete genome of subgroup A hepatitis B virus isolates from South Africa. Journal of General Virology, 83, pp. 835-839.

Kurbanov, F., Tanaka, Y., Fujiwara, K., Sugauchi, F., Mbanya, D., Zekeng, L., Ndembi, N., Ngansop, C., Kaptue, L., Miura, T., Ido, E., Hayami, M., Ichimura, H. and Mizokami, M., 2005. A new subtype (subgenotype) Ac (A3) of hepatitis B virus and recombination between genotypes A and E in Cameroon. Journal of General Virology, 86(7), pp. 2047-2056.

Kurbanov, F., Tanaka, Y., Kramvis, A., Simmonds, P. and Mizokami, M., 2008. When should "I" consider a new hepatitis B virus genotype? Journal of Virology, 82(16), pp. 8241-8242.

L

Lander, E.S. and Waterman, M.S., 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2(3), pp. 231-239.

Lau, G.K.K., Piratvisuth, T., Luo, K.X., Marcellin, P., Thongsawat, S., Cooksley, G., Gane, E., Fried, M.W., Chow, W.C. and Paik, S.W., 2005. Peginterferon Alfa-

2a, lamivudine, and the combination for HBeAg-positive chronic hepatitis B. New England Journal of Medicine, 352(26), pp. 2682-2695.

Legler, K., Strohmeyer, H., Ritter, S., Gerlich, W. and Thomssen, R., 1983. Kinetics, subtype specificity and immunoglobulin class of anti-HBs induced by hepatitis B vaccine. Developments in Biological Standardization, 54, pp. 179.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26(5), pp. 589-595.

Liaw, Y.F., Gane, E., Leung, N., Zeuzem, S., Wang, Y., Lai, C.L., Heathcote, E.J., Manns, M., Bzowej, N. and Niu, J., 2009. 2-Year GLOBE trial results: telbivudine Is superior to lamivudine in patients with chronic hepatitis B. Gastroenterology, 136(2), pp. 486.

Lo, Y.M. and Chiu, R.W., 2009. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. Clinical Chemistry, 55(4), pp. 607-608.

Locarnini, S.A. and Yuen, L., 2010. Molecular genesis of drug-resistant and vaccine-escape HBV mutants. Antiviral Therapy, 15(3 Pt B), pp. 451-461.

Lok, A.S., Akarca, U. and Greene, S., 1994. Mutations in the pre-core region of hepatitis B virus serve to enhance the stability of the secondary structure of the pre-genome encapsidation signal. Proc.Natl.Acad.Sci.U.S.A., 91(9), pp. 4077-4081.

Lukhwareni, A., 2008. Exploring the impact of Human Immunodeficiency Virus on Hepatitis B Virus diagnosis, Prevantion and Control in Co-infected Adult South African Patients on Highly Active Anti-Retroviral Therapy, University of Limpopo.

Lusida, M.I., Nugrahaputra, V.E., Soetjipto, Handajani, R., Nagano-Fujii, M., Sasayama, M., Utsumi, T. and Hotta, H., 2008. Novel subgenotypes of hepatitis B virus genotypes C and D in Papua, Indonesia. J.Clin.Microbiol., 46(7), pp. 2160-2166.

M

Machida, A. and Nakamura, T., 1991. Hepatitis B Vaccine. United States of America: Google Patents.

Makondo, E., Bell, T.G. and Kramvis, A., 2012. Genotyping and Molecular Characterization of Hepatitis B Virus from Human Immunodeficiency Virus-Infected Individuals in Southern Africa. PloS one, 7(9), pp. e46345.

Marcellin, P., Heathcote, E.J., Buti, M., Gane, E., de Man, R.A., Krastev, Z., Germanidis, G., Lee, S.S., Flisiak, R. and Kaita, K., 2008. Tenofovir disoproxil fumarate versus adefovir dipivoxil for chronic hepatitis B. New England Journal of Medicine, 359(23), pp. 2442-2455.

Mason, W., Gerlich, W., Taylor, J., Kann, M., Loeb, D., Sureau, S., Magnius, L. and Norder, H., 2012. Family - Hepadnaviridae. Virus Taxonomy, ninth report of

the international committee on taxonomy of viruses. Amsterdam: Elsevier, pp. 445-455.

Mayaphi, S.H., Roussow, T.M., Masemola, D.P., Olorunju, S.A., Mphahlele, M.J. and Martin, D.J., 2012. HBV/HIV co-infection: the dynamics of HBV in South African patients with AIDS. South African Medical Journal, 102(3 Pt 1), pp. 157-162.

Mayaphi, S.H., Martin, D.J., Mphahlele, M.J., Blackard, J.T. and Bowyer, S.M., 2013. Variability of the preC/C region of hepatitis B virus genotype A from a South African cohort predominantly infected with HIV. Journal of Medical Virology, 85, pp. 1883–1892.

McMahon, B.J., 2009. The influence of hepatitis B virus genotype and subgenotype on the natural history of chronic hepatitis B. Hepatology International, 3(2), pp. 334-342.

Mimms, L., 2005. Hepatitis B virus escape mutants: “pushing the envelope” of chronic hepatitis B virus infection. Hepatology, 21(3), pp. 884-887.

Mirandola, S., Sebastiani, G., Rossi, C., Velo, E., Erne, E.M., Vario, A., Tempesta, D., Romualdi, C., Campagnolo, D. and Alberti, A., 2012. Genotype-specific mutations in the polymerase gene of hepatitis B virus potentially associated with resistance to oral antiviral therapy. Antiviral Research, 96(3), pp. 422-429.

Moskovitz, D., Osiowy, C., Giles, E., Tomlinson, G. and Heathcote, E., 2005. Response to long-term lamivudine treatment (up to 5 years) in patients with severe chronic hepatitis B, role of genotype and drug resistance. Journal of Viral Hepatitis, 12(4), pp. 398-404.

Mulders, M.N., Venard, V., Njayou, M., Edorh, A.P., Bola Oyefolu, A.O., Kehinde, M.O., Muyembe Tamfum, J.J., Nebie, Y.K., Maiga, I., Ammerlaan, W., Fack, F., Omilabu, S.A., Le Faou, A. and Muller, C.P., 2004. Low genetic diversity despite hyperendemicity of hepatitis B virus genotype E throughout West Africa. Journal of Infectious Diseases, 190(2), pp. 400-408.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H., 1986. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction, Cold Spring Harbor Symposia on Quantitative Biology 1986, Cold Spring Harbor Laboratory Press, pp. 263-273.

Mulyanto, Depamede, S.N., Surayah, K., Tsuda, F., Ichiyama, K., Takahashi, M. and Okamoto, H., 2009. A nationwide molecular epidemiologicalstudyon hepatitis B virus in Indonesia: identification of two novel subgenotypes, B8 and C7. Archives of Virology, 154(7), pp. 1047-1059.

N

Nishijima, N., Marusawa, H., Ueda, Y., Takahashi, K., Nasu, A., Osaki, Y., Kou, T., Yazumi, S., Fujiwara, T. and Tsuchiya, S., 2012. Dynamics of hepatitis B virus quasispecies in association with nucleos (t) ide analogue treatment determined by ultra-deep sequencing. PloS one, 7(4), pp. e35052.

Norder, H., Courouce, A.M., Coursaget, P., Echevarria, J.M., Lee, S.D., Mushahwar, I.K., Robertson, B.H., Locarnini, S. and Magnius, L.O., 2004. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology, 47(6), pp. 289-309.

O

Okamoto, H., Tsuda, F., Akahane, Y., Sugai, Y., Yoshiba, M., Moriyama, K., Tanaka, T., Miyakawa, Y. and Mayumi, M., 1994. Hepatitis B virus with mutations in the core promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. Journal of Virology, 68(12), pp. 8102-8110.

Olinger, C.M., Venard, V., Njayou, M., Oyefolu, A.O., Maiga, I., Kemp, A.J., Omilabu, S.A., le Faou, A. and Muller, C.P., 2006. Phylogenetic analysis of the precore/core gene of hepatitis B virus genotypes E and A in West Africa: new subtypes, mixed infections and recombinations. Journal of General Virology, 87(Pt 5), pp. 1163-1173.

Osiowy, C., Gordon, D., Borlang, J., Giles, E. and Villeneuve, J.P., 2008. Hepatitis B virus genotype G epidemiology and co-infection with genotype A in Canada. Journal of General Virology, 89(Pt 12), pp. 3009-3015.

Owiredu, W.K.B.A., Kramvis, A. and Kew, M.C., 2001. Hepatitis B virus DNA in serum of healthy black African adults positive for hepatitis B surface antibody alone: possible association with recombination between genotypes A and D. Journal of Medical Virology, 64(4), pp. 441-454.

P

Packianathan, C., Katen, S.P., Dann, C.E. and Zlotnick, A., 2010. Conformational changes in the hepatitis B virus core protein are consistent with a role for allostery in virus assembly. Journal of Virology, 84(3), pp. 1607-1615.

Pan, X., Wei, L., Han, J., Ma, H., Deng, K. and Cong, X., 2011. Artificial recombinant cell-penetrating peptides interfere with envelopment of hepatitis B virus nucleocapsid and viral production. Antiviral Research, 89(1), pp. 109-114.

Perkins, J.A., 2002. Hepatitis B Virus. Medical and Scientific illustrations.

Perni, R., Conway, S., Ladner, S., Zaifert, K., Otto, M. and King, R., 2000. Phenylpropenamide derivatives as inhibitors of hepatitis B virus replication. Bioorganic Medicinal Chemistry Letters, 10(23), pp. 2687-2690.

Previsani, N. And Lavanchy, D., 2002-last update, Hepatitis B [Homepage of Global Alert and Response], [Online]. Available: URL: http://www.who.int/csr/disease/hepatitis/whocdscsrlyo20022/en/index.html [accessed November 28, 2013].

Prosperi, M.C. and Salemi, M., 2012. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics, 28(1), pp. 132-133.

Prosperi, M.C., Yin, L., Nolan, D.J., Lowe, A.D., Goodenow, M.M. and Salemi, M., 2013. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Scientific reports, 3: 2837. DOI: 10.1038/srep02837

Purcell, R.H. and Gerin, J.L., 1975. Hepatitis B subunit vaccine: a preliminary report of safety and efficacy tests in chimpanzees. American Journal of Medical Science, 270(2), pp. 395.

Q

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics, 13(1), pp. 341.

R

Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C. and Hall, N., 2012. Application of next-generation sequencing technologies in virology. Journal of General Virology, 93(Pt 9), pp. 1853-1868.

Raimondi, S., Maisonneuve, P., Bruno, S. and Mondelli, M.U., 2010. Is response to antiviral treatment influenced by hepatitis B virus genotype? Journal of Hepatology, 52(3), pp. 441-449.

Ribeiro, R., Germanidis, G., Powers, K., Pellegrin, B., Nikolaidis, P., Perelson, A. and Pawlotsky, J., 2010. Hepatitis B virus kinetics under antiviral therapy sheds light on differences in hepatitis B e antigen positive and negative infections. Journal of Infectious Diseases, 202(9), pp. 1309-1318.

Richterich, P., 1998. Estimation of errors in “raw” DNA sequences: a validation study. Genome Research, 8(3), pp. 251-259.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P., 2011. Integrative genomics viewer. Nature Biotechnology, 29(1), pp. 24-26.

S

Sauer, N.J., 1992. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? Social Science and Medicine, 34(2), pp. 107-111.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4), pp. 406-425.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. National Academy of Science, 74(12), pp. 5463-5467.

Schädler, S. and Hildt, E., 2009. HBV Life Cycle: Entry and Morphogenesis. Viruses, 1(2), pp. 185-209.

Schaefer, S., 2005. Hepatitis B virus: significance of genotypes. Journal of Viral Hepatitis, 12(2), pp. 111-124.

Schaefer, S., Magnius, L. and Norder, H., 2009. Underconstruction: classification of hepatitis B virus genotypes and subgenotypes. Intervirology, 52(6), pp. 323-325.

