Fakultät für Medizin

Estimation of the global distribution of hepatitis B virus genotypes and clinically relevant variants

Stoyan Velkov

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Dr. Stefan Engelhardt

Prüfende der Dissertation:

1. Univ.-Prof. Dr. Ulrike Protzer
2. Univ.-Prof. Dr. Dmitrij Frishman

Die Dissertation wurde am 01.10.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 07.06.2022 angenommen.

Table of Contents

1 Abstract

Worldwide, an estimated 257 million people are chronically infected with hepatitis B virus (HBV). HBV can be divided into nine genotypes, termed A – I. So far, an in dept investigation of the global distribution of these genotypes has not been conducted. To address this lacking information, a literature review of publications describing genotyping data throughout the world was performed. A total of 900 publications were manually evaluated from which 213 records of genotyping data were extracted covering 125 countries. Based on previously published HBV prevalence and population data, the global genotype distribution of chronic HBV infections was approximated. The genotypes that caused the most chronic hepatitis B cases were A – 17%, B – 14%, C – 26%, D – 22% and E – 18%. In total these genotypes were found to be responsible for 96% of the chronic HBV infections worldwide, while the four genotypes F – I contributed together to less than 2%. The remaining infections were accounted to inter-genotype recombinants or mixed infections with more than one genotype or could not be defined. The study describes the up-to-date overview over global genotyping data and is the first of its kind to estimate the global genotype distribution of HBV. The results suggest that at least genotypes A – E should be taken into account when establishing novel therapies, vaccines and treatment models for HBV due to their broad distribution [1].

The second project of this thesis covers the clinically relevant HBV variants. Distinct mutations on the viral genome can affect the course of disease and the treatment options available. Additionally, the different genotypes contribute to different outcomes of HBV infection. Therefore, we performed a comprehensive *in-situ* analysis of the HBV variants by genotype and region. Publicly available full genome sequences were used to evaluate the occurrence of mutations of the HBV surface antigen (HBsAg), reverse transcriptase (RT) and the basal core promotor/pre-core region. The data resulted in an association of genotypes to distinct, clinically relevant mutations, and additionally showed that populations of certain world regions were at least partially susceptible to specific HBV variants. Three HBV mutations, which were initially proposed to convey clinically relevant phenotypes, were not found in any sequence, R122P in the S; P177G and F249A in the polymerase gene – questioning their real-world significance. Regional accumulation of mutations which convey resistance to nucleos(t)ides analogs, or promote disease progression, can be retrieved from the data which can serve as a primary orientation for selecting suitable antiviral drugs and help improve diagnostic tests.

The third project of this thesis focused on developing a database for hepatitis D virus (HDV) sequences, termed HDVdb. Studies regarding HDV are of great demand, as HDV is the causing agent of the most severe and rapidly progression form of viral hepatitis,

leading in 70% of the cases to cirrhosis and accounting for 15 – 20 million worldwide chronic carriers. As an all-in-one webservice, the HDVdb webpage, was developed covering all publicly available viral genome sequences. In addition, a genotyping tool for HDV was developed, accompanied by additional services like sequence alignment, primer design, phylogenetic analysis and visualization. HDVdb is aiming to become the most important platform for bioinformatic approaches regarding HDV, and was therefore built to provide future updates and upgrades to its users.

