We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 8

The Study of Hepatitis B Virus Using Bioinformatics

Trevor Graham Bell and Anna Kramvis

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63076

Abstract

Hepatitis refers to the inflammation of the liver. A major cause of hepatitis is the hepatotropic virus, hepatitis B virus (HBV). Annually, more than 786,000 people die as a result of the clinical manifestations of HBV infection, which include cirrhosis and hepatocellular carcinoma. Sequence heterogeneity is a feature of HBV, because the viral-encoded polymerase lacks proof-reading ability. HBV has been classified into nine genotypes, A to I, with a putative 10th genotype, "J," isolated from a single individual. Comparative analysis of HBV strains from various geographic regions of the world and from different eras can shed light on the origin, evolution, transmission and response to anti-HBV preventative, and treatment measures. Bioinformatics tools and databases have been used to better understand HBV mutations and how they develop, especially in response to antiviral therapy and vaccination. Despite its small genome size of ~3.2 kb, HBV presents several bioinformatic challenges, which include the circular genome, the overlapping open reading frames, and the different genome lengths of the genotypes. Thus, bioinformatics tools and databases have been developed to facilitate the study of HBV.

Keywords: alignments, computation, databases, genotypes, phylogenetics

1. Introduction

Primarily, bioinformatics is the use of computational science to study biological and clinical data using statistics, mathematics, and information theory. This field is developing and evolving; thus, the definition cannot be precise. Moreover, the field is broad, ranging from the study of DNA and proteins, to structural biology, drug design and comparative genomics, transcriptomics, proteomics, and metagenomics. The optimization of computational technology is paramount in order to handle, store, manage, and analyze the large volumes of data generat-

INTECH
open science | open minds

© 2016 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

ed in the last decade. The data include molecular sequencing data of host and pathogen genomes and their associations to demographic and clinical records, laboratory test results, as well as information on treatment. Moreover, bioinformatics can aid in the investigation of virus–host genome and environmental interactions and in the identification of both host and viral biomarkers. This analysis can lead to a better understanding of clinical manifestation of disease and effective design of preventative and treatment measures [1].

In the first section, we describe the unique genomics and molecular biology of hepatitis B virus (HBV). Using illustrative examples, we showed how bioinformatics analyses can facilitate the understanding of the origin, evolution, transmission, and response to antiviral agents of HBV. Next, we described the bioinformatics challenges posed by HBV and present the public databases and tools currently available for the study of HBV.

2. Hepatitis B virus

2.1. Hepatitis

Hepatitis refers to the inflammation of the liver. A major cause of hepatitis is the hepatotropic virus, HBV. HBV infection is a public health problem of worldwide importance. Globally, 2 billion people have been exposed to this virus at some stage of their lives, and 240 million are chronic carriers of the virus [2].

This infection can lead to a spectrum of clinical consequences. In the majority of cases, the infection is subclinical and transient, whereas in 25% of cases, it can cause self-limited acute hepatitis and in 1% of these progress to acute liver failure. The virus can persist in 90% of neonates and 5–10% of adults, leading to chronic infection that can progress to either chronic hepatitis or an asymptomatic carrier state. Both of these states can ultimately develop liver cancer or hepatocellular carcinoma (HCC), with or without the intermediate cirrhotic stage. Annually, more than 786,000 people die as a result of these clinical manifestations of HBV infection [3].

2.2. Prevalence

The prevalence of HBV in a community can be estimated by the proportion of the population, who are hepatitis B surface antigen (HBsAg)-positive carriers. HBV prevalence varies widely in the world [3]. The prevalence is low (<1%) in northern Europe, Australia, New Zealand, Canada, and the United States of America. Northern Asia, the Indian subcontinent, parts of Africa, Eastern and south-eastern Europe, and parts of Latin America are areas of intermediate prevalence (1–5%). The high prevalence areas (5–20%) include East and Southeast Asia, the Pacific Islands, and sub-Saharan Africa.

2.3. Classification and structure

HBV, the prototype member of the family *Hepadnaviridae*, belongs to the genus *Orthohepadnavirus*. With a diameter of 42 nm and a DNA genome of ~3.2 kilobases (kb), it is the smallest

http://dx.doi.org/10.5772/63076

DNA virus infecting man. The genome is circular and partially double stranded. One DNA strand is complete, except for a small nick (the minus strand), and the other is short and incomplete (the plus strand). The minus strand contains four overlapping open reading frames (ORFs; **Figure 1**) [4] that represent: (1) the *preS*/*S* gene that codes for the envelope proteins, large, middle, and small HBsAgs; (2) the *P* gene for DNA polymerase/reverse transcriptase (POL); (3) the *X* gene for the X protein, a key regulator during the natural infection process, which has transcriptional trans-activation activity and is required to initiate and maintain HBV replication [5]; and (4) the *precore*/*core* gene that codes for the HBcAg or core protein that forms the capsid and for an additional protein known as HBeAg, which is not incorporated into the virus itself but is expressed on the liver cells and secreted into the serum. **Figure 2** illustrates the structure of the hepatitis B virion.

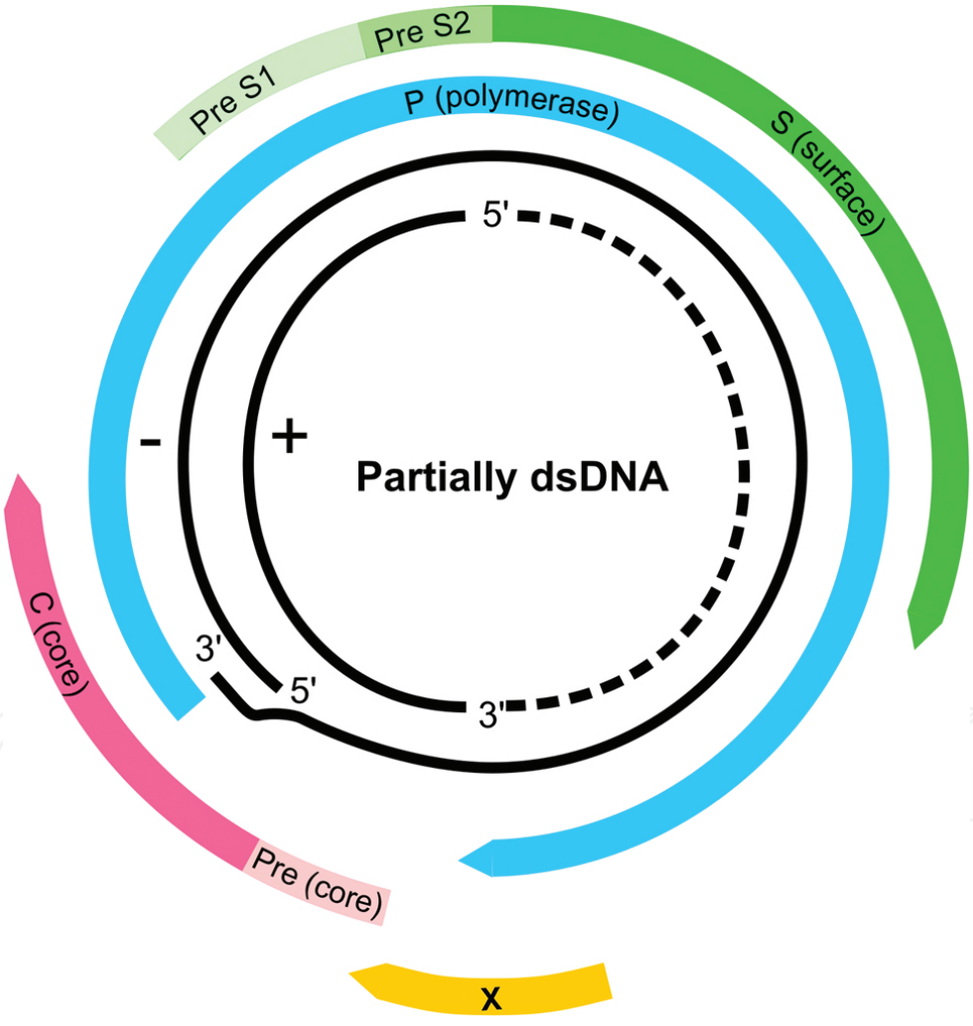

Figure 1. The genome of hepatitis B virus (HBV). The partially double-stranded DNA (dsDNA) with the complete minus (−) strand and the incomplete (+) strand. The four open reading frames (ORFs) are shown: *precore/core (preC/C)* that encodes the e antigen (HBeAg) and core protein (HBcAg); *P* for polymerase (reverse transcriptase), *PreS1*/*PreS2*/*S* for surface proteins [three forms of HBsAg, small (S), middle (M), and large (L)] and *X* for a transcriptional trans-activator protein.