Schultz, A., Bulla, I., Abdou-Chekaraou, M., Gordien, E., Morgenstern, B., Zoulim, F., Deny, P. and Stanke, M., 2012. jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. Nucleic Acids Research, 40(W1), pp. W193-W198.

Schultz, U., Grgacic, E. And Nassal, M., 2004. Duck Hepatitis B Virus: An Invaluable Model System for HBV Infection. Advances in Virus Research, pp. 1-70.

Seddigh-Tonekaboni, S., Waters, J.A., Jeffers, S., Gehrke, R., Ofenloch, B., Horsch, A., Hess, G., Thomas, H.C. and Karayiannis, P., 2000. Effect of variation in the common “a” determinant on the antigenicity of hepatitis B surface antigen. Journal of Medical Virology, 60(2), pp. 113-121.

Seeger, C., Zoulim, F. and Mason, W., 2007. Hepadnaviruses. In: M. Knipe and P.M. Howley, eds, Field's virology. 4th edition edn. Philadelphia, PA: Lippincott Williams & Wilkins, pp. 2977-3029.

Sheldon, J., Ramos, B., Garcia-Samaniego, J., Rios, P., Bartholomeusz, A., Romero, M., Locarnini, S., Zoulim, F. and Soriano, V., 2007. Selection of hepatitis B virus (HBV) vaccine escape mutants in HBV-infected and HBV/HIV-coinfected patients failing antiretroviral drugs with anti-HBV activity. JAIDS Journal of Acquired Immune Deficiency Syndromes, 46(3), pp. 279-282.

Sheldon, J. and Soriano, V., 2008. Hepatitis B virus escape mutants induced by antiviral therapy. Journal of Antimicrobial Chemotherapy, 61(4), pp. 766-768.

Shen, T., Gao, J.M., Zou, Y.L., Dong, H. and Yan, X.M., 2009. Novel hepatitis B virus subgenotype in the southern Yunnan Province of China. Intervirology, 52(6), pp. 340-346.

Simmonds, P. & Midgley, S. 2005, "Recombination in the genesis and evolution of hepatitis B virus genotypes", Journal of virology, vol. 79, no. 24, pp. 15467-15476.

Simmonds, P., 2012. SSE: a nucleotide and amino acid sequence analysis platform. BMC Research Notes, 5, pp. 50.

Stannard, L.M., 1995-last update, Hepatitis B Virus [Homepage of University of Cape Town], [Online]. Available: http://web.uct.ac.za/depts/mmi/stannard/hepb.html [Accessed: 14 December, 2013].

Stramer, S.L., Wend, U., Candotti, D., Foster, G.A., Hollinger, F.B., Dodd, R.Y., Allain, J. and Gerlich, W., 2011. Nucleic acid testing to detect HBV infection in blood donors. New England Journal of Medicine, 364(3), pp. 236-247.

Stuyver, L., De Gendt, S., Van Geyt, C., Zoulim, F., Fried, M., Schinazi, R.F. and Rossau, R., 2000. A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. Journal of General Virology, 81(1), pp. 67-74.

Szpara, M.L., Parsons, L. and Enquist, L., 2010. Sequence variability in clinical and laboratory isolates of herpes simplex virus 1 reveals new mutations. Journal of Virology, 84(10), pp. 5303-5313.

T

Takahashi, K., Brotman, B., Usuda, S., Mishiro, S. and Prince, A.M., 2000. Full-genome sequence analyses of hepatitis B virus (HBV) strains recovered from chimpanzees infected in the wild: implications for an origin of HBV. Virology, 267(1), pp. 58-64.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution, 28(10), pp. 2731-2739.

Tapparel, C., Cordey, S., Junier, T., Farinelli, L., Van Belle, S., Soccal, P.M., Aubert, J., Zdobnov, E. and Kaiser, L., 2011. Rhinovirus genome variation during chronic upper and lower respiratory tract infections. PloS one, 6(6), pp. e21163.

Tatematsu, K., Tanaka, Y., Kurbanov, F., Sugauchi, F., Mano, S., Maeshiro, T., Nakayoshi, T., Wakuta, M., Miyakawa, Y. and Mizokami, M., 2009. A genetic variant of hepatitis B virus divergent from known human and ape genotypes isolated from a Japanese patient and provisionally assigned to new genotype J. Journal of Virology, 83(20), pp. 10538-10547.

Tanaka, Y., Hasegawa, I., Kato, T., Orito, E., Hirashima, N., Acharya, S.K., Gish, R.G., Kramvis, A., Kew, M.C. and Yoshihara, N., 2004. A case-control study for differences among hepatitis B virus infections of genotypes A (subtypes Aa and Ae) and D. Hepatology, 40(3), pp. 747-755.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. 14(2), pp. 178-192.

Toh, S.T., Jin, Y., Liu, L., Wang, J., Babrzadeh, F., Gharizadeh, B., Ronaghi, M., Toh, H.C., Chow, P.K. and Chung, A.Y., 2013. Deep sequencing of the hepatitis B virus in hepatocellular carcinoma patients reveals enriched integration events, structural alterations and sequence variations. Carcinogenesis, 34(4), pp. 787-798.

Tran, T.T., Trinh, T.N. and Abe, K., 2008. New complex recombinant genotype of hepatitis B virus identified in Vietnam. Journal of Virology, 82(11), pp. 5657-5663.

V

Vieth, S., Manegold, C., Drosten, C., Nippraschk, T. and Günther, S., 2002. Sequence and Phylogenetic Analysis of Hepatitis B Virus Genotype G Isolated in Germany. Virus Genes, 24(2), pp. 153-156.

Viral Zone, 2011-last update, HBV envelope proteins [Homepage of Swiss Institute for Bioinformatics], [Online]. Available:

http://viralzone.expasy.org/all_by_protein/1405.html [Accessed: 14 December, 2013]

Vrancken, B., Lequime, S., Theys, K. and Lemey, P., 2010. Covering all bases in HIV research: unveiling a hidden world of viral evolution. AIDS Reviews, 12(2), pp. 89-102.

W

Wang, G.F., Shi, L.P., Ren, Y.D., Liu, Q.F., Liu, H.F., Zhang, R.J., Li, Z., Zhu, F.H., He, P.L., Tang, W., Tao, P.Z., Li, C., Zhao, W.M. and Zuo, J.P., 2009. Anti-hepatitis B virus activity of chlorogenic acid, quinic acid and caffeic acid in vivo and in vitro. Antiviral Research, 83(2), pp. 186-190.

Waters, J., Kennedy, M., Voet, P., Hauser, P., Petre, J., Carman, W. and Thomas, H., 1992. Loss of the common" A" determinant of hepatitis B surface antigen by a vaccine-induced escape mutant. Journal of Clinical Investigations, 90(6), pp. 2543.

Watts, N., Vethanayagam, J., Ferns, R., Tedder, R., Harris, A., Stahl, S., Steven, A. and Wingfield, P., 2010. Molecular basis for the high degree of antigenic cross-reactivity between hepatitis B virus capsids (HBcAg) and dimeric capsid-related protein (HBeAg): insights into the enigmatic nature of the e-antigen. Journal of Molecular Biology, 398(4), pp. 530-541.

Weber, B., 2005. Genetic variability of the S gene of hepatitis B virus: clinical and diagnostic impact. Journal of Clinical Virology, 32(2), pp. 102-112.

Weber, O., Schlemmer, K.-., Hartmann, E., Hagelschuer, I., Paessens, A., Graef, E., Deres, K., Goldmann, S., Niewoehner, U., Stoltefuss, J., Haebich, D., Ruebsamen-Waigmann, H. and Wohlfeil, S., 2002. Inhibition of human hepatitis B virus (HBV) by a novel non-nucleosidic compound in a transgenic mouse model. Antiviral Research, 54(2), pp. 69-78.

Wong, V.C.W., Reesink, H.W., Ip, H.M.H., Nco Lelie, P., Reerink-Brongers, E.E., Yeung, C. and Ma, H., 1984. Prevention of the HBsAg carrier state in newborn infants of mothers who are chronic carriers of HBsAg and HBeAg by administration of hepatitis-B vaccine and hepatitis-B immunoglobulin: double-blind randomised placebo-controlled study. The Lancet, 323(8383), pp. 921-926.

World Health Organization, July, 2013-last update, Hepatitis B Fact sheet [Homepage of World Health Organization], [Online]. Available: http://www.who.int/mediacentre/factsheets/fs204/en/index.html [accessed: 30 September, 2013].

Wynne, S.A., Crowther, R.A. and Leslie, A.G.W., 1999. The Crystal Structure of the Human Hepatitis B Virus Capsid. Molecular Cell, 3(6), pp. 771-780.

Y

Yamamoto, K., Horikita, M., Tsuda, F., Itoh, K., Akahane, Y., Yotsumoto, S., Okamoto, H., Miyakawa, Y. and Mayumi, M., 1994. Naturally occurring escape mutants of hepatitis B virus with various mutations in the S gene in carriers seropositive for antibody to hepatitis B surface antigen. Journal of Virology, 68(4), pp. 2671-2676.

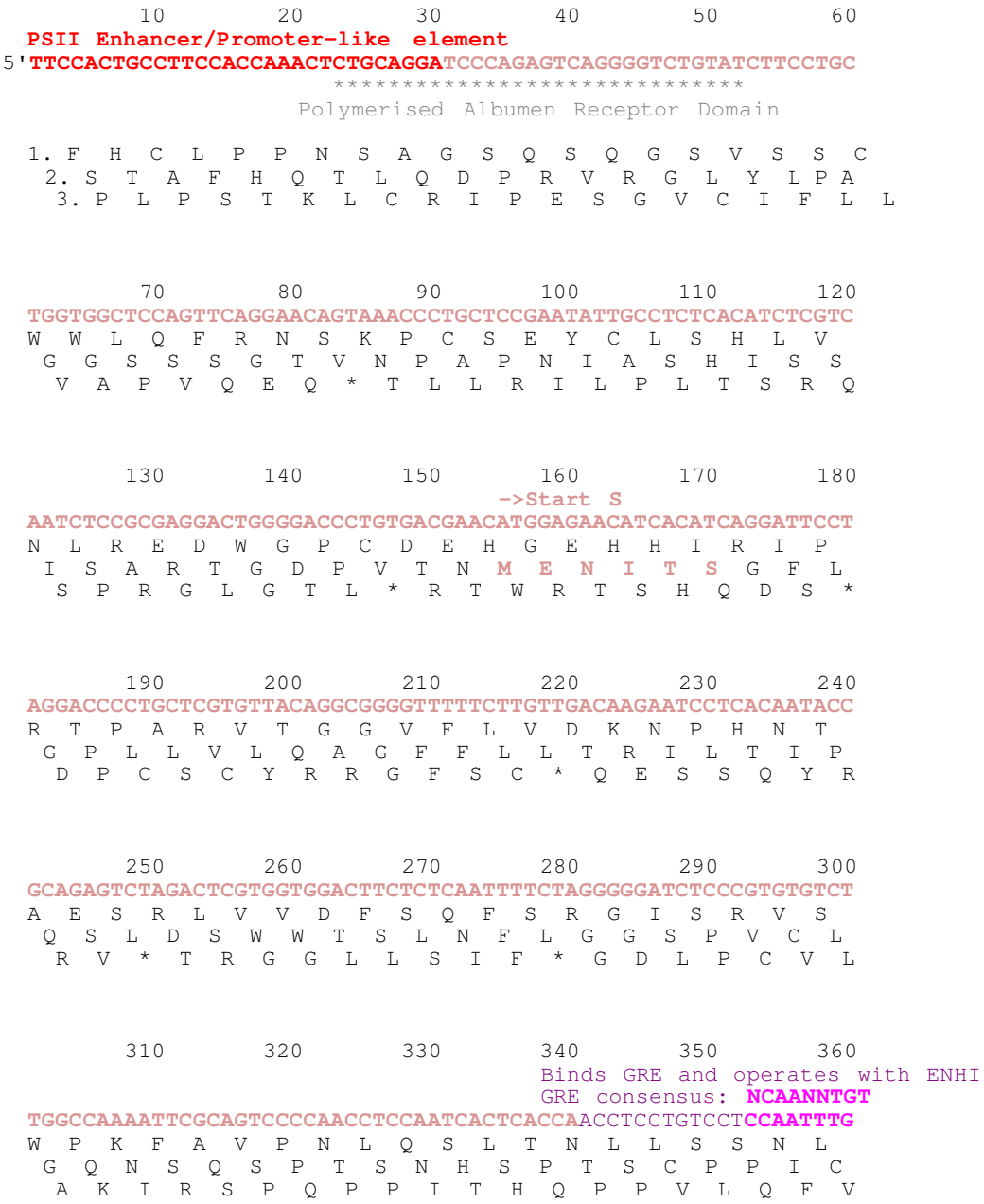
Z

Zhu, Y., Jin, Y., Guo, X., Bai, X., Chen, T., Wang, J., Qian, G., Groopman, J.D., Gu, J., Li, J. and Tu, H., 2010. Comparison study on the complete sequence of hepatitis B virus identifies new mutations in core gene associated with hepatocellular carcinoma. Cancer Epidemiology, Biomarkers and Prevention, 19(10), pp. 2623-2630.