2 Zusammenfassung

Weltweit sind schätzungsweise 257 Millionen Menschen chronisch mit dem Hepatitis-B-Virus (HBV) infiziert. HBV lässt sich in neun Genotypen unterteilen, die mit A – I bezeichnet werden. Eine genaue Charakterisierung der globalen Verteilung dieser Genotypen stand bisher noch aus. Deshalb wurde zunächst eine Literaturrecherche bezüglich der Genotypisierungsdaten und deren regionaler Verteilung durchgeführt. Anschließend wurden insgesamt 900 Publikationen manuell ausgewertet, aus denen 213 Datensätze mit Genotypisierungsdaten extrahiert wurden, welche 125 Länder abdecken. Basierend auf zuvor veröffentlichten HBV-Prävalenz- und Bevölkerungsdaten wurde die globale Genotypen-Verteilung chronischer HBV-Infektionen ermittelt. Die häufigsten Genotypen waren A – 17 %, B – 14 %, C – 26 %, D – 22 % und E – 18 %. Insgesamt sind diese Genotypen für 96 % der chronischen HBV-Infektionen weltweit verantwortlich, während die restlichen vier Genotypen F – I zusammen zu weniger als 2 % der Infektionen beitragen. Die übrigen Infektionen wurden auf rekombinante Genotypen oder Mischinfektionen mit mehr als einem Genotyp zurückgeführt oder konnten nicht definiert werden. Die Studie beschreibt die aktuell geschätzte globale Genotypen-Verteilung und ist die erste ihrer Art für HBV. Die Ergebnisse legen nahe, dass zumindest die Genotypen A – E aufgrund ihrer weiten Verbreitung bei der Etablierung neuer Therapien, Impfstoffe und Behandlungsmodelle für HBV berücksichtigt werden sollten [1].

Das zweite Projekt dieser Arbeit befasst sich mit den klinisch relevanten HBV-Varianten. Unterschiedliche Mutationen im viralen Genom haben Einfluss auf den Krankheitsverlauf und die verfügbaren Behandlungsmöglichkeiten. Zusätzlich tragen die verschiedenen Genotypen zu unterschiedlichen Verläufen der HBV-Infektion bei. Daher wurde eine umfassende in-situ-Analyse der HBV-Varianten nach Genotyp und Region durchgeführt. Öffentlich verfügbare Vollgenomsequenzen wurden verwendet, um das Auftreten von Mutationen des HBV-Oberflächenantigens (HBsAg), der Reversen Transkriptase (RT) und der basalen Core-Promotor/Pre-Core-Region zu beschreiben. Die Daten ergaben eine Assoziation von Genotypen zu verschiedenen klinisch relevanten Mutationen und zeigten zusätzlich, dass die Populationen in den jeweiligen Regionen der Welt teilweise für konkrete HBV-Varianten anfällig waren. Drei dieser als klinisch relevant charakterisierten Mutationen wurden in keiner Sequenz gefunden, R122P im S-Gen sowie P177G und F249A im RT-Gen – was ihre Bedeutung in Frage stellt. Regionale Häufungen von Resistenzen gegen Nukleos(t)idsanaloga lassen sich aus den Daten ablesen und können als erste Orientierung für die Auswahl geeigneter Therapien dienen und das Design zukünftiger diagnostischer Tests verbessern.

Das letzte Thema in dieser Thesis basiert auf der bioinformatischen Analyse viraler genomischer Sequenzen, um eine Hepatitis-D-Virus (HDV)-Sequenzdatenbank, genannt HDVdb, zu erstellen. Studien zu HDV sind von großer Bedeutung, da HDV der Erreger der schwersten und am schnellsten fortschreitenden Form der viralen Hepatitis ist, die in 70% der Fälle zu einer Zirrhose führt und für 15 - 20 Millionen chronische Träger weltweit verantwortlich ist. Als ein Komplettpaket wurde die HDVdb-Webseite entwickelt, die alle öffentlich verfügbaren viralen Genomsequenzen umfasst. Darüber hinaus wurde ein Genotypisierungs-Tool für HDV entwickelt und bereitgestellt, gefolgt von zusätzlichen Diensten wie Sequenz-Alignment, Primer-Design, phylogenetische Analysen und Visualisierung. HDVdb hat das Ziel, die wichtigste Plattform für bioinformatische Ansätze in Bezug auf HDV zu werden und ist daher darauf ausgelegt, seinen Nutzern zukünftige Updates und Upgrades zur Verfügung zu stellen.

3 Introduction

3.1 Hepatitis B Virus

An estimated 257 million humans are chronically infected with Hepatitis B virus (HBV) [2], reflecting 3.5 % of the world's population. HBV is therefore a leading cause of death attributed to infectious diseases. Chronic HBV is accounting for 887.000 deaths per year by causing liver cirrhosis and hepatocellular carcinoma (HCC) [2]. Though prophylactic vaccines have been around for approximately 40 years, providing a high effective protection against the virus [3], HBV still remains a major public health problem. Especially in certain regions like East and Southeast Asia, the Pacific Islands and sub-Saharan Africa the prevalence is between 5 % – 25 % of the total population [4].