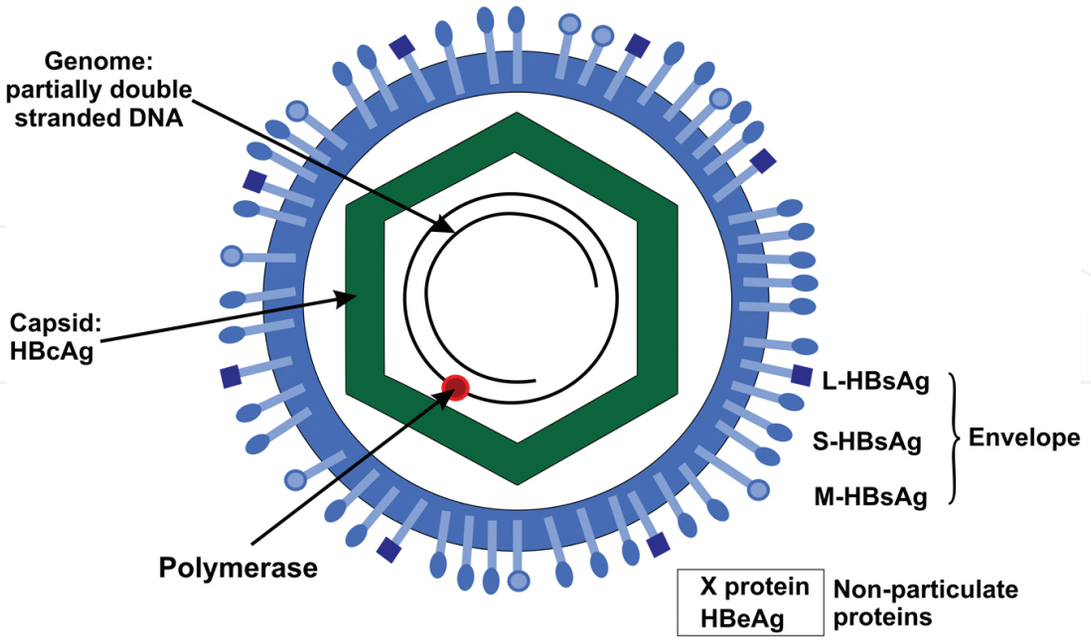

Figure 2. Schematic representation of hepatitis B virus (HBV), showing the structure of the virion, composed of a partially double-stranded DNA genome, enclosed by a capsid, comprised of HBcAg and surrounded by a lipid envelope containing large (L)-HBsAg, middle (M)-HBsAg, and small (S)-HBsAg. The virus also expresses two non-particulate proteins X protein and HBeAg.

2.4. Regulatory elements of HBV

Every single nucleotide of the HBV genome is necessary for the translation of a protein and may also be part of one of the regulatory elements of HBV, which overlap with protein expressing regions. The regulatory elements include the S1 and S2 promoters, which overlap both the preS region and polymerase ORFs; the preC/pregenomic promoter, which includes the basic core promoter (BCP) and overlaps the X and preC ORF; and the X promoter. There are two enhancers (enhancer I and enhancer II) as well as *cis*-acting negative regulatory elements (URR: upper regulatory region, CURS: core upstream regulatory sequence, NRE: negative regulatory element). These regulatory elements control transcription (reviewed in [6, 7]).

2.5. Replication of HBV

HBV and other members of the family *Hepadnaviridae* have an unusual replication cycle. These DNA viruses replicate by reverse transcription of a RNA intermediate known as the pregenomic RNA (pgRNA) [8]. Entry into the cell is via the sodium taurocholate cotransporting polypeptide (NTCP), a multiple transmembrane transporter predominantly expressed in the liver [9]. After entry, the virion is uncoated and the core particle is actively transported to the nucleus [10], where the partially double strand relaxed circular DNA molecule is released. The single-stranded gap is closed by the viral polymerase to yield a covalently closed circular molecule of DNA (cccDNA) [11], which is the template for transcription by the host RNA polymerase II [12]. The mRNAs are transported into the cytoplasm where they are translated

into the seven viral proteins. In addition to being translated into the polymerase and the core protein, the pgRNA is packaged into immature core particles by the process known as encapsidation. In order to be encapsidated, the 5′ end of the pgRNA has to be folded into a particular secondary structure known as the encapsidation signal (ε) [13].

The encapsidation signal (ε) is a bipartite stem-loop structure, consisting of an upper and lower stem, the bulge, and an apical loop. Besides encapsidation, ε has a number of other functions (reviewed in [13]) and references therein. It acts in template restriction so that not any piece of RNA is encapsidated, and it also plays a role in the activation of the viral polymerase, so that there is no indiscriminate reverse transcription. It is also involved in the initiation of reverse transcription. The polymerase or reverse transcriptase acts as a primer of RNA-directed DNA synthesis by the binding of the polymerase to the bulge of ε. The first three nucleotides of the negative stand of DNA are synthesized at the bulge and are transferred to an acceptor site on the 3′ end of the pgRNA, where DNA synthesis proceeds toward the 5′ end of the pgRNA [14], giving rise to the immature virion. The virus matures by acquiring its glycoprotein envelope, containing HBsAg, in the endoplasmic reticulum and is exported by vesicular transport from the cell [15].

2.6. Genotypes and subgenotypes of HBV

Sequence heterogeneity is a feature of HBV, because the viral-encoded polymerase lacks proof-reading ability as mentioned above [16]. Using phylogenetic analysis of the complete genome of HBV and an intergroup divergence of greater than 7.5%, HBV has been classified into nine genotypes, A to I [17, 18, 19], with a putative 10th genotype, "J," isolated from a single individual [20]. With between ~4 and ~8% intergroup nucleotide difference across the complete genome and good bootstrap support, genotypes A–D, F, H, and I are classified further into at least 35 subgenotypes [21]. The genotypes differ in genome length, the size of ORFs and the proteins translated [17], as well as the development of various mutations [22]. Generally, the genotypes, and in some cases the subgenotypes, have a distinct geographic distribution (**Table 1**).