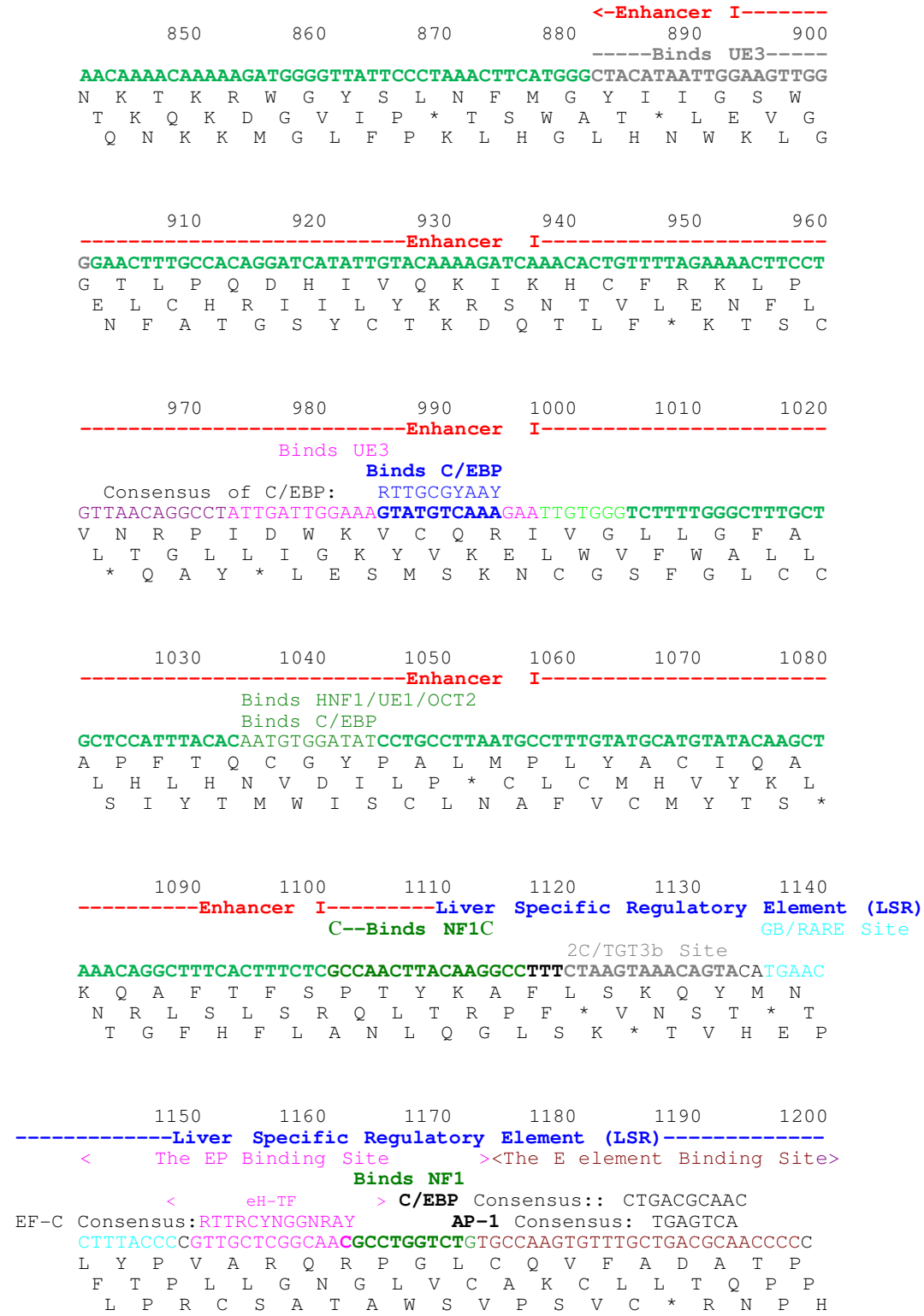
Zöllner, B., Petersen, J., Schäfer, P., Schröter, M., Laufs, R., Sterneck, M. & Feucht, H.H. 2002, "Subtype-dependent response of hepatitis B virus during the early phase of lamivudine treatment", Clinical infectious diseases, 34(9), pp. 1273-1277.

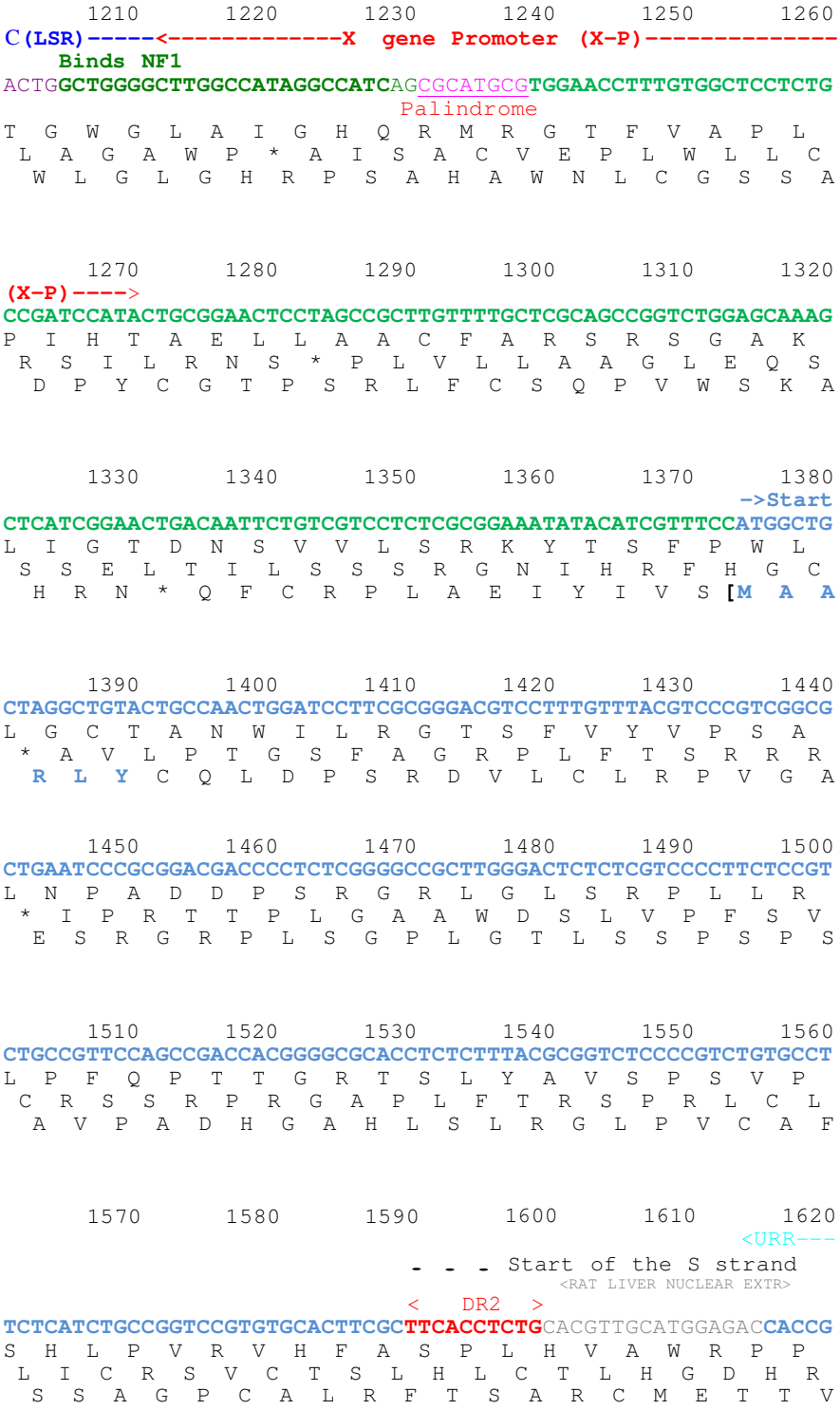
APPENDIX A

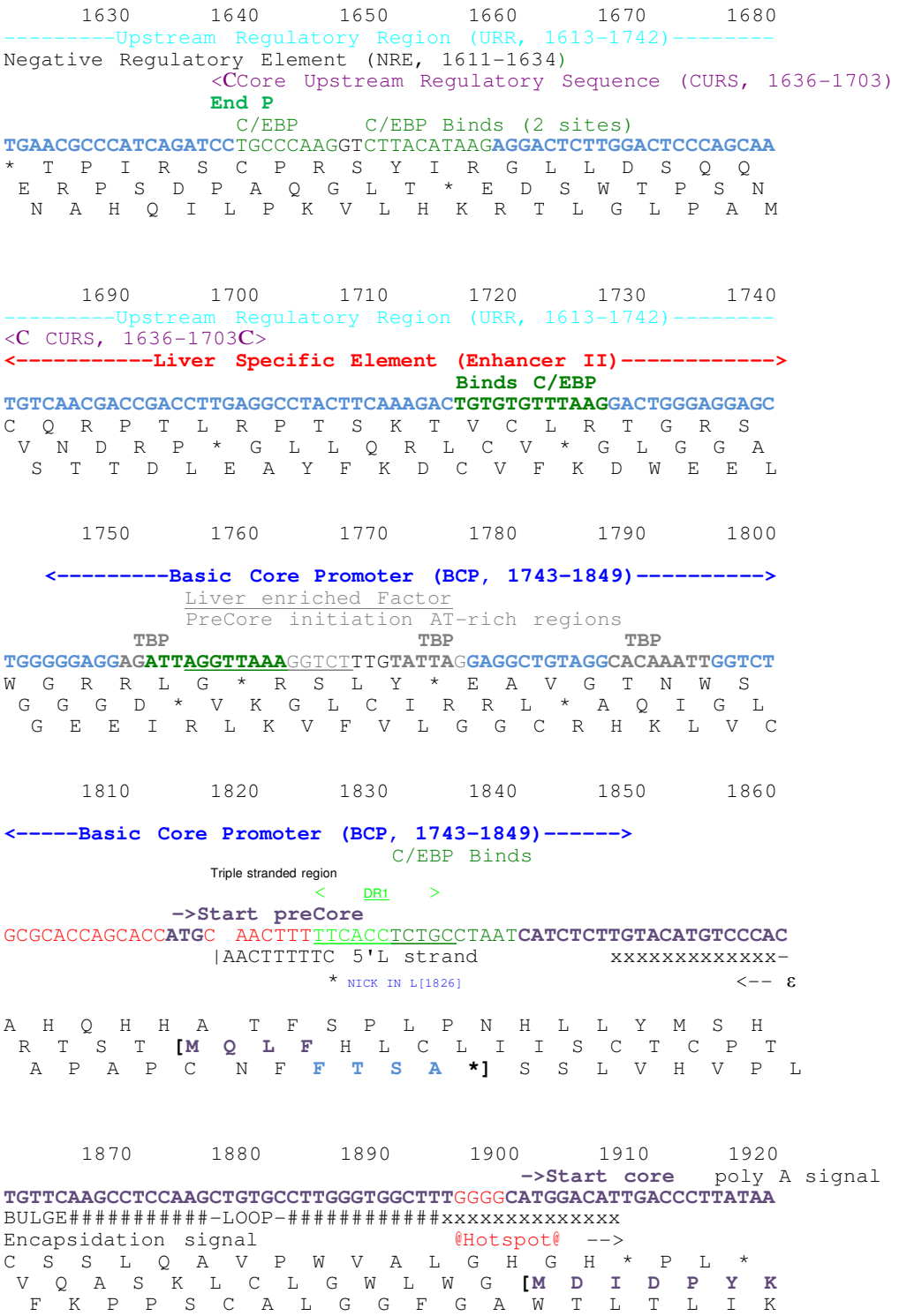
THE POSITION OF ORFs AND BINDING SITES ON THE HBV GENOME SHOWN ON GENBANK SEQUENCE ACCESSION NUMBER X02763