3.1.1 Classification

HBV is a relatively small (42 nm of diameter) enveloped, partially double-stranded, relaxed circular DNA (rcDNA) virus and replicates via an RNA intermediate. It is a hepatotropic prototype member of the family of *Hepadnaviridae,* precisely the genus *Orthohepadnavirida* that is infectious to mammals [5]. Known hosts for human HBV, besides *Homo sapiens*, are *Pan troglodytes* (chimpanzee) [6] and *Tupaia belangeri* (northern tree shrew) [7].

3.1.2 Genomic organization and Proteins

The HBV rcDNA genome is around 3.2 kilobases (kb) long and its minus strand consists of four overlapping open reading frames (ORFs) – preS1/preS2/S, preC/C, P, X (Figure 1) [8], which encode for seven proteins. The preS1 ORF encodes for the longest HBV envelope protein (HBs) – L, followed by preS2 – M, and S. The three surface proteins have the same C-terminal domain that is being extended by the N-terminal resulting in the longer versions M and L. The longer product of the preC ORF encodes for the hepatitis B e protein (HBe), the shorter N-terminal C for the core protein (HBc). The other two ORFs P and X are encoding for the viral polymerase protein (P) and respectively for the X protein (X; HBx). HBs and HBc are the only structural proteins of the virus. Based on the dense overlapping structure and size of the HBV genome, every nucleotide has a coding function. Therefore, a point mutation might have an effect on more than one HBV protein.

The HBs proteins are located on the virions' lipid bilayer envelope surrounding the capsid which encapsulates the rcDNA (Figure 2). HBc forms into a T3 or T4 structure which consists of 180 or 240 HBc subunits. The HBV capsid is a dimer of that T3 or T4 structure

[9]. The polymerase, which functions as reverse transcriptase, RNaseH and primer during rcDNA synthesis [10], is covalently bound to the 5' end of the rcDNA's minus strand [11]. HBe is supposedly responsible for immune regulation [12], while HBx is initiating and maintaining the viral transcription [13].

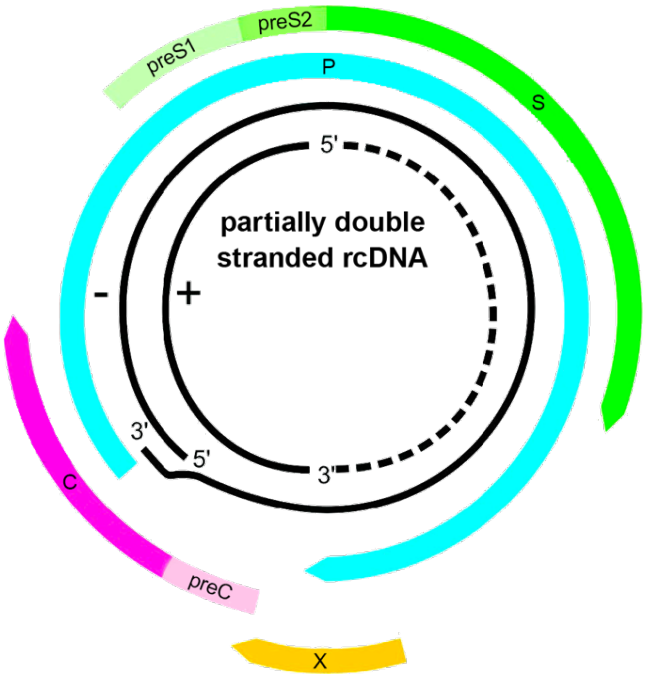

Figure 1: **HBV genomic structure.** The HBV genome is a partially double stranded rcDNA with four ORFs on its minus strand: preS1/preS2/S, preC/C, P and X. (adapted from Bell & Kramvis [14])