Genotype	Length	Differentiating features	Subgenotypes	Geographic distribution	Serological subtype	Transmission route
A	3221	6-nucleotide insert at carboxyl end of core gene	A1	Africa[#]	*adw*2/*ayw*2	Horizontal: parenteral or sexual
			A2	Europe/North America	*adw*2	
			Quasi-subgenotype A3[(A3,A4,A5)§]	Africa, Haiti	*ayw*1	
			A4[(A6)§]	Africa	*ayw*1 *adw*4*	

Genotype	Length	Differentiating features	Subgenotypes	Geographic distribution	Serological subtype	Transmission route
B	3215		B1	Japan	*adw*2	
			B2	China	*adw*2	
			Quasi-subgenotype B3[(B3,B5,B7–B9,B6(China)§]	Indonesia Philippines/China	*adw*2	Perinatal
			B4	Vietnam/Cambodia/ France	*ayw*1/*adw*2	
			B5[(B6)§]	Eskimos/Inuits	*adw*2	
					*adr*fntab1_3	
C	3215		C1	Thailand/Myanmar/ Vietnam	*adr**	
			Quasi-subgenotype C2[(C2,C14, undefined sequences)§]	Japan/China/Korea	*adr*	
			C3	New Caledonia/ Polynesia	*adr*	Perinatal
			C4	Australian Aborigines	*ayw*2/*ayw*3	
			C5	Philippines/ Indonesia	*adw*2	
			C6–C12	Indonesia/ Philippines	*adr*	
			C13–C15	Indonesia	*adr*	
			C16	Indonesia	*ayr**	
D	3182	33-nucleotide deletion at the amino terminus of the preS1 region	D1	Middle East, Central Asia	*ayw*2	
			D2	Europe/Japan/Lebanon	*ayw*3	
			D3	Worldwide	*ayw*2/*ayw*3	
			D4	Australian aborigines, Micronesians, Papua New Guineans, Arctic Denes	*ayw*2	Horizontal: parenteral with intravenous drug use being a risk factor
			D5	India	*ayw*3/*ayw*2	

Genotype	Length	Differentiating features	Subgenotypes	Geographic distribution	Serological subtype	Transmission route
			D6	Tunisia/Nigeria	*ayw*2	
					*ayw*4*/*adw*3*	
E	3212	3-nucleotide deletion at the amino terminus of the preS1 region	–	Western/Central Africa	*ayw*4	Horizontal
F	3215		F1	Argentina/Costa Rica/El Salvador, Alaska	*adw*4	
			F2	Nicaragua/ Venezuela/Brazil	*adw*4	Horizontal
			F3	Venezuela/Colombia	*adw*4	
			F4	Argentina	*adw*4	
					*adw*2*/*ayw*4*	

[¶]Summarizes data compiled from Kramvis [21] and references cited therein.
[§]Earlier subgenotype designation.
*Rare serological subtype for that genotype.
[#]And in regions outside Africa where there was historical forced migration as a result of the slave trade [23].
[¥]Vietnamese residing in Canada [24].

Table 1. Comparison of the virological and clinical characteristics of the genotypes and subgenotypes of HBV[¶].

2.7. Genotyping and subgenotyping methods

HBV genotypes, and in some cases subgenotypes and various mutations, can influence the clinical course of disease [22] as well as response to antiviral therapy [25] and can be used to show transmission [26] and to trace human migrations [23]. Thus, HBV genotyping is becoming increasingly relevant in the clinical setting and may contribute to future personalized treatment [27] and may be important in epidemiological and transmission studies. Bioinformatics has played a major role in the development of various tools that can be used for identifying genotypes/subgenotypes and detecting various mutations. Therefore, a number of methods have been developed [28, 29].

Although analysis of the HBV *S* gene sequence is sufficient to classify HBV into genotypes [30], the complete genome sequence provides additional information with respect to phylogenetic relatedness [31, 32], including the identification of recombinants. Furthermore, even though complete genome analysis is the gold standard for genotyping, it does not allow for rapid and direct analysis on a large scale basis [17] and requires expertise and thus capacity development in computer processing coupled with phylogenetic analyses. In order to expedite and facilitate genotyping, a number of methods have been developed [17, 28, 29]. Each one has its advantages

and disadvantages [17, 28, 29], which should be taken into account, when selecting the genotyping method appropriate for a particular study or application.

2.8. Phylogenetic analyses of HBV

Although, as already mentioned, the error-prone polymerase of HBV leads to sequence heterogeneity [16], the degree, at which this can occur, is constrained by the partially overlapping ORFs and the presence of secondary RNA structures, such as ε, coded by non-overlapping regions [33, 34]. The HBV genome has been estimated to evolve with an error rate of ~10^{-3}–10^{-6} nucleotide substitutions/site/year [35–41], although this rate is not constant within the different regions of the HBV genome [41]. The progress of computers and information technology has played an important role in the development of phylogenetic analysis as a powerful tool in the analysis of the molecular evolution of viruses.

As exemplified in the next sections, comparative analysis of HBV strains from various geographic regions of the world and from different eras can shed light on the origin, evolution, transmission, and response to anti-HBV preventative and treatment measures.

2.9. Origin

The origin and age of the family *Hepadnaviridae* remains controversial. However, until the issues with the estimation of the substitution rate of HBV [41] are overcome, the debate on the origin of HBV will continue ([17, 41] and references cited therein). Nonetheless, bioinformatics, coupled with growing number of hepadnaviral sequences in the databases, with accurate sampling times, and advances in phylogenetic and coalescent methodology [42], is beginning to shed light on this issue. For example, according to Suh and colleagues [43], analysis of the endogenous sequences in the zebra finch provides direct evidence that the compact genomic organization of hepadnaviruses has not changed during the last 482 million years of hepadnaviral evolution. Furthermore, phylogenetic analyses and distribution of HBV relics suggest that birds potentially are the ancestral hosts of the family *Hepadnaviridae* and that mammalian hepatitis B viruses probably emerged after a bird–mammal host switch [43].

2.10. Evolution

Genetic variation is important in viral evolution. The sequence heterogeneity displayed by HBV because of the lack of proof-reading ability of the polymerase is limited by functional constraints [33], leading to non-random variation [44]. Moreover, mutations can be affected by host–virus interaction and selective pressure, imposed endogenously by the immune system and exogenously by vaccination and antiviral treatment [17]. Phenotypic resistance to antiviral drugs occurs because of mutations in the reverse transcriptase of POL, whereas mutations in the *BCP*/*preC* and *preS* regions have been implicated as risk factors for the development of HCC. Mutations in the *S* region coding for HBsAg can lead to both vaccine and detection escape of HBV. At any time, the virus population can be composed of a number of different mutants referred to as "quasispecies" [45]. Direct sequencing and more recently next generation sequencing (NGS), parallel with bioinformatics, provide us with powerful

tools to study the evolution of the various HBV mutations. NGS or ultra-deep sequencing generates large volumes of data, which can only be analyzed using bioinformatics tools and provides large coverage that can detect minor quasispecies populations of HBV [46–51] that may be important in understanding HBV pathogenicity and response to treatment. In order to minimize the number of artifactual calls of single-nucleotide variations in NGS, it is important that the correct reference sequences are used [51, 52].

By designing a circular construct, Homs and co-workers [53] were able to use NGS to study evolution of both the precore and polymerase regions. They demonstrated the presence of precore mutants in HBeAg-positive phase, wild-type precore in the HBeAg-negative phase as well as lamivudine resistance strains in treatment naïve patients. This demonstrates that viral strains occurring at low frequencies can act as reservoirs or memory genomes, which are selected and evolve in response to both intrinsic (host immune response) and extrinsic (drug administration) factors.

2.11. Transmission and tracing human migrations

Sequencing and bioinformatics have played an important role in demonstrating transmission routes, for which previous evidence could only be anecdotal. For example, molecular characterization of HBV together with phylogenetic analysis was used to demonstrate inter-spousal transmission of HBV even after long marriages, in two Japanese patients, who developed acute liver failure [54]. Similarly, the first known case of transfusion-transmitted HBV infection by blood screened using individual donor nucleic acid testing was confirmed by the 99.7% sequence homology between the complete genome sequences of the donor and the recipient HBV strains [26]. When migration events were estimated by ancestral state reconstruction using the criterion of parsimony, it was shown that Africa was the most probable source of dispersal of subgenotype A1 of HBV globally and its dispersal to Asia and Latin America occurred as a result of the slave and trade routes [23, 55].