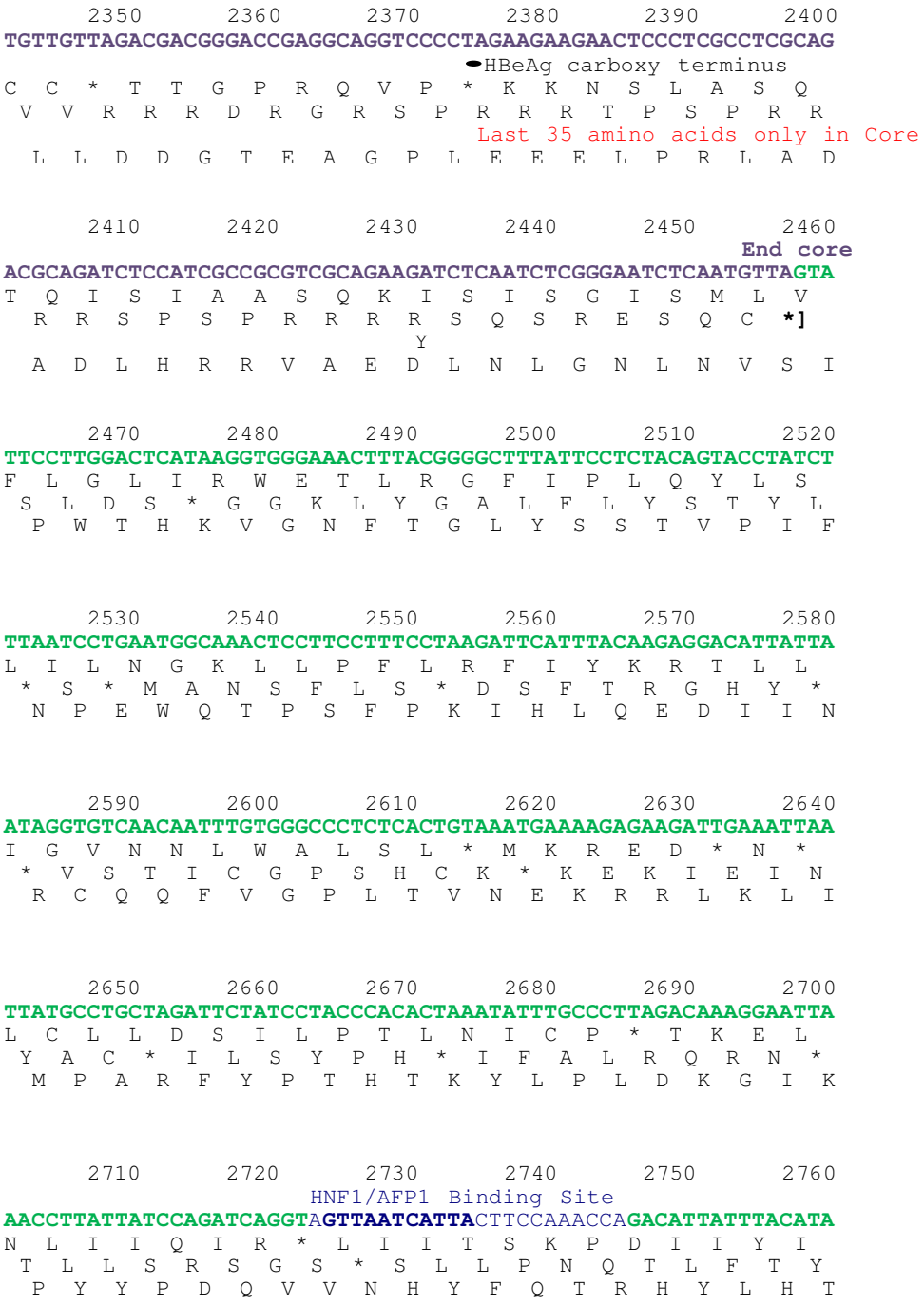
```
       370        380        390        400        410        420
TCCTGGTTATCGCTGGATGTGTCTGCGGCGTTTTATCATATTCCTCTTCATCCTGCTGCT
S  W  L  S  L  D  V  S  A  A  F  Y  H  I  P  L  H  P  A  A
 P  G  Y  R  W  M  C  L  R  R  F  I  I  F  L  F  I  L  L  L
  L  V  I  A  G  C  V  C  G  V  L  S  Y  S  S  S  S  C  C  Y

       430        440        450        460        470        480
ATGCCTCATCTTCTTATTGGTTCTTCTGGATTATCAAGGTATGTTGCCCGTTTGTCCTCT
M  P  H  L  L  I  G  S  S  G  L  S  R  Y  V  A  R  L  S  S
 C  L  I  F  L  L  V  L  L  D  Y  Q  G  M  L  P  V  C  P  L
  A  S  S  S  Y  W  F  F  W  I  I  K  V  C  C  P  F  V  L  *

       490        500        510        520        530        540
                                        d/y
AATTCCAGGATCAACAACAACCAGTACGGGACCATGCAAAACCTGCACGACTCCTGCTCA
N  S  R  I  N  N  N  Q  Y  G  T  M  Q  N  L  H  D  S  C  S
 I  P  G  S  T  T  T  S  T  G  P  C  K  T  C  T  T  P  A  Q
  F  Q  D  Q  Q  Q  P  V  R  D  H  A  K  P  A  R  L  L  L  K

       550        560        570        580        590        600
            <-aa 139-147 is often referred to as the 'a' epitope B>
AGGCAACTCTATGTTTCCCTCATGTTGCTGTACAAAACCTACGGATGGAAATTGCACCTG
R  Q  L  Y  V  S  L  M  L  L  Y  K  T  Y  G  W  K  L  H  L
 G  N  S  M  F  P  S  C  C  C  T  K  P  T  D  G  N  C  T  C
  A  T  L  C  F  P  H  V  A  V  Q  N  L  R  M  E  I  A  P  V

       610        620        630        640        650        660
                                  r/w
TATTCCCATCCCATCGTCCTGGGCTTTCGCAAAATACCTATGGGAGTGGGCCTCAGTCCG
Y  S  H  P  I  V  L  G  F  R  K  I  P  M  G  V  G  L  S  P
 I  P  I  P  S  S  W  A  F  A  K  Y  L  W  E  W  A  S  V  R
  F  P  S  H  R  P  G  L  S  Q  N  T  Y  G  S  G  P  Q  S  V

       670        680        690        700        710        720
TTTCTCTTGGCTCAGTTTACTAGTGCCATTTGTTCAGTGGTTCGTAGGGCTTTCCCCCAC
F  L  L  A  Q  F  T  S  A  I  C  S  V  V  R  R  A  F  P  H
 F  S  W  L  S  L  L  V  P  F  V  Q  W  F  V  G  L  S  P  T
  S  L  G  S  V  Y  *  C  H  L  F  S  G  S  *  G  F  P  P  L

       730        740        750        760        770        780
TGTTTGGCTTTCAGCTATATGGATGATGTGGTATTGGGGGCCAAGTCTGTACAGCATCGT
C  L  A  F  S  Y  M  D  D  V  V  L  G  A  K  S  V  Q  H  R
 V  W  L  S  A  I  W  M  M  W  Y  W  G  P  S  L  Y  S  I  V
  F  G  F  Q  L  Y  G  *  C  G  I  G  G  Q  V  C  T  A  S  *

       790        800        810        820        830        840
                                                    End S
GAGTCCCTTTATACCGCTGTTACCAATTTTCTTTTGTCTCTGGGTATACATTTAAACCCT
E  S  L  Y  T  A  V  T  N  F  L  L  S  L  G  I  H  L  N  P
 S  P  F  I  P  L  L  P  I  F  F  C  L  W  V  Y  I  *]  T  L
  V  P  L  Y  R  C  Y  Q  F  S  F  V  S  G  Y  T  F  K  P  *
```