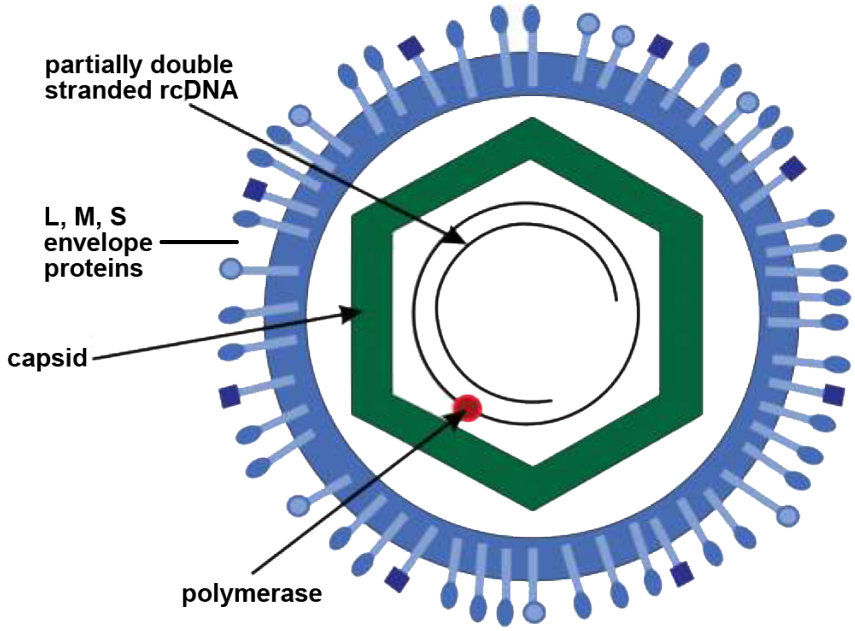

Figure 2: **HBV virion.** Inside the capsid the polymerase can be found that is attached to the rcDNA. The lipid bilayer envelope is covered with L, M, S proteins. (adapted from Bell & Kramvis [14])

3.1.3 Life Cycle

Figure 3 shows the viral entry of HBV into susceptible cells and its corresponding life cycle. Infectious HBV virions, termed “Dane particles” [15], first interact with heparan sulfate proteoglycan (HPSG) on the cell surface by attaching to their glycosaminoglycan side chains [16]. Next, they bind to the recently identified HBV entry receptor sodium-dependent taurocholate co-transporting polypeptide (NTCP) by an interaction with the preS1 domain of the L protein [17]. NTCP is expressed exclusively on hepatocytes and serves as an integral membrane glycoprotein for the uptake of bile acids [18]. In primary human hepatocytes, HBV uptake is mediated via endocytosis of the clathrin pathway [19]. In other cell types like HepaRG, this is not the case, as caveolin-1 plays the important role there [20]. After the uncoating of the HBV particle, the capsid is being transported to the nucleus. By attaching to the nuclear basket its dimer T structure is loosen, causing the rcDNA to be injected into the nucleus [21]. There, the rcDNA genome is repaired causing the establishment of stable, double stranded, covalently closed circular DNA (cccDNA) [22]. This persistent form of HBV resides in infected cells and is further used as the viral transcription template.

The pgRNA transcript is used as the template for the HBV polymerase and core production. Additionally, in the cytoplasm, it is reversely transcribed into rcDNA by the polymerase for the integration in *de novo* capsids [23]. Those newly formed capsids can follow two different routes, either translocation into the endoplasmatic reticulum or reshutteling into the nucleus and thereby increasing the cell’s cccDNA pool.

At the endoplasmatic reticulum capsids are enveloped with a lipid bilayer harboring the three HBs proteins before being secreted in the extracellular space [24]. The secretion is assisted by components of the multivesicular body machinery [25].

The precore RNA is translated into HBe, which is then secreted following the same pathway of the infectious particles [24].