2.12. Treatment response and resistance to treatment

According to international chronic hepatitis B treatment guidelines, the most desirable endpoint of treatment is HBsAg loss. Following HBsAg loss, patients have better clinical outcomes, including decreased risk of developing cirrhosis and HCC, and death [56]. However, the currently available treatments, which include either nucleos(t)ide analogues (NAs) for direct inhibition of the viral polymerase or pegylated interferon (PegIFN) for immune-mediated HBV control, generally achieve HBV DNA suppression and HBeAg loss only, which are not enduring. In an attempt to identify viral factors associated with HBsAg loss, Charuworn et al. [57] demonstrated that viral diversity could differentiate those patients, who would lose HBsAg when treated with tenofovir disoproxil fumarate. Lower diversity was seen in the protein-encoding regions of HBV from patients who lost HBsAg compared to those who did not. On the other hand, higher diversity in regulatory elements of HBV was found to be a predictor of HBsAg loss [57]. These findings need to be confirmed by studies incorporating larger numbers of patients, as well as genotypes other than A and D.

The high mutation rate of HBV means that it can evolve to develop resistance against NAs that target the viral DNA polymerase. Drug-resistant mutants develop under drug pressure in order for HBV to survive in the presence of the NA. The development of drug resistance mutations can be affected by HBV DNA levels at baseline, rate of viral suppression, length of NA treatment, and prior exposure to NA treatment [58]. Sequential treatment with different NAs, following drug failure, can lead to the development of multidrug resistance, which cannot be treated using currently available drugs [59]. The most frequent lamivudine drug resistance mutants are rtM204V/I, which are also selected by the L-pyrimidine analogues, emitricitabine, clevudine, and telbivudine but are susceptible to the purine analogues adefovir and tenofovir [59]. rtA181V develops following lamivudine treatment but is sensitive to other NAs, whereas rtN236T is resistant to adefovir only. In deciding on treatment options, the detection of genotypic resistance, which is defined as the detection of viral mutations conferring drug resistance, is a priority in clinics. Direct and NGS of the polymerase region of the HBV genome can detect both well-defined and novel mutations.

Bioinformatics tools and databases have been used to better understand HBV mutations and how they develop, especially in response to antiviral therapy and vaccination. Although laboratory methods have been used to study mutations, they are both labor intensive and expensive and limited in the degree of complexity they can investigate. As a more economical alternative, bioinformatics and computer simulation can use available biological data, such as the protein sequence and structural information, to investigate interactions by virus, host, and the environment [60]. Thus, Shen et al. [60] showed that most mutations develop in the hydrophobic regions of HBsAg and POL and that the amino acids that are more likely to be mutated are serine and threonine [60]. Understanding how amino acids mutations develop in HBV proteins can facilitate the rational design of both vaccines and drugs [60], for the prevention and treatment of HBV infection, respectively. By the use of bioinformatics to compare viral and host genomic patterns, together with clinical information, to data from databases can lead to enhanced and individualized antiviral therapy.

3. Bioinformatics tools and databases

3.1. Bioinformatics challenges of HBV

Despite its small genome size of ~3.2 kb, HBV presents several bioinformatic challenges:

1. The genome is circular, with position 1 conventionally taken to be the first "T" nucleotide in the *Eco*R1 restriction site ("GAATTC"). Historically, position 1 was the start of the "Core" region, which is position 1901 in the current numbering system. Therefore, a number of sequences deposited earlier in the public databases are numbered using this outdated system and thus require processing before they can be used in alignments, together with more recently submitted sequences.

2. Four overlapping reading frames are encoded in the circular genome, whereas nucleotides or amino acids are sequenced and processed linearly. Extracting nucleotide or amino acid

sequences for the *S* and *POL* ORFs, which span the *EcoRI* site, from full-length or subgenomic fragments, requires additional processing.

3. The differences in genome lengths between the nine HBV genotypes (ranging from 3182 to 3248 base pairs in length) mean that direct comparison of loci between genotypes is not always possible using the current numbering system. These differences in genome lengths result in genotype alignments containing several regions of gaps, ranging from 3 to 33 nucleotides in length. A possible solution is the implementation of a standardized "universal numbering system" for all HBV genotypes, which we are currently developing.
4. Sequence variability is a feature of HBV. It is, therefore, essential to check all sequences carefully, to distinguish between artifacts and true variation (mutations). Variation within a population at a locus may result in two overlapping peaks on a chromatogram. Superinfections or co-infections with different strains may result in mixed populations, which appear as multiple or misaligned peaks on sequencing chromatograms. Disambiguating these is essential for robust downstream analyses.

3.2. Public sequence databases

The first public sequence database, "GenBank," was established in 1982, having arisen from the earlier Los Alamos database, established in 1979 [61, 62]. Since then, the number of nucleotides in GenBank has doubled approximately every 18 months [63]. The International Nucleotide Sequence Database Collaboration (INSDC) is a collection of three publicly available nucleotide (DNA or RNA) sequence databases, which synchronize data daily [64]. The collection consists of the DNA DataBank of Japan (DDBJ, located in Japan), the European Molecular Biology Laboratory (EMBL, located in the United Kingdom) and GenBank (located in the United States of America). The latest release of the database (release 211.0, from 15 December, 2015; [65]) contains 189,232,925 loci and 203,939,111,071 bases, from 189,232,925 sequences, totaling approximately 742 gigabytes. In addition to the INSDC, many other databases exist, including genome databases, protein sequence, structure and interaction databases, microarray databases, and meta-databases. A list of biological databases on Wikipedia includes over 200 entries [66].

When searching for "hepatitis b virus" across all fields, the GenBank database [63], accessed on 27th January 2016, contained 105,745 sequences. When searching for "hepatitis b virus" in the "organism" field only, 84,119 sequences were found, with the oldest sequence submitted in the early 1980s. Refining this search to include only sequences of 200 nucleotides or longer, and excluding words such as "recombinant," "clone," and "patent," resulted in 68,762 sequences. When this same query was previously executed on 29 November 2015, 67,893 sequences were returned. Therefore, in the 59 days between the two queries, 869 new sequences (of at least 200 nucleotides in length, and not containing the words mentioned previously) were uploaded to GenBank. On average, this equates to almost 15 new HBV sequences added to GenBank per day.

Making use of these sequences in downstream applications, such as multiple sequence alignments or phylogenetic analyses, is often challenging, as it is difficult to query for sufficient

sequences, of the correct genotype, or subgenotype, and covering the required genomic region. In order to overcome this limitation, we have developed a bioinformatics solution, whereby all sequences matching a query are downloaded, curated, and aligned. The algorithm developed allows for the generation of a multiple sequence alignment for each genotype, which contains all the available sequences matching the query and in their correct position and orientation [67].

3.3. Bioinformatics tools for HBV

Workflow	Tool name and description
Chromatograms	**Quality score analyzer**
	• *Plots chromatogram quality scores*
	Automatic contig generator tool
	• *Generates a contig from a forward and reverse chromatogram*
Alignment	**Automatic alignment clean-up tool**
	• *Eliminates "gap-columns" and disambiguate ambiguous bases*
	Mind the gap
	• *Splits FASTA file based on gap threshold per column*
Analysis	**Babylon**
	• *Extracts HBV protein sequences (ORFs)*
	Wild-type 2 × 2
	• *Calculates 2 × 2 wild-type/mutant contingency tables*
	Divergence calculator*
	• *Intra- and Inter-group divergence with custom groups*
	Rafael*
	• *Generates random subsets from an input FASTA file*
Serotyping	**HBV serotyper tool**
	• *Determines HBV serotype*
Phylogenetics	**Pipeline: TreeMail**
	• *Generates a phylogenetic tree*
GenBank Submission	**PadSeq**
	• *Places two HBV sequence fragments on a backbone template*

¶Table modified from Bell and Kramvis [68].
*Described for the first time here.

Table 2. List of the online tools developed and the workflow process at which each would be used¶.