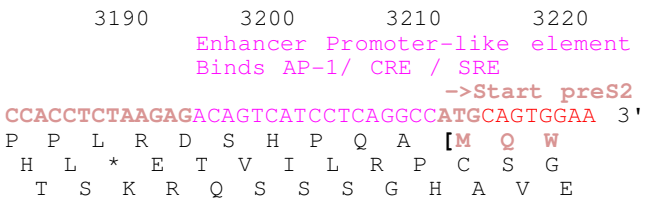




```
      1930       1940       1950       1960       1970       1980
                                GT cluster involved in poly A addition
AGAATTTGGAGCTACTGTGGAGTTACTCTCGTTTTTGCCTTCTGACTTCTTTCCTTCCGT
R  I  W  S  Y  C  G  V  T  L  V  F  A  F  *  L  L  S  F  R
 E  F  G  A  T  V  E  L  L  S  F  L  P  S  D  F  F  P  S  V
  N  L  E  L  L  W  S  Y  S  R  F  C  L  L  T  S  F  L  P  S

      1990       2000       2010       2020       2030       2040
CAGAGATCTCCTAGACACCGCCTCAGCTCTGTATCGAGAAGCCTTAGAGTCTCCTGAGCA
Q  R  S  P  R  H  R  L  S  S  V  S  R  S  L  R  V  S  *  A
 R  D  L  L  D  T  A  S  A  L  Y  R  E  A  L  E  S  P  E  H
  E  I  S  *  T  P  P  Q  L  C  I  E  K  P  *  S  L  L  S  I


      2050       2060       2070       2080       2090       2100
TTGCTCACCTCACCATACTGCACTCAGGCAAGCCATTCTCTGCTGGGGGGAATTGATGAC
L  L  T  S  P  Y  C  T  Q  A  S  H  S  L  L  G  G  I  D  D
 C  S  P  H  H  T  A  L  R  Q  A  I  L  C  W  G  E  L  M  T
  A  H  L  T  I  L  H  S  G  K  P  F  S  A  G  G  N  *  *  L


      2110       2120       2130       2140       2150       2160
                                        <--------  e1 epitope---
TCTAGCTACCTGGGTGGGTAATAATTTGGAAGATCCAGCATCTAGGGATCTTGTAGTAAA
S  S  Y  L  G  G  *  *  F  G  R  S  S  I  *  G  S  C  S  K
 L  A  T  W  V  G  N  N  L  E  D  P  A  S  R  D  L  V  V  N
  *  L  P  G  W  V  I  I  W  K  I  Q  H  L  G  I  L  *  *  I


      2170       2180       2190       2200       2210       2220
--->
TTATGTTAATACTAACGTGGGTTTAAAGATCAGGCAACTATTGTGGTTTCATATATCTTG
L  C  *  Y  *  R  G  F  K  D  Q  A  T  I  V  V  S  Y  I  L
 Y  V  N  T  N  V  G  L  K  I  R  Q  L  L  W  F  H  I  S  C
  M  L  I  L  T  W  V  *  R  S  G  N  Y  C  G  F  I  Y  L  A


      2230       2240       2250       2260       2270       2280
CCTTACTTTTGGAAGAGAGACTGTACTTGAATATTTGGTCTCTTTCGGAGTGTGGATTCG
P  Y  F  W  K  R  D  C  T  *  I  F  G  L  F  R  S  V  D  S
 L  T  F  G  R  E  T  V  L  E  Y  L  V  S  F  G  V  W  I  R
  L  L  L  E  E  R  L  Y  L  N  I  W  S  L  S  E  C  G  F  A


      2290       2300       2310       2320       2330       2340
       <------------e2  epitope--------->
                          ->Start P
CACTCCTCCAGCCTATAGACCACCAAATGCCCCTATCTTATCAACACTTCCGGAAACTAC
H  S  S  S  L  *  T  T  K  C  P  Y  L  I  N  T  S  G  N  Y
 T  P  P  A  Y  R  P  P  N  A  P  I  L  S  T  L  P  E  T  T
  L  L  Q  P  I  D  H  Q  [M  P  L  S  Y  Q  H  F  R  K  L  L
```



```
      2770       2780       2790       2800       2810       2820
                          (The) TATA BOX
CTCTTTGGAAGGCTGGTATTCTATATAAGCGGGAAACCACACGTAGCGCATCATTTTGCG
L  F  G  R  L  V  F  Y  I  S  G  K  P  H  V  A  H  H  F  A
 S  L  E  G  W  Y  S  I  *  A  G  N  H  T  *  R  I  I  L  R
  L  W  K  A  G  I  L  Y  K  R  E  T  T  R  S  A  S  F  C  G

      2830       2840       2850       2860       2870       2880
                                       -----33-bp-deletion------
                                  ->preS1 Start
GGTCACCATATTCTTGGGAACAAGAGCTACAGCATGGGAGGTTGGTCATCAAAACCTCGC
G  H  H  I  L  G  N  K  S  Y  S  [M  G  G  W  S  S  K  P  R
 V  T  I  F  L  G  T  R  A  T  A  W  E  V  G  H  Q  N  L  A
  S  P  Y  S  W  E  Q  E  L  Q  H  G  R  L  V  I  K  T  S  Q

      2890       2900       2910       2920       2930       2940
AAAGGCATGGGGACGAATCTTTCTGTTCCCAATCCTCTGGGATTCTTTCCCGATCATCAG
K  G  M  G  T  N  L  S  V  P  N  P  L  G  F  F  P  D  H  Q
 K  A  W  G  R  I  F  L  F  P  I  L  W  D  S  F  P  I  I  S
  R  H  G  D  E  S  F  C  S  Q  S  S  G  I  L  S  R  S  S  V

      2950       2960       2970       2980       2990       3000
TTGGACCCTGCATTCGGAGCCAACTCAAACAATCCAGATTGGGACTTCAACCCCGTCAAG
L  D  P  A  F  G  A  N  S  N  N  P  D  W  D  F  N  P  V  K
 W  T  L  H  S  E  P  T  Q  T  I  Q  I  G  T  S  T  P  S  R
  G  P  C  I  R  S  Q  L  K  Q  S  R  L  G  L  Q  P  R  Q  G

      3010       3020       3030       3040       3050       3060
     NF1 Binding Site
GACGACTGGCCAGCAGCCAACCAAGTAGGAGTGGGAGCATTCGGGCCAAGGCTCACCCCT
D  D  W  P  A  A  N  Q  V  G  V  G  A  F  G  P  R  L  T  P
 T  T  G  Q  Q  P  T  K  *  E  W  E  H  S  G  Q  G  S  P  L
  R  L  A  S  S  Q  P  S  R  S  G  S  I  R  A  K  A  H  P  S

      3070       3080       3090       3100       3110       3120
CCACACGGCGGTATTTTGGGGTGGAGCCCTCAGGCTCAGGGCATATTGACCACAGTGTCA
P  H  G  G  I  L  G  W  S  P  Q  A  Q  G  I  L  T  T  V  S
 H  T  A  V  F  W  G  G  A  L  R  L  R  A  Y  *  P  Q  C  Q
  T  R  R  Y  F  G  V  E  P  S  G  S  G  H  I  D  H  S  V  N

      3130       3140       3150       3160       3170       3180
Activates PreS/S Promoter in liver cells but deactivates HELA cells
ACAATTCCTCCTCCTGCCTCCACCAATCGGCAGTCAGGAAGGCAGCCTACTCCCATCTCT
T  I  P  P  P  A  S  T  N  R  Q  S  G  R  Q  P  T  P  I  S
 Q  F  L  L  L  P  P  P  I  G  S  Q  E  G  S  L  L  P  S  L
  N  S  S  S  C  L  H  Q  S  A  V  R  K  A  A  Y  S  H  L  S
```



*Adapted from the thesis of S.M. Bowyer (1997)

APPENDIX B

EXAMPLE OF A QURE RUN

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.


C:\Users\Stephane>cd Documents


C:\Users\Stephane\Documents>cd QuRe_v0.99971


C:\Users\Stephane\Documents\QuRe_v0.99971>"C:\Program
Files\Java\jdk1.7.0_25\bin\java" -cp . -Xmx8G QuRe
"C:\Users\Stephane\Documents\QuRe_v0.99971\1. Data\Sample
12\3358.fasta" "C:\Users\Stephane\Documents\QuRe_v0.99971\2.
Refs\AY233277 A1(D).fas"

----------------------------------------------------------------------
----------

----------------------------------------------------------------------
----------

parallel processing enabled: no. of cores available = 7

parsing "C:\Users\Stephane\Documents\QuRe_v0.99971\1. Data\Sample
12\3358.fasta"

 read file 100%

        148152 reads

        average (st.dev.) read length is 227 (48)

parsing "C:\Users\Stephane\Documents\QuRe_v0.99971\2.
Refs\AY233277 A1(D).fas" r

Reference genome file 100%

        >gi|32330490|gb|AY233277.1| Hepatitis B virus isolate 1848
complete genome read (3221 bases)

        building dictionary 100%

        calculating quasi-random alignment score distribution 100%

        average (st.dev) quasi-random score is 67 (14)
```

```
aligning reads to reference genome 100%
        time employed = 511588 ms
        removing 147515 reads with alignment p-value > 0.01
        637 reads retained
reconstructing consensus genome and variations 100%
        average (st.dev.) coverage of each mapped base is 45 (14)
        correcting mapped reads, SNP and indel list 100%
        post-alignment (st.dev.) read length is 221 (68)
        reference genome is covered from position 1 to 3221
        407 reads spanning the high-coverage window
alignment and mapping time = 568527 ms
starting Quasispecies Reconstruction (QuRe)
        phase 0: fixed-size sliding window overlaps 100%
        phase 1: random overlaps 100%
        phase 3: assessing best a-posteriori overlaps set
                overlaps space n=10082
                avg. (std) min. interval coverage 6.5 (3.72)
                avg. (std) interval coverage 26.91 (3.56)
                avg. (std) min. overlaps diversity 0.0 (0.0)
                avg. (std) overlaps diversity 0.06 (0.02)
                avg. (std) frac.non-zerodivers.overl. 0.82 (0.07)
                avg. (std) min. overlap length 10.01 (26.37)
                avg. (std) overlap length 44.77 (18.66)
                avg. (std) num. intervals 29.53 (36.39)
                avg. (std) min amplicon length 51.03 (23.01)
                avg. (std) amplicon length 112.62 (17.2)
                max. a-posteriori overlaps set:
                min. interval coverage 6.0 (post.prob. = 0.6)
                avg. interval coverage 28.78 (post.prob. = 0.73)
```

```
min. overlap diversity 0.0 (post.prob. = 0.92)
avg. overlap diversity 0.07 (post.prob. = 0.85)
frac.non-zerodivers.overl.0.91(post.prob. = 0.85)
min. overlap length 6.66 (post.prob. = 0.82)
avg. overlap length 44.74 (post.prob. = 0.66)
number of intervals = 23 (post.prob. = 0.7)
min. amplicon length = 50.0 (post.prob. = 0.89)
avg. amplicon length = 117.14 (post.prob. = 0.7)
       1-164
       139-275
       242-340
       317-471
       449-615
       487-704
       635-729
       705-860
       854-992
       955-1025
       972-1143
       1076-1155
       1134-1326
       1284-1368
       1309-1445
       1429-1487
       1445-1524
       1505-1583
       1544-1594
       1554-1619
       1561-1637
```

```
                    1575-1679

                    1582-1711

                    ---------

          executing core reconstruction algorithm

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                   reconstruction(s) done.

                  | reconstruction(s) done.

                  initial number of variants = 4

                  final clustering (random search + BIC* selection)