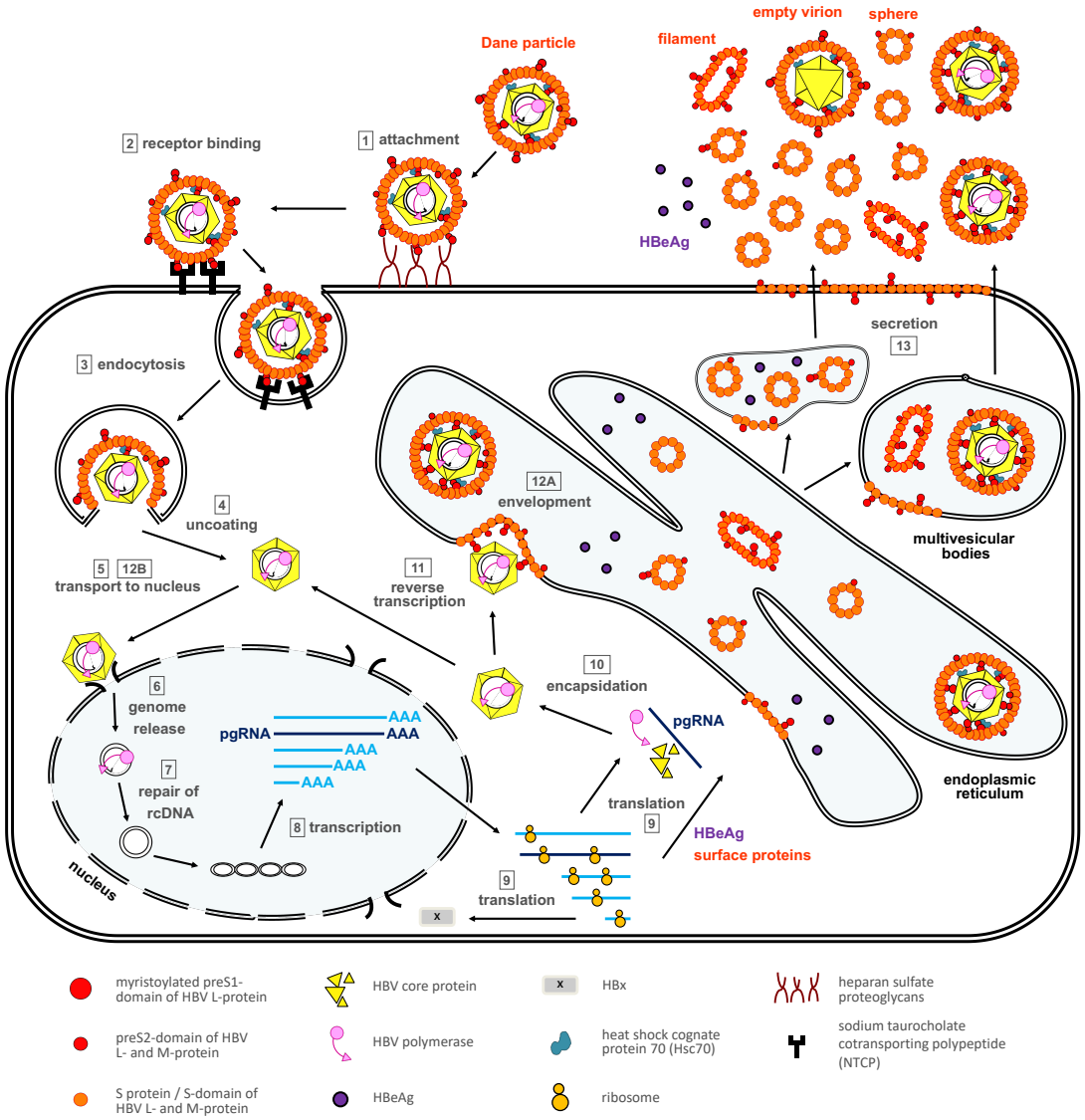

Figure 3: **HBV replication cycle.** After attaching to heparan sulphate proteoglycan, HBV binds to NTCP. It is then taken up via endocytosis and uncoated. The capsid is then attached to the nuclear basket of the nucleus and rcDNA is being released into it. After cellular repair mechanism are applied, cccDNA is being established that is then further used for the template for all viral transcripts. pgRNA is reverse transcribed and is encapsulated with P into new capsids that are later reshuttled into the nucleus or enveloped with HBs proteins in the endoplasmatic reticulum and released via multivesicular bodies. (Ko *et al*. [26]).