A standard molecular biology laboratory workflow includes DNA extraction, polymerase chain reaction (PCR) amplification, direct DNA sequencing, viewing and checking of chromatograms, preparation of curated sequences, multiple sequence alignment, sequence analysis, serotyping, genotyping, phylogenetic analysis, and preparation of sequences for submission to the GenBank public sequence database [68]. Each of these steps presents data processing challenges, many of which have been addressed by the development of a suite of online tools (**Table 2**) [68].

Name	URL	Usage	References
Geno2Pheno [hbv]	http://www.genafor.org/g2p_hbv/index.php	Drug resistance mutations, escape mutant analysis	[69]
HBV Blast Search	http://www.bioafrica.net/blast/hbvblast.html	Genotyping, drug resistance database	[70]
HBV STAR	http://www.vgb.ucl.ac.uk/starn.shtml	Genotyping	[71]
HBVRegDB	http://lancelot.otago.ac.nz/HBVRegDB/	Annotation, alignments, information about conserved regions	[72]
HBVdb	http://hbvdb.ibcp.fr/	Genotyping, annotation, drug resistance database	[73]
HBVseq	http://hivdb.stanford.edu/HBV/HBVseq/development/HBVseq.html		
Hepatitis virus database	http://s2as02.genes.nig.ac.jp/	Genotyping, sequence alignment and map viewing	
Hepatitis virus database	http://www.ibibiobase.com/projects/hepatitis/index.htm		
NCBI genotyping tool	http://www.ncbi.nlm.nih.gov/projects/genotyping/	Genotyping	[74]
Oxford HBV subtyping tool	http://www.bioafrica.net/rega-genotype/html/subtypinghbv.html	Genotyping	[70]
RegaDB	http://regaweb.med.kuleuven.be/software/regadb		
SeqHepB	http://www.seqhepb.com/	Sequence analysis, genotyping, detection of clinically important mutations	[75]

¶Table modified from [67].

Table 3. Currently available HBV websites and databases¶.

Any operating system platform from any location with an internet connection can be used to access stand-alone, web-based tools. There is no requirement to install and learn new bioin-

formatics software, as these tools can be used when required. A system for processing ultra-deep pyrosequencing (amplicon resequencing) data has also been developed [51]. In addition, a number of HBV-specific websites and databases are currently available, a selection of which are represented in **Table 3**.

3.4. New bioinformatics tools for HBV

Here, we present two newly developed tools for the bioinformatic analysis of HBV.

3.4.1. Divergence calculator [http://hvdr.bioinf.wits.ac.za/divergence/]

One method of classifying HBV sequences into genotype or subgenotype is to examine nucleotide sequence divergence between sequences. This divergence calculation is performed by totaling the number of nucleotides, which differ, between two aligned sequences and computing the percentage difference. The divergence calculator (**Figure 3**) performs various divergence calculations on groups of sequences from nucleotide or amino acid multiple sequence alignments in FASTA format. A minimum of one group containing two sequences, or two groups containing one sequence each, must be specified.

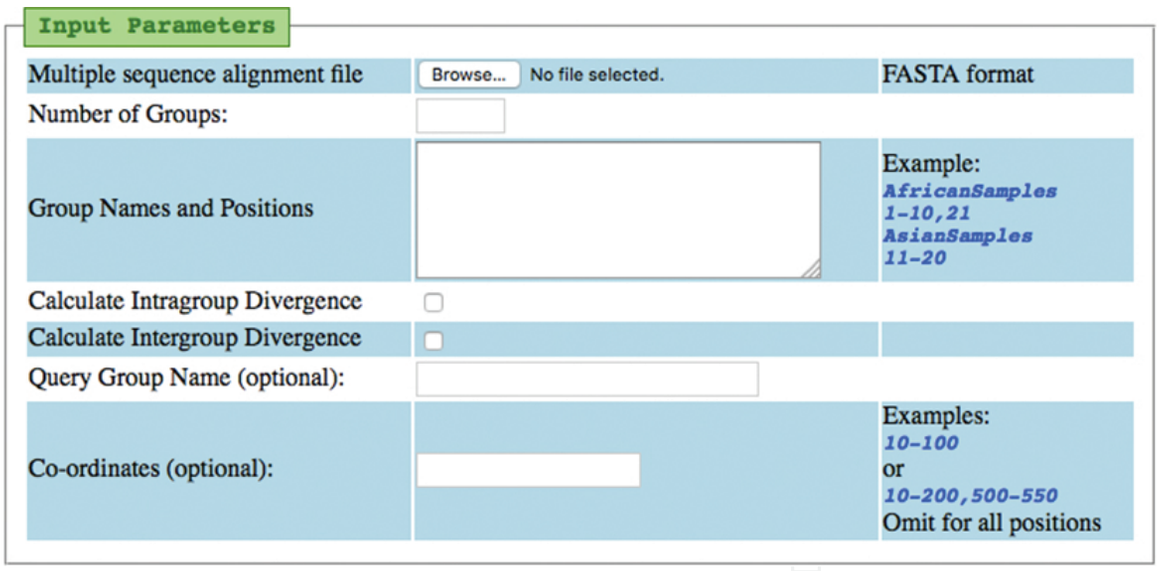

Figure 3. The input screen of the divergence calculator in which sequences are extracted and allocated to groups and other parameters specified.

As an example, consider an alignment of 10 genotype A sequences (group 1) and 10 genotype D sequences (group 2). Intra-group divergence, for each group, is calculated by comparing each sequence in group 1 with each other sequence in group 1 and then calculating the median, mean, and standard deviation of the divergences. This is then repeated for group 2. The inter-group divergence compares each sequence in group 1 with each sequence in group 2, and then

calculates the median, mean, and standard deviation. If more than two groups are specified, the calculations iterate over all groups in turn.

If the optional "query" group is specified, the tool compares each sequence in the query group with each sequence in the other group or groups, but outputs statistics for each sequence in the query group individually. This method would typically be used with a set of unknown query sequences and one or more groups of reference sequences. A comprehensive list of descriptive statistics is included on the output page for each analysis.

3.4.2. Random FASTA extraction and allocation (RAFAEL) [http://hvdr.bioinf.wits.ac.za/rafael/]

In some analyses, particularly when constructing phylogenetic trees, it may be desirable to extract one or more random subsets of sequences from a master or reference alignment. The "RAFAEL" tool was designed to perform this task. This tool takes an input file in FASTA format, which does not have to be aligned and generates one or more subsets of the file, each containing a random selection of the specified number of sequences. The number of sequences may be specified as a count, or as a percentage of the number of sequences in the input file. There are guaranteed to be no duplicate sequences within each subset. However, duplicates may exist in multiple subsets, as subsets are not unique.

3.5. Open-source software

In addition to biological databases, a large variety of biological analysis software, which is generally genome agnostic, is available. As with software in any field, the licensing terms and commercial costs of these packages vary widely. Packages, which may be free of cost, may not necessarily be open-source, for example.

The Free Software Foundation (FSF) [76, 77] defines free software as software which "respects the users' freedom" in the sense that "users have the freedom to run, copy, distribute, study, change, and improve the software". As such, "free" is "a matter of liberty, not price". Free software, therefore, does not necessarily have to be made available at no cost or be a non-commercial project. Furthermore, software, which is provided at no cost, may not be "free" in the sense described above.

The term "open-source" is often used when referring to "free" software. However, the two terms are not synonymous, although there is some overlap. Open-source software may, or may not, be free software, depending on the restrictions placed on users by the software. If the user is not free to distribute, change, and improve the software, even if it is open source, then it cannot be considered to be free software. Most software, for which a license is purchased, is not free, or open source. The user does not have the freedom to distribute the software, or to use it on any computer chosen.

3.6. Recommended software

A list of recommended freely available download software is presented in **Table 4**. Comprehensive lists of open-source bioinformatics software can be found elsewhere [78].