final number of variants = 3

amplicon estimation and quasispecies reconstruction time = 329348
ms

total time employed = 897875 ms

--------------------------------------------------------------------
----------

--------------------------------------------------------------------
----------
```

*BIC – Bayesian Information Criterion

APPENDIX C

TABLE OF REFERENCES USED IN ANALYSIS

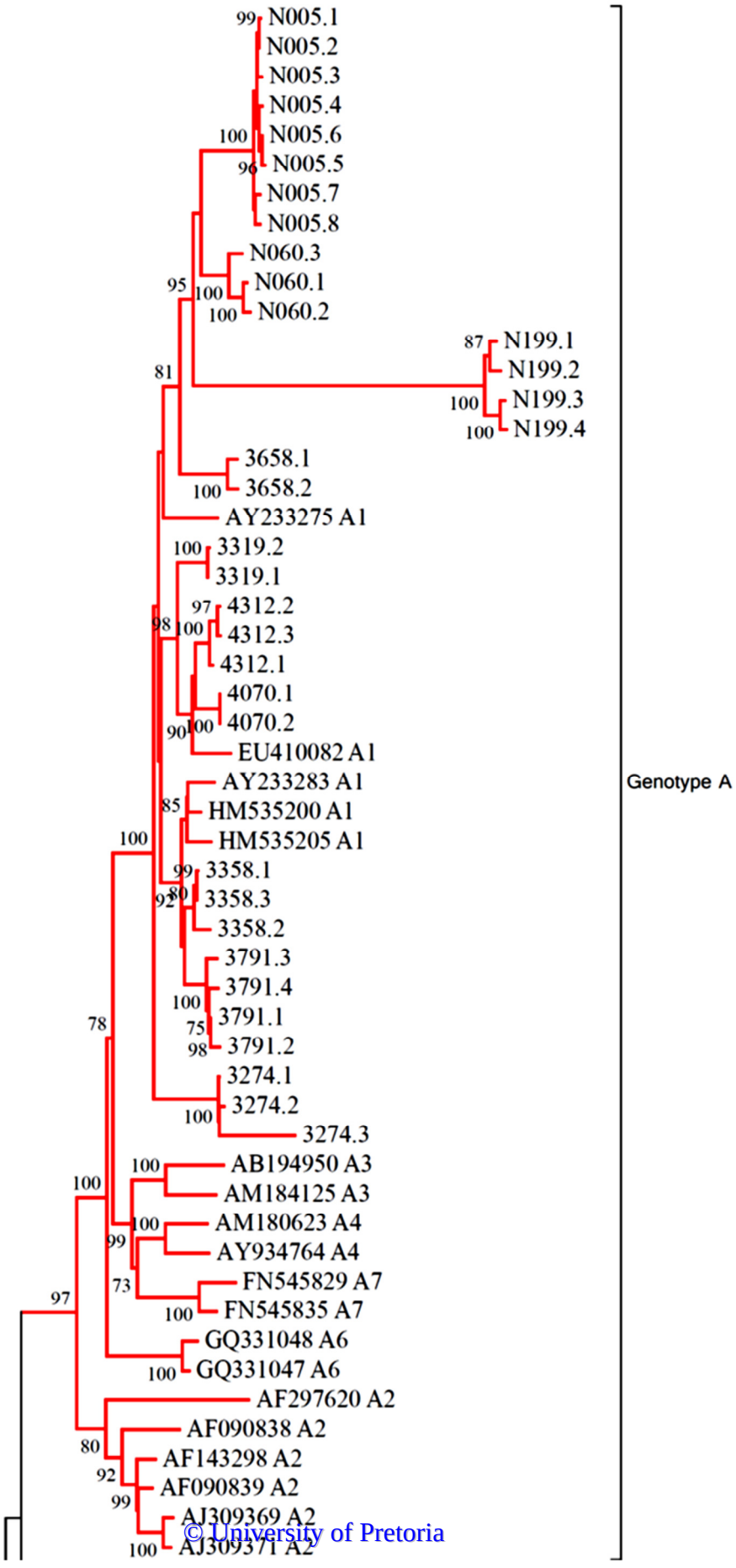
Accession number	Origin	Author	Year	Classification
AB116086	India	Sugauchi,F.	2006	A1
AB116088	Nepal	Sugauchi,F.	2003	A1
AB116091	Philippines	Sugauchi,F.	2003	A1
AB116094	Philippines	Sugauchi,F.	2003	A1
AB194950	Cameroon	Kurbanov,F.	2004	A3
AB241115	Philippines	Sakamoto,T.	2006	A1
AF090838	Belgium	Stuyver,L.	2000	A2
AF090839	Belgium	Stuyver,L.	2000	A2
AF143298	Germany	Preikschat,P	1999	A2
AF297620	South Africa	Owiredu,W.K.	2001	A2R/D
AF297621	South Africa	Owiredu,W.K.	2001	A1
AF297622	South Africa	Owiredu,W.K.	2001	A2R/C
AJ309369	France	Kay,A.C.	2001	A2
AJ309371	France	Kay,A.C.	2001	A2
AM180623	Mali	Olinger,C.M.	2006	A4
AM184125	Gabon	Roques,P.	2006	A3
AY090458	Costa Rica	Arauz-Ruiz,P.	2002	F
AY233275	South Africa	Kimbi,G.C.	2004	A1
AY233276	South Africa	Kimbi,G.C.	2004	A1
AY233277	South Africa	Kimbi,G.C.	2004	A1
AY233278	South Africa	Kimbi,G.C.	2004	A1
AY233281	South Africa	Kimbi,G.C.	2004	A1
AY233283	South Africa	Kimbi,G.C.	2004	A1
AY233284	South Africa	Kimbi,G.C.	2004	A1
AY233285	South Africa	Kimbi,G.C.	2004	A1
AY233287	South Africa	Kimbi,G.C.	2004	A1
AY233289	South Africa	Kimbi,G.C.	2004	A1
AY233290	South Africa	Kimbi,G.C.	2004	A1
AY934764	Gambia	Hannoun,C.	2005	A4
AY934766	Somalia	Hannoun,C.	2005	A1
AY934769	Somalia	Hannoun,C.	2005	A1
AY934774	Philippines	Hannoun,C.	2005	A1
DQ020003	United Arab Emirates	Hannoun,C.	2005	A1
EU410082	Philippines	Cavinta,L.	2009	A1
FJ692557	Haiti	Andernach,I.E.	2009	A1
FJ692557	Haiti	Andernach,I.E.	2009	A1
FJ692560	Haiti	Andernach,I.E.	2009	A1
FJ692566	Haiti	Andernach,I.E.	2009	A1
FJ692583	Haiti	Andernach,I.E.	2009	A1
FJ692585	Haiti	Andernach,I.E.	2009	A1
FJ692590	Haiti	Andernach,I.E.	2009	A1
FM199979	Rwanda	Hubschen,J.M.	2009	A1

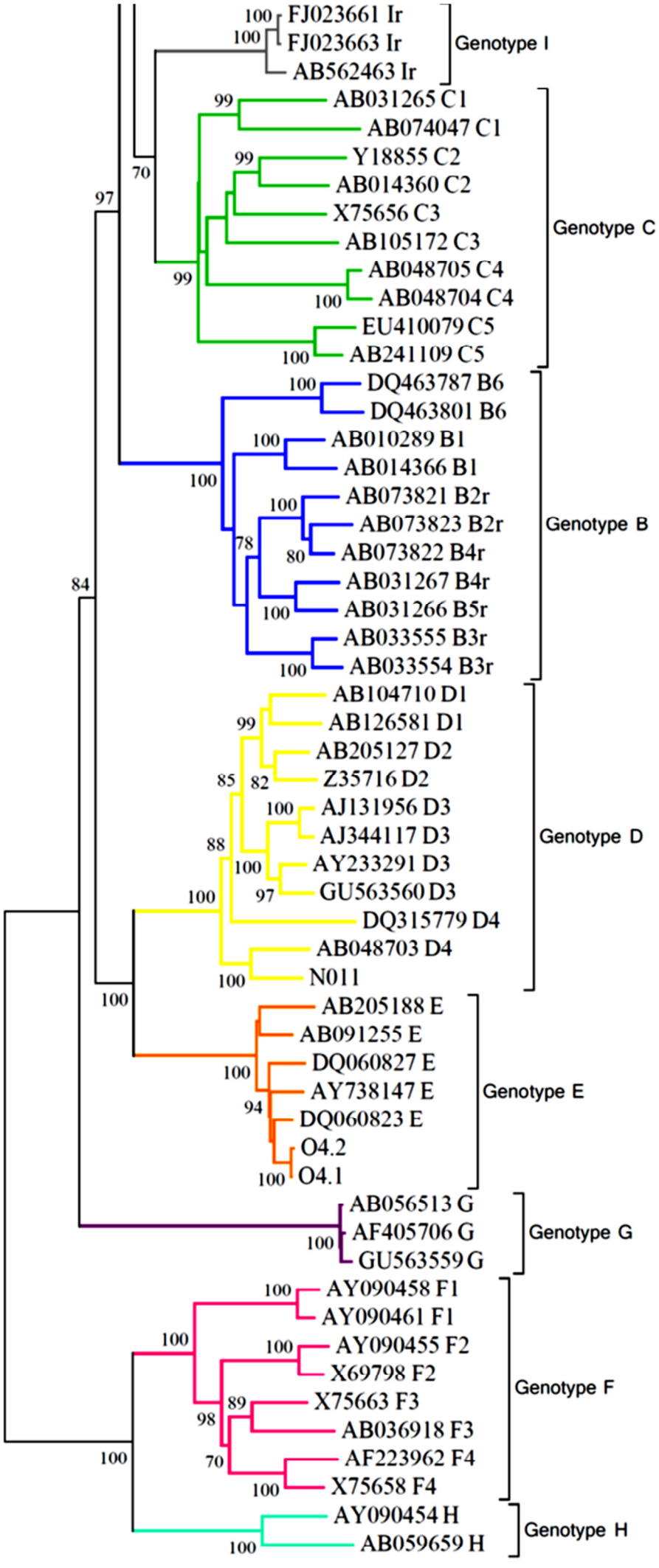
Accession number	Origin	Author	Year	Classification
FN545829	Cameroon	Hubschen,J.M.	2011	A7
FN545835	Cameroon	Hubschen,J.M.	2011	A7
GQ331047	Belgium	Pourkarim,M.R.	2009	A6
GQ331048	Belgium	Pourkarim,M.R.	2009	A6
GU563545	Belgium	Pourkarim,M.R.	2011	A1
HE974362	Martinique	Brichler,S.	2012	A1
HE974363	Martinique	Brichler,S.	2012	A1
HM535200	Zimbabwe	Gulube,Z.	2011	A1
HM535205	Zimbabwe	Gulube,Z.	2011	A1
JN315779	Korea	Bar-Gal,G.K.	2012	Ancient C
JX154581	Kenya	Kiyaba,R.M.	2013	A1
JX154582	Kenya	Kiyaba,R.M.	2013	A1
U87742	South Africa	Bowyer,S.M.	2002	A1
V00866	Japan	Ono,Y.	1983	A1

The references used in both phylogenetic analyses, recombination analyses as well as assessing positional variations at the DNA and protein level are listed in the table along with the country of origin and original authors.

APPENDIX D

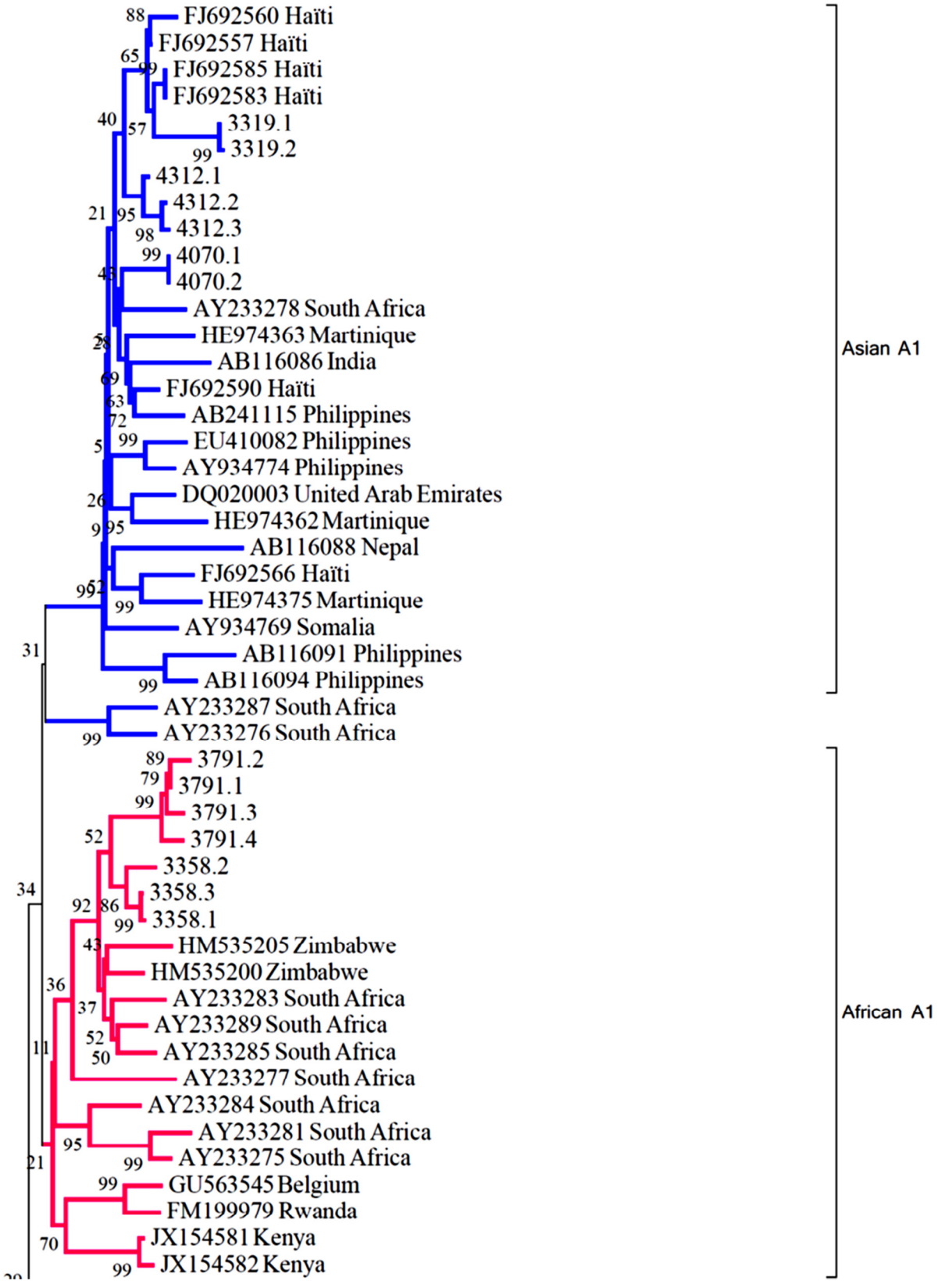
PHYLOGENETIC TREE FOR HBV/A TO I

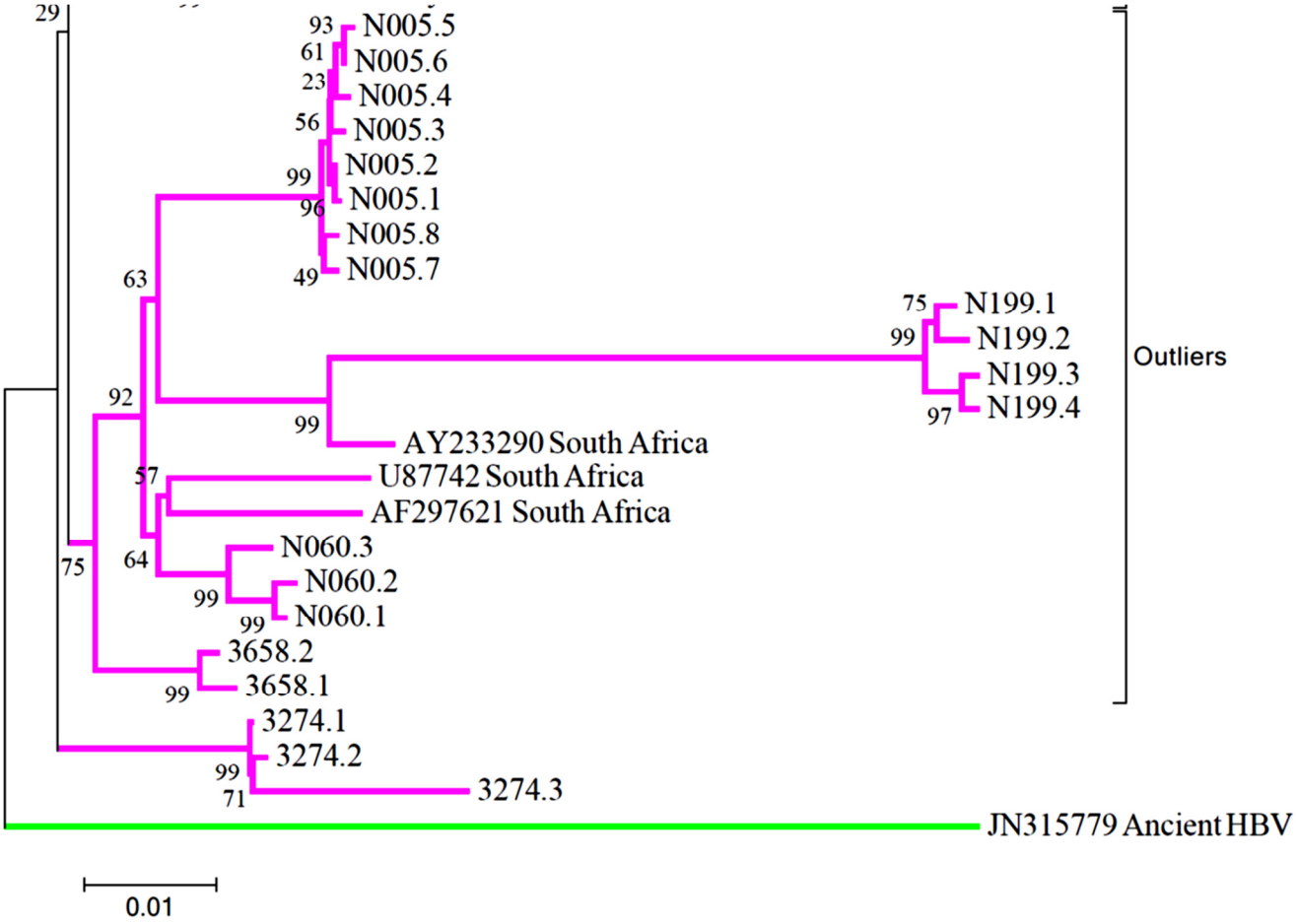




APPENDIX E

PHYLOGENETIC TREE FOR HBV/A1

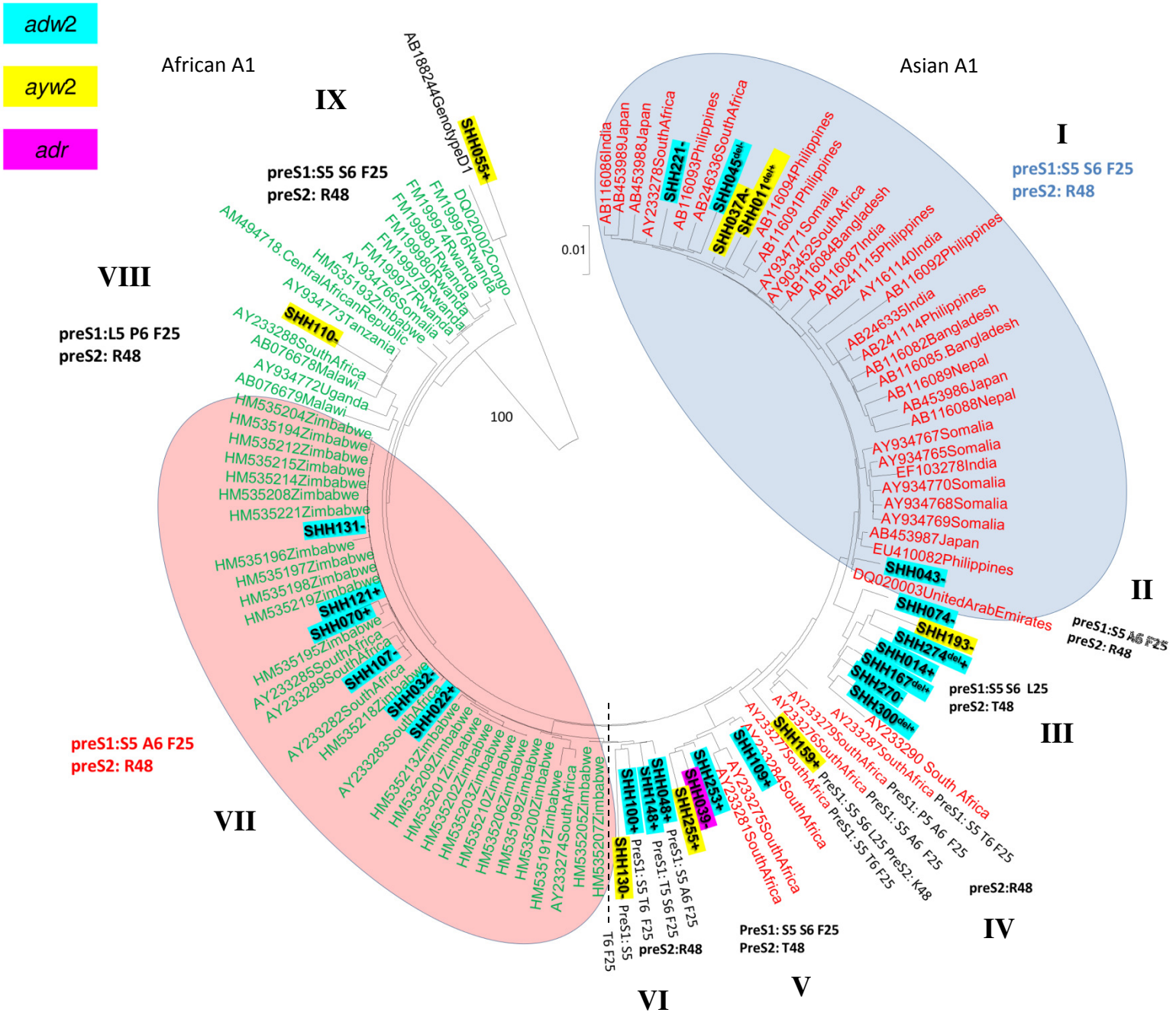




Phylogenetic trees were constructed for both HBV genotypes A to I (Appendix D) as well as for A1 (Appendix E) alone with the Neighbour-Joining method in MEGA 5 with 1000 bootstrap repeats.

APPENDIX F

ADAPTED FIGURE 1 FROM Makondo et al. (2012)



Phylogenetic relationship of complete pre-S1/pre-S2/S sequences (nt 2854–835 from the EcoRI site, numbering according to GenBank accession #AY233274) of 29 HBV isolates from HIV infected particpants [isolate number in bold, +: HBsAg+ve, −:HBsAg−ve, del:

deletion mutant] to sequences of other African (green) and “Asian” (red) subgenotype A1 HBV isolates obtained from GenBank established using neighbour-joining (Makondo et al. 2012). Each individual group that partitioned separately and shared common changes in Pre-S1/Pre-S2 are numbered from left to right in a clock-wise manner (I-IX) where, I to VI represent Asian A1 sequences and VII to IX represent African A1 sequences.

APPENDIX G

TABLE OF SAMPLE SPECIFIC VARIATION

3791 171 variable sites

DNA change	Protein change	Region	Region	Protein change
C81T	**T150I**	SURFACE	POLYMERASE	
T102C	I157T			
T192C	L187*(13)*P			
A201G	Q190*(16)*R			
A221G	T197*(23)*A			
A286C*				**I53L**
T454C*				**S109P**
A493T*				**N122Y**
C732T*	**S367*(192)*L**			
C1165T				
A1467G*	**R32G**	X-GENE		
A1484C				
A1508T				
T1512A*	S47T			
T1544A				
T1574A				
T1631C				
G1635A				
C1637A				
C1638T				
A1727G				
T1740C				
T1809G	**S146A**			
T1812C	**S147P**			
A1850T*	**T18S**	CORE		
C1858T				
A2136T				
G2495A*			POL	
T2852G				
T2869G*	**S6A**			
C3133A*	P94T			

Core gene amino acids

	FROM	TO
PRE-C (HBeAg)	1	214
C (HBc)	30	214

Surface gene amino acids:

	FROM	TO
PRE-S1	1	401
PRE-S2	120	401
S (HBsAg)	175	401

Stop codon

N199 365 variable sites

DNA change	Protein change	Region		Protein change
A7C		SURFACE	POLYMERASE	
T55C				
C81T*	**T150I**			
G134C*	**R167T**			
C147T*	**A172V**			
A159G*	E176*(2)*G			
A161T*	N177*(3)*Y			E11V
C166T*				H13C
A167G*	T179*(5)*A			
A169C*				
T170A*	S180*(6)*T			
A179C*				I14H
T180A*				
G194A*	V188*(14)*I			R22H
G196A*				V23I
A199G*				T24V
C200T*	Q190stop			
G225A/T*	R198*(24)*H/K			
T228A*	I199*(25)*N			N33K
C229T*				P34S
T231A*	L200*(26)*Q			
C232A*				H35N
A236C*	I202*(28)*L			N36T
A238G*				T37V
C239T*	P203*(29)*F			
C240T*				
G241T*				A38L
C242T*	Q204stop			
A245G*	S205*(31)*N/D			
G246A*				
C255T*	S208*(34)*L			
T279C*	L216*(42)*P			
A286C*				**I53P**
T287C*	S219*(45)*P			
C343T*				L72F
G351A*	P240*(66)*Q			
T352A*				S75I
C353T*	P241*(67)*S			
A355G*				
A356G*	I242*(68)*A			N76G
T357C*				

T358C*				
G360A*	C243*(69)*Y			
T435G*	L268stop			
T442A*				**S105T**
G458A*	G276*(102)*S			R110K
T465C*	L278*(104)*S			
G466A*				A113T
G587A*	G319*(145)*E/K			
G588A*				W153stop
A589G*				K154E
C595T*				H156C
A596G*	T322*(148)*A			
T614C*	S328*(154)*P			I162T
T702C*	F357(*183*)S			
C703T*				R192F
G704T*	V358*(184)*F			
A706T*				R193W
C714G*	S361*(187)*C			F195L
C717G*	P362*(188)*R			
G765A*	**S378*(204)*N**			
T777C*	I382*(208)*T			
C787A*				L220I
T792C*	I387*(213)*T			
C839A*				P237Q
T840G*				
G854A*				R242K
G858T*				W243C
C876T*				F249S
G925A*				
C969G*				
T1042C*				
A1043G/C*				
C1044T/A*				
C1045T*				
C1046T/G*				
T1422C*	C17P	X-GENE		
G1423C*				
G1429A*	R19H			
T1430C*				
C1432T*	P20L			
G1434A*	V21K			
T1435A*				
C1436A*				

C1439T*	
C1441T*	A23V
T1442C*	
G1443A*	E24R
A1444G*	
T1446A*	S25T
C1448T*	
C1449G*	R26A
G1450C*	
C1451A*	
G1452T*	G27stop
G1453A*	
C1455G*	R28G
C1458T*	P29Y
C1459A*	
C1460T*	
C1461G*	L30A
T1462C*	
C1463T*	
G1466T*	
G1469C*	P33stop
C1470T*	
C1471A*	
G1472A*	
C1473A*	L34N
T1474A*	
G1476T*	G35C
G1478C*	
C1480G*	T36S
C1482G*	L37A
T1483C*	
C1527A*	H52N
A1612C*	**E80A**
C1665T*	L98F
A1727G*	
T1753C*	I127T
A1762T*	**K130M**
G1764A*	**V131I**
T1800C*	C143Y/L
C1802T*	
T1803C*	
G1804A/T*	

C1805G*				
G1806C/T*	A144P/L/V			
C1807T*				
A1808T*				
T1809G*	P145A			
T1812G	**S146A**			
T1815G	**S147P**			
A1841C*	I10L	CORE		
C1843T*				
T1845T*	S11F			
T1847C*	C12H			
G1848A*				
A1859T*	T16S			
T1863C*	**V17A**			
T1864C*				
C1865T*	Q18stop			
G1866T*				
C1946T*	L45F			
C1948T*				
T1952C*	F47P			
T1953C*				
T1955C*				
T1961A*	S50N			
C1962A*				
G1964C*	D51P			
A1965C*				
C1966T*				
C1969T*				
T1971C*	F53S			
T1972A*				
C1981A*				
A1990G*				
T1992C*	L60P			
G1994A*	D61N			
A1997C*	T62P			
A1999T*				
G2000T*	A63F			
C2001T*				
C2002T*				
G2006C*	A65P			
A2092T*	**E93D**			
A2095T*	**L94V**			

C2100A*	**T96N**			
A2108T*	T99S			
A2131C*	E106D			
T2151A*	**L113*(84)*Q**			
C2191T*				
G2237C*	E142Q			
T2434C*				
C2435T*	Q208stop			
T2440C*				
C2503T*				
T2507A*				
A2508C*				
G2511C*				
C2513G*				
C2519G*				
T2926C*	**F25L**			
G3115T*	**V88L**			
G3121A*	**A90T**			

N005

257 variable sites

DNA change	Protein change	Region		Protein change
T12G*	F127C			
C16T				
C26A*	Q132K			
T123C*	I164T			
G134C*	**R167T**			
C147T*	**A172V**			
G381A/T*	C250*(76)*F/Y			
T442A*				**S105T**
T491A*	S287*(113)*T			I121N
T705C*	**V358*(184)*A**			
T770C*	Y380*(206)*H			V214A
A849T				
C873T				
A906G				
C1020T				
G1249T*				
T1386G*	L5V			
G1413A*	D14N			
T1425C				
G1437A*	**G22S**			
C1465A*	A31E			

A1612C*	**E80A**			
A1727G				
T1741C*	L123S			
T1754G*	I127M			
A1762T*	**K130M**			
G1764A*	**V131I**			
T1812G	**S146A**			
A1908G*		CORE		
A/G1951T				
C1981A				
C1988T*				
C2023A				
A2092T*	**E93D**			
A2095T*	**L94V**			
C2100A*	**T96N**			
C2710T*				
A2121C*	N103T			
T2134C*				
T2151A*	**L113*(84)*Q**			
C2158T*				
C2172T*	T120*(91)*I			
C2183A*	L124(95)I			
C2191T				
C2222A*	L137I			
G2257A				
C2260T*				
A2297G*	R162G			
A2324T*	T171S			
C2354T*	R181stop			
A2358G*	D182G			
G2364A*	**G184D**			
A2375G*	R188G			
T2518A			POLYMERASE	
C2519T				
A2569G*				
T2614C*				
T2678C				
A2687G*				
T2717C*				
A2745C*				
A2871C*		S		

T2926C*	**F25L**			
T3076C*				
G3115T*	**V88L**			
G3124A*	V91I			
A3160C*				
G3213A*				

3319 72 variable sites

DNA change	Protein change	Region		Protein change
A286G*		SURFACE	POLYMERASE	**I53V**
T454C*				**S109P**
C/A493T*				**N122Y**
C732T*	**S367*(193)*L**			
T779C*				
G852A				
T873C				
A951G				
T975C				
G1437A*	**G22S**	X - GENE		
C1461T*	**L30F**			
C1470T*	**P33S**			
C1810T*	S146L			
T1844A*	S11T	CORE		
G1848T*	C12F			
A1850T*				
A2145T*				
T2957C		S	P	
A2995G*	**I48V**			
G3032A*				
A3128G*				

4070 169 variable sites

DNA change	Protein change	Region		Protein change
C117T*	S162L	SURFACE	POLYMERASE	
C150T*	**L173P**			
C346T*				
A493C*				N122H