Software name	Software description	Website (http://)	Lin*	Mac*	Win*	References
Unipro UGene	Integrated bioinformatics suite	ugene.net	Yes	Yes	Yes	[79]
MEGA 6	Integrated bioinformatics suite; command-line version available	megasoftware.net	Emu	Yes	Yes	[80]
BioEdit	Multiple sequence alignment viewer and editor	www.mbio.ncsu.edu/bioedit/bioedit.html	No	No	Yes	
SeaView	Multiple sequence alignment viewer and editor, and molecular phylogenetics; opens GeneDoc "MSF" files	doua.prabi.fr/software/seaview	Yes	Yes	Yes	[81]
AliView	Multiple sequence alignment viewer and editor	www.ormbunkar.se/aliview	Yes	Yes	Yes	[82]
GeneDoc	Multiple sequence alignment viewer and editor, and shading utility	iubio.bio.indiana.edu/soft/molbio/ibmpc/genedoc-readme.html	No	No	Yes	
PHYLIP	Programs for interring phylogenies; website includes comprehensive list of other phylogenetics programs	evolution.genetics.washington.edu/phylip.html	Yes	Yes	Yes	[83]
MrBayes	Bayesian inference and model choice using Markov Chain Monte Carlo	mrbayes.sourceforge.net	Com	Yes	Yes	[84]
BEAST	Bayesian analysis of molecular sequences using Markov Chain Monte Carlo	beast.bio.ed.ac.uk	Yes	Yes	Yes	[85]
FigTree	Graphical viewer and editor of phylogenetic trees	tree.bio.ed.ac.uk/software/figtree/	Yes	Yes	Yes	
Archaeopteryx	Graphical viewer and editor of phylogenetic trees	sites.google.com/site/cmzmasek/home/software/archaeopteryx	Yes	No	Yes	[86]
EMBOSS	A suite of command-line tools for molecular biology	emboss.sourceforge.net/	Yes	Yes	Emu	
JalView	Multiple sequence alignment viewer and editor	www.jalview.org	Yes	Yes	Yes	[87]
Finch TV	DNA sequence trace (chromatogram) viewer	www.geospiza.com/Products/finchtv.shtml	Yes	Yes	Yes	

Software name	Software description	Website (http://)	Lin*	Mac*	Win*	References
Chromas	DNA sequence trace (chromatogram) viewer	technelysium.com.au/?page_id=13	No	No	Yes	

*"GUI" = graphical user interface, "CL" = command line interface, "OSS" = open-source software, "Lin" = GNU/Linux, "Mac" = Apple MacIntosh, "Win" = Microsoft Windows, "Emu" = emulator or virtual machine recommended by authors, "Com" = compilation from source code required.

Table 4. Bioinformatics software available free of charge for various computer operating system platforms.

4. Conclusion

The unique genome structure and molecular biology of HBV pose a number of challenges, and thus, the development of bioinformatic tools has facilitated a more comprehensive and detailed analysis and understanding of the origin, evolution, transmission, and response to antiviral agents of HBV and its interaction with the host. There are a wide range of free and commercially available tools, which have been developed for different applications. The availability and applications of high-throughput sequencing techniques and the advancement of "-omics" will continue to provide additional challenges, which will need to be addressed by further computational solutions.

Acknowledgements

Trevor Bell is the recipient of a National Research Foundation (NRF) Scarce Skills Post-Doctoral Fellowship (GUN#86215) and Anna Kramvis received funding from the National Research Foundation (GUN#65530, GUN#93516).

Author details

Trevor Graham Bell* and Anna Kramvis

*Address all correspondence to: TrevorGrahamBell@gmail.com

Hepatitis Virus Diversity Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

References

[1] Qi H, Wu NC, Du Y, Wu TT, Sun R. High-resolution genetic profile of viral genomes: why it matters. Curr Opin Virol. 2015;14:62–70. PubMed PMID: 26364133.

[2] Hepatitis B Fact sheet no. 204 [Internet]. 2015. Available from: http://www.who.int/mediacentre/factsheets/fs204/en/[Accessed: 2016-01-16]

[3] Ott JJ, Stevens GA, Groeger J, Wiersma ST. Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. Vaccine. 2012;30(12):2212–9. PubMed PMID: 22273662. Epub 2012/01/26. eng.

[4] Tiollais P, Pourcel C, Dejean A. The hepatitis B virus. Nature. 1985;317(6037):489–95. PubMed PMID: 2995835.

[5] Lucifora J, Arzberger S, Durantel D, Belloni L, Strubin M, Levrero M, et al. Hepatitis B virus X protein is essential to initiate and maintain virus replication after infection. J Hepatol. 2011;55(5):996–1003. PubMed PMID: 21376091.

[6] Kramvis A, Kew MC. The core promoter of hepatitis B virus. J Viral Hepat. 1999;6(6):415–27. PubMed PMID: 10607259. Epub 1999/12/22. eng.

[7] Moolla N, Kew M, Arbuthnot P. Regulatory elements of hepatitis B virus transcription. J Viral Hepat. 2002;9(5):323–31. PubMed PMID: 12225325. Epub 2002/09/13. eng.

[8] Summers J, Mason WS. Replication of the genome of a hepatitis B-like virus by reverse transcription of an RNA intermediate. Cell. 1982;29(2):403–15. PubMed PMID: 6180831. Epub 1982/06/01. eng.

[9] Yan H, Zhong G, Xu G, He W, Jing Z, Gao Z, et al. Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. eLife. 2012;1:e00049. PubMed PMID: 23150796. Pubmed Central PMCID: 3485615.

[10] Rabe B, Glebe D, Kann M. Lipid-mediated introduction of hepatitis B virus capsids into nonsusceptible cells allows highly efficient replication and facilitates the study of early infection events. J Virol. 2006;80(11):5465–73. PubMed PMID: 16699026. Pubmed Central PMCID: 1472160.

[11] Kock J, Rosler C, Zhang JJ, Blum HE, Nassal M, Thoma C. Generation of covalently closed circular DNA of hepatitis B viruses via intracellular recycling is regulated in a virus specific manner. Plos Pathogens. 2010;6(9):e1001082. PubMed PMID: 20824087. Pubmed Central PMCID: 2932716.

[12] Beck J, Nassal M. Hepatitis B virus replication. World J Gastroenterol. 2007;13(1):48–64. PubMed PMID: 17206754. Pubmed Central PMCID: 4065876.

[13] Kramvis A, Kew MC. Structure and function of the encapsidation signal of hepadnaviridae. J Viral Hepat. 1998;5(6):357–67. PubMed PMID: 9857345. Epub 1998/12/19. eng.

[14] Kramvis A, Kew M, Francois G. Hepatitis B virus genotypes. Vaccine. 2005;23(19):2409–23. PubMed PMID: 15752827. Epub 2005/03/09. eng.

[15] Wang GH, Seeger C. Novel mechanism for reverse transcription in hepatitis B viruses. J Virol. 1993;67(11):6507–12. PubMed PMID: 7692081. Pubmed Central PMCID: 238087.

[16] Glebe D, Bremer CM. The molecular virology of hepatitis B virus. Semin Liver Dis. 2013;33(2):103–12. PubMed PMID: 23749666.

[17] Steinhauer DA, Holland JJ. Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J Virol. 1986;57(1):219–28. PubMed PMID: 3001347. Pubmed Central PMCID: 252718. Epub 1986/01/01. eng.

[18] Yu H, Yuan Q, Ge SX, Wang HY, Zhang YL, Chen QR, et al. Molecular and phylogenetic analyses suggest an additional hepatitis B virus genotype "I". Plos One. 2010;5(2):e9297. PubMed PMID: 20174575. Pubmed Central PMCID: 2824819. Epub 2010/02/23. eng.

[19] Norder H, Courouce AM, Coursaget P, Echevarria JM, Lee SD, Mushahwar IK, et al. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology. 2004;47(6):289–309. PubMed PMID: 15564741. Epub 2004/11/27. eng.