T735C*	**V368*(194)*A**			
T1055A*				
C1171A*				
T1386G*	**L5V**	X		
T1425C				

C1470T*	**P33S**			
A1617T*	**T82S**			
G1862T*	**V17F**	C		
T2149C*				
A2155T*				
G2777A				
A2995G*	**I48V**	S	P	
C2810T				

4312 63 variable sites

DNA change	Protein change	Region		Protein change
G8A*	A126T	SURFACE	POLYMERASE	
C117T				
C150T*	**L173P**			
A286G*				**I53V**
C287A				
A493C*				N122H
T735C*	**V368*(194)*A**			
A882G				
G925A				
A993G				
T1092C				
T1218C				
A1320C				
A1368C				
G1437A*	**G22S**	X-GENE		
C1461T*	**L30F**			
C1470T*	**P33S**			
G1479A*				
G1862T*	**V17F**	C		
T2119C*				
A2200G				
A2460C*			POLYMERASE	
T2488G*				
T2516C*				
T2543C*				
A2654G				
T2613C				
G2672T/C				
T/A2720C*				
A2995G*	**I48V**	S		
G3000A*				

3274 265 variable sites

DNA change	Protein change	Region		Protein change
C13T*		SURFACE	POLYMERASE	
C96A*	P155Q			
A97G*				
C105T*	A158V			
G148A*				
T192C*	L187*(13)*P			
G241A*				
T259C				
C290A*	P220*(46)*T			T54N
T344C*	S238*(64)*P			
G348T*	C239*(65)*F			
C353T*	P241*(67)*S			
A356G*	I242*(68)*V			
A357G				
T358C*				
T359C*	C243*(69)*R			
C373T*				
G379C*	M249*(75)*I			
G381A*	**C250*(76)*Y**			
T382C*				
C383T*				
C386A*	R252*(78)*I			
G387T*				
G388A*				
T390A*	R253*(79)*H			
C427T*				
T429G*	I266*(92)*S			
T432C*	F267*(93)*S			
C433T*				
T434G*	L268*(94)*V			
T438C*	L269*(95)*S			
G449A*	D273*(99)*K			
T451G*				
T452C*	Y274*(100)*P			
A453C*				
C455T*	Q275stop			
A457G*				
G458A*	G276*(102)*S			
A461C*	M277*(103)*L			
G463C*				

T464C*	L278*(104)*Q			
T465A*				
C467A*	P279*(105)*S			
C468G*				
G470A*	V280*(106)*I			
T473G*	C281*(107)*A			
G474C*				
T478C*				
A481C*				
T483C*	I284*(110)*T			
T484C*				
C485T*	P285*(111)*S			
A493C*				N122H
T592C*				
T684C*	V351*(177)*A			
T777C*	**I382*(208)*T**			
A834G*				
A895G*				
A912G*				
A987C*				
A1083G*				
A1104C*				
C1221A*				
C1258T*				
A1368C*				
A/G1467C		X-GENE		
C1470T*	**P33S**			
C1649A*				
G1658A*				
C1703A*				
T1753C*	I127T			
A1762T*	**K130M**			
G1764A*	**V131I**			
A1934T*	T41S	CORE		
C2004T*	S64L			
T2035G/A*				
C2063A*	L84I			
C2100T*	T96I			
C2102T*	L97F			
A2104T*				
A2131G*				
T2167C				

T/C2191A*				
C2293T*				
A2302C*				
C2304A*	P164Q			
A2326T*				
A2335G*				
A2358G*	**D182G**			
T2498G*				
A2555G*				
G2585A*				
C2609T*				
G2668A*				
T2684A*				
G2792A*				
T2869G*	**S6A**	SURFACE	POLYMERASE	
T2916C*				
A2922G*				
T3104C*	I84T			
C3163A*	Q104K			

3658 241 variable sites

DNA change	Protein change	Region		Protein change
T53C*	F141R	SURFACE	POLYMERASE	
T54G*				
T84A*	L151H			
A97C*				
G132A*	R167K			
A286C*				**I53L**
T454C*				**S109P**
A493C*				N122H
C502G*				Q125E
A541G*				R138E
G542A*	G304*(130)*N			
G543A*				
T705C*	**V358*(184)*A**			
C717T*	P362*(188)*L			
T729C*				
C732T*				
G765A*	**S378*(204)*N**			
T777C*	**I382*(208)*T**			
A972T*				

G1062A				
A1317G				
A1479G		X-GENE		
T1544A				
T1574A				
A1612C*	**E80A**			
G1613A				
A1635G*	I88V			
C1638T				
C1653T*	H94Y			
A1762T*	**K130M**			
C1766T*				
T1768A*	F132Y			
T1815G*	**S147P**			
A1850T*	**T13S**	CORE		
C1858T				
C2002T*				
C2078T				
T2167C				
T2278A				
C2519T			POLYMERASE	
T2684C				
C2685T				
T2852C				
T2869G*	**S6A**	SURFACE		
T2926C*	**F25L**			
C3021T				
T3111C				
G3115T*	**V88L**			

N060 194 variable sites

DNA change	Protein change	Region		Protein change
G134C*	**R167T**	S	POLYMERASE	
C147T*	**A172V**			
T442A*				**S105T**
A457G*				R110G
T735C*	**V368*(194)*A**			
A849T				
T903G*				
G915A*				
C940T				
C967A				

T975C				
G1080A				
G1122A				
A1234C*				
A1368G				
T1425C*	R26C	X-GENE		
C1449T				
A1467G*	**R32G**			
A1612C*	**E80A**			
T1636G*	I88S			
T1740G				
T1809G*				
T1815G*	**S147P**			
A1850T*	**T13S**	CORE		
C1858T				
G1896A*	W28stop			
T1909C*				
G1931T*	A40S			
C1978A				
C1981A				
T1993G*				
G2011A*				
G2017A*				
C2022G*	A70G			
G2032A				
C2034A/T*	P74Q/L			
T2035G				
A2047C				
A2059G*				
A2075G*	I88V			
G2129C*	E106Q			
A2137T				
C2191T*				
C2245T				
G2257A				
C2266T				
T2278A				
C2325T*	T171I		POL	
A2326C*				
T2447C*	S212P			
C2519T				
T2552A				

C2573T				
G2613A*				
A2616C*				
G2629A*				
T2648C				
A2741G				
T2831C				
A2851C*				
C2910A		SURFACE		
T2926C*	**F25L**			
G3115T*	**V88L**			
G3121A*	**A90T**			
T3154A*	S101T			

The tables indicate the observed changes at nucleotide level, as compared to the references listed in appendix C and used in phylogenetic analyses, detected in MEGA which were either unique (*) or sparse with common variation highlighted in bold. The associated unique changes at the amino acid level are also indicated along with the respective genes; the numbering scheme used for amino acids is also indicated. The bright coral red blocks indicate stop codon mutations; text highlighted in red indicates change in the primer regions