[20] Tatematsu K, Tanaka Y, Kurbanov F, Sugauchi F, Mano S, Maeshiro T, et al. 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. J Virol. 2009;83(20):10538–47. PubMed PMID: 19640977. Pubmed Central PMCID: 2753143. Epub 2009/07/31. eng.

[21] Kramvis A. Genotypes and genetic variability of hepatitis B virus. Intervirology. 2014;57(3–4):141–50. PubMed PMID: 25034481. Epub 2014/07/19. eng.

[22] Kramvis A, Kew MC. Relationship of genotypes of hepatitis B virus to mutations, disease progression and response to antiviral therapy. J Viral Hepat. 2005;12(5):456–64. PubMed PMID: 16108759. Epub 2005/08/20. eng.

[23] Kramvis A, Paraskevis D. Subgenotype A1 of HBV—tracing human migrations in and out of Africa. Antivir Ther. 2013;18(3 Pt B):513–21. PubMed PMID: 23792935. Epub 2013/06/26. eng.

[24] Osiowy C, Kaita K, Solar K, Mendoza K. Molecular characterization of hepatitis B virus and a 9-year clinical profile in a patient infected with genotype I. J Med Virol. 2010;82(6): 942–8. PubMed PMID: 20419807. Epub 2010/04/27. eng.

[25] Wiegand J, Hasenclever D, Tillmann HL. Should treatment of hepatitis B depend on hepatitis B virus genotypes? A hypothesis generated from an explorative analysis of published evidence. Antivir Ther. 2008;13(2):211–20. PubMed PMID: 18505172.

[26] Vermeulen M, Dickens C, Lelie N, Walker E, Coleman C, Keyter M, et al. Hepatitis B virus transmission by blood transfusion during 4 years of individual-donation nucleic acid testing in South Africa: estimated and observed window period risk. Transfusion. 2012;52:880–92. PubMed PMID: 21981386. Epub 2011/10/11. Eng.

[27] Martinot-Peignoux M, Marcellin P. Virological and serological tools to optimize the management of patients with chronic hepatitis B. Liver Int. 2016;36 Suppl 1:78–84. PubMed PMID: 26725902.

[28] Bartholomeusz A, Schaefer S. Hepatitis B virus genotypes: comparison of genotyping methods. Rev Med Virol. 2004;14(1):3–16. PubMed PMID: 14716688.

[29] Guirgis BS, Abbas RO, Azzazy HM. Hepatitis B virus genotyping: current methods and clinical implications. Int J Infect Dis IJID. 2010;14(11):e941–53. PubMed PMID: 20674432.

[30] Kramvis A, Arakawa K, Yu MC, Nogueira R, Stram DO, Kew MC. Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. J Med Virol. 2008;80(1):27–46. PubMed PMID: 18041043. Epub 2007/11/28. eng.

[31] Hu X, Margolis HS, Purcell RH, Ebert J, Robertson BH. Identification of hepatitis B virus indigenous to chimpanzees. Proc Natl Acad Sci USA. 2000;97(4):1661–4. PubMed PMID: 10677515. Pubmed Central PMCID: 26492. Epub 2000/03/04. eng.

[32] Norder H, Courouce AM, Magnius LO. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology. 1994;198(2):489–503. PubMed PMID: 8291231. Epub 1994/02/01. eng.

[33] Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, Gojobori T. Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol. 1997;44 Suppl 1:S83–90. PubMed PMID: 9071016. Epub 1997/01/01. eng.

[34] Torres C, Fernandez MD, Flichman DM, Campos RH, Mbayed VA. Influence of overlapping genes on the evolution of human hepatitis B virus. Virology. 2013;441(1): 40–8. PubMed PMID: 23541083. Epub 2013/04/02. eng.

[35] Tedder RS, Bissett SL, Myers R, Ijaz S. The 'Red Queen' dilemma–running to stay in the same place: reflections on the evolutionary vector of HBV in humans. Antivir Ther. 2013;18(3 Pt B):489–96. PubMed PMID: 23792884. Epub 2013/06/26. eng.

[36] Andernach IE, Hunewald OE, Muller CP. Bayesian Inference of the Evolution of HBV/E. Plos One. 2013;8(11):e81690. PubMed PMID: 24312336. Pubmed Central PMCID: 3843692. Epub 2013/12/07. eng.

[37] Fares MA, Holmes EC. A revised evolutionary history of hepatitis B virus (HBV). J Mol Evol. 2002;54(6):807–14. PubMed PMID: 12029362. Epub 2002/05/25. eng.

[38] Orito E, Mizokami M, Ina Y, Moriyama EN, Kameshima N, Yamamoto M, et al. Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc Natl Acad Sci USA. 1989;86(18):7059–62. PubMed PMID: 2780562. Pubmed Central PMCID: 297993. Epub 1989/09/01. eng.

[39] Gunther S, Sommer G, Von Breunig F, Iwanska A, Kalinina T, Sterneck M, et al. Amplification of full-length hepatitis B virus genomes from samples from patients with low levels of viremia: frequency and functional consequences of PCR-introduced

mutations. J Clin Microbiol. 1998;36(2):531–8. PubMed PMID: 9466771. Pubmed Central PMCID: 104572. Epub 1998/02/18. eng.

[40] Paraskevis D, Magiorkinis G, Magiorkinis E, Ho SY, Belshaw R, Allain JP, et al. Dating the origin and dispersal of hepatitis B virus infection in humans and primates. Hepatology. 2013;57(3):908–16. PubMed PMID: 22987324. Epub 2012/09/19. eng.

[41] Bouckaert R, Alvarado-Mora MV, Pinho JR. Evolutionary rates and HBV: issues of rate estimation with Bayesian molecular methods. Antivir Ther. 2013;18(3 Pt B):497–503. PubMed PMID: 23792904. Epub 2013/06/26. eng.

[42] Alvarado Mora MV, Romano CM, Gomes-Gouvea MS, Gutierrez MF, Botelho L, Carrilho FJ, et al. Molecular characterization of the Hepatitis B virus genotypes in Colombia: a Bayesian inference on the genotype F. Infect Genet Evol. 2011;11(1):103–8. PubMed PMID: 20951841. Epub 2010/10/19. eng.

[43] Suh A, Brosius J, Schmitz J, Kriegs JO. The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses. Nat Commun. 2013;4:1791. PubMed PMID: 23653203.

[44] Yang Z, Lauder IJ, Lin HJ. Molecular evolution of the hepatitis B virus genome. J Mol Evol. 1995;41(5):587–96. PubMed PMID: 7490773.

[45] Domingo E, Sabo D, Taniguchi T, Weissmann C. Nucleotide sequence heterogeneity of an RNA phage population. Cell. 1978;13(4):735–44. PubMed PMID: 657273.

[46] Buti M, Tabernero D, Mas A, Homs M, Prieto M, Rodriguez-Frias F, et al. Hepatitis B virus quasispecies evolution after liver transplantation in patients under long-term lamivudine prophylaxis with or without hepatitis B immune globulin. Transpl Infect Dis. 2015;17(2):208–20. PubMed PMID: 25641570.

[47] Homs M, Buti M, Tabernero D, Quer J, Sanchez A, Corral N, et al. Quasispecies dynamics in main core epitopes of hepatitis B virus by ultra-deep-pyrosequencing. World J Gastroenterol. 2012;18(42):6096–105. PubMed PMID: 23155338. Pubmed Central PMCID: 3496886.

[48] Homs M, Caballero A, Gregori J, Tabernero D, Quer J, Nieto L, et al. Clinical application of estimating hepatitis B virus quasispecies complexity by massive sequencing: correlation between natural evolution and on-treatment evolution. Plos One. 2014;9(11):e112306. PubMed PMID: 25393280. Pubmed Central PMCID: 4231103.

[49] Ramirez C, Gregori J, Buti M, Tabernero D, Camos S, Casillas R, et al. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antivir Res. 2013;98(2):273–83. PubMed PMID: 23523552.

[50] Rodriguez-Frias F, Buti M, Tabernero D, Homs M. Quasispecies structure, cornerstone of hepatitis B virus infection: mass sequencing approach. World J Gastroenterol.

2013;19(41):6995–7023. PubMed PMID: 24222943. Pubmed Central PMCID: 3819535. Epub 2013/11/14. Eng.

[51] Yousif M, Bell TG, Mudawi H, Glebe D, Kramvis A. Analysis of ultra-deep pyrosequencing and cloning based sequencing of the basic core promoter/precore/core region of hepatitis B virus using newly developed bioinformatics tools. Plos One. 2014;9(4):e95377. PubMed PMID: 24740330. Epub 2014/04/18. eng.

[52] Liu WC, Lin CP, Cheng CP, Ho CH, Lan KL, Cheng JH, et al. Aligning to the sample-specific reference sequence to optimize the accuracy of next-generation sequencing analysis for hepatitis B virus. Hepatol Int. 2016;10(1):147–57. PubMed PMID: 26208819. Pubmed Central PMCID: 4722079.

[53] Homs M, Buti M, Quer J, Jardi R, Schaper M, Tabernero D, et al. Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome. Nucleic Acids Res. 2011;39(19):8457–71. PubMed PMID: 21742757. Pubmed Central PMCID: 3201856. Epub 2011/07/12. eng.

[54] Okamoto D, Nakayama H, Ikeda T, Ikeya S, Nagashima S, Takahashi M, et al. Molecular analysis of the interspousal transmission of hepatitis B virus in two Japanese patients who acquired fulminant hepatitis B after 50 and 49 years of marriage. J Med Virol. 2014;86(11):1851–60. PubMed PMID: 25132075.

[55] Lago BV, Mello FC, Kramvis A, Niel C, Gomes SA. Hepatitis B virus subgenotype A1: evolutionary relationships between Brazilian, African and Asian Isolates. Plos One. 2014;9(8):e105317. PubMed PMID: 25122004. Pubmed Central PMCID: 4133366. Epub 2014/08/15. eng.

[56] Liu J, Yang HI, Lee MH, Lu SN, Jen CL, Batrla-Utermann R, et al. Spontaneous seroclearance of hepatitis B seromarkers and subsequent risk of hepatocellular carcinoma. Gut. 2014;63(10):1648–57. PubMed PMID: 24225939.

[57] Charuworn P, Hengen PN, Aguilar Schall R, Dinh P, Ge D, Corsa A, et al. Baseline interpatient hepatitis B viral diversity differentiates HBsAg outcomes in patients treated with tenofovir disoproxil fumarate. J Hepatol. 2015;62(5):1033–9. PubMed PMID: 25514556.

[58] Gish RG, Given BD, Lai CL, Locarnini SA, Lau JY, Lewis DL, et al. Chronic hepatitis B: virology, natural history, current management and a glimpse at future opportunities. Antivir Res. 2015;121:47–58. PubMed PMID: 26092643.

[59] Zoulim F, Durantel D, Deny P. Management and prevention of drug resistance in chronic hepatitis B. Liver Int. 2009;29 Suppl 1:108–15. PubMed PMID: 19207973. Epub 2009/02/12. eng.

[60] Shen K, Shen L, Wang J, Jiang Z, Shen B. Understanding amino acid mutations in hepatitis B virus proteins for rational design of vaccines and drugs. Adv Protein Chem Struct Biol. 2015;99:131–53. PubMed PMID: 26067819.

[61] Kanehisa M, Fickett JW, Goad WB. A relational database system for the maintenance and verification of the Los Alamos sequence library. Nucleic Acids Res. 1984;12(1 Pt 1): 149–58. PubMed PMID: 6694899. Pubmed Central PMCID: 320992.

[62] Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, et al. The GenBank genetic sequence databank. Nucleic Acids Res. 1986;14(1):1–4. PubMed PMID: 3945546. Pubmed Central PMCID: 339347.

[63] Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2015;43(Database issue):D30–5. PubMed PMID: 25414350. Pubmed Central PMCID: 4383990.

[64] Karsch-Mizrachi I, Nakamura Y, Cochrane G, International Nucleotide Sequence Database C. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2012;40(Database issue):D33–7. PubMed PMID: 22080546. Pubmed Central PMCID: 3244996.

[65] Genetic Sequence Data Bank [Internet]. 2016. Available from: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt [Accessed: 2016-01-27]

[66] List of biological databases [Internet]. 2016. Available from: https://en.wikipedia.org/wiki/List_of_biological_databases [Accessed: 2016-01-27]

[67] Bell, T.G., Yousif, Y., Kramvis, A. Bioinformatic curation and alignment of genotyped hepatitis B virus (HBV) sequence data from the GenBank public database. Submitted.

[68] Bell TG, Kramvis A. Bioinformatics tools for small genomes, such as hepatitis B virus. Viruses. 2015;7(2):781–97. PubMed PMID: 25690798. Epub 2015/02/19. eng.

[69] Beerenwinkel N, Daumer M, Oette M, Korn K, Hoffmann D, Kaiser R, et al. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res. 2003;31(13):3850–5. PubMed PMID: 12824435. Pubmed Central PMCID: 168981.

[70] de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics. 2005;21(19):3797–800. PubMed PMID: 16076886.

[71] Myers R, Clark C, Khan A, Kellam P, Tedder R. Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. J Gen Virol. 2006;87(Pt 6):1459–64. PubMed PMID: 16690910.

[72] Panjaworayan N, Roessner SK, Firth AE, Brown CM. HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in hepatitis B virus sequences. Virol J. 2007;4:136. PubMed PMID: 18086305. Pubmed Central PMCID: 2235840. Epub 2007/12/19. eng.

[73] Hayer J, Jadeau F, Deleage G, Kay A, Zoulim F, Combet C. HBVdb: a knowledge database for hepatitis B virus. Nucleic Acids Res. 2013;41(Database issue):D566–70. PubMed PMID: 23125365. Pubmed Central PMCID: 3531116.

[74] Rozanov M, Plikat U, Chappey C, Kochergin A, Tatusova T. A web-based genotyping resource for viral sequences. Nucleic Acids Res. 2004;32(Web Server issue):W654–9. PubMed PMID: 15215470. Pubmed Central PMCID: 441557.

[75] Yuen LK, Ayres A, Littlejohn M, Colledge D, Edgely A, Maskill WJ, et al. SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. Antivir Res. 2007;75(1):64–74. PubMed PMID: 17215050.

[76] Free Software Foundation [Internet]. 2016. Available from: http://www.fsf.org/ [Accessed: 2016-01-27]

[77] What is free software? [Internet]. 2016. Available from: http://www.gnu.org/philosophy/free-sw.html [Accessed: 2016-01-27]

[78] List of open-source bioinformatics software [Internet]. 2016. Available from: https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software [Accessed: 2016-01-28]

[79] Okonechnikov K, Golosova O, Fursov M, team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–7. PubMed PMID: 22368248.

[80] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–9. PubMed PMID: 24132122. Pubmed Central PMCID: 3840312.

[81] Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27(2):221–4. PubMed PMID: 19854763.

[82] Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276–8. PubMed PMID: 25095880. Pubmed Central PMCID: 4221126.

[83] Phylogeny Programs [Internet]. 2016. Available from: http://evolution.genetics.washington.edu/phylip/software.html [Accessed: 2016-03-07]

[84] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–5. PubMed PMID: 11524383.

[85] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73. PubMed PMID: 22367748. Pubmed Central PMCID: 3408070.

[86] Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinform. 2009;10:356. PubMed PMID: 19860910. Pubmed Central PMCID: 2774328.

[87] Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91. PubMed PMID: 19151095. Pubmed Central PMCID: 2672624.