3.1.4 Epidemiology

HBV is either transmitted horizontally via body fluids during sexual intercourse, by blood or blood products, or vertically from mother to child during birth [27]. HBV infection in healthy adults often progresses asymptomatic. One third of the cases result in an acute hepatitis and 0.5 – 1% of these have an increased risk to develop a fulminant hepatitis that can result in liver failure. Only 5 – 10% of healthy adults develop a chronic HBV infection, that is defined by HBsAg being detectable for more than six months. These

rates are drastically increased in immune compromised patients (30-90%) and children up to the age of six months (90%) [28]. Thus, vertical transmission during birth is the major cause of chronic HBV in most countries [29].

Thereby, disease progression following HBV infection can differ, depending on. environmental factors such as alcohol consumption [30] and aflatoxin [31], key factors such as immune status, age of time of infection, human leukocyte antigen type (HLA) [32] but also ethnicity [33].

Viral factors which can influence the natural course of infection are the HBV titer / viral load during infection [34] and also the associated sero- and genotype [35]. Especially the genotypes are known to influence the primary transmission route, progression, treatment and mortality of HBV infection [36,37].

The clinical difference and importance of the genotypes have still to be shown in detail by future clinical trials, but data suggest that it might play a major role in transmission, decease progression, treatment and fatality. A more common vertical transmission can be observed with genotype B, C and I, while a horizontal transmission is more common with genotypes A, D and G [37-40]. A higher chronification rate can be seen with genotypes A and C [40]. Liver cirrhosis and hepatocellular carcinoma (HCC) are favored with genotypes C, D and F, while treatment of genotype A and B with interferon-a (IFNα) shows better results than with the other genotypes [40,41].

3.1.5 Serotypes

The HBV serotype describes the reactivity of antibodies towards HBsAg. The identification and categorization of HBV was initially done by serotyping, before genotypes were introduced [42]. The interaction with reference antibodies was used to determine the serotype. Four serotypes based on the antigenic determinants were initially used for classification: *adr, adw, ayr, ayw* [43]. Lately, those were expanded to nine: *ayw1, ayw2, ayw3, ayw4, ayr, adw2, adw4, adrq+, adrq-* [44].

The role of serotypes is relatively low in contemporary patient care. Though they should be still taken into account concerning vaccines efficacy, antibodies-based therapies and discovery by diagnostic tests.

3.1.6 Genotypes

HBV is classified into at least nine genotypes, from A – I, and there is additionally a supposed genotype ‘J’ [45] which was only found once in a patient [37]. The patient of Japanese descent was residing on Borneo Island and the phylogenetic analysis revealed that the origin of this variant was closer to gibbon HBV than the human one [46]. Since

further sequences are still missing, ‘J’ was not accepted as a new human HBV genotype. Genotype I was the last one added to the HBV genotypes. It was first discovered 2008 in a Vietnamese patient [47]. Nowadays it is even categorized into two sub-genotypes – I1 and I2 [37].

The nine genotypes are defined by a variation of at least 7.5% of the nucleotide sequence. Those are further categorized into sub-genotypes, which are defined by 4% difference of the nucleotides of their sequence [37,48]. The sub-genotypes are accounted for over 50. An exact number of sub-genotypes is not available, as per varying definitions of recombinants, low quality single sequences representing sub-genotypes and an ever-growing full genome sequence pool, literature tends to vary [37,49].

To allow discrimination of HBV genotypes different methods like non-sequencing-based (e.g., enzyme immune assays), partial genome sequencing (covering one or several proteins) and full genome sequencing were applied. Depending on the method type and the time of the categorization, the results vary not only in the accuracy of the sequence but also newer genotypes are falsely accounted for recombinants, a false genotype or worse – being undetected. This could be observed in the past like for genotype I, being the latest genotype added, which evolved as a recombinant variant from genotype A, C and G [47].

3.1.7 Treatments

Current treatment options for chronic HBV consist of pegylated IFNα and various nucleos(t)ide analogues. While interferon is known for its antiviral and immunomodulatory properties, it also has major side-effects like the induction of autoimmune diseases such as psoriasis, thyroiditis, systemic lupus erythematosus [50], but also depression and tiredness [51]. Lamivudine, telbivudine, adefovir, entecavir and tenofovir are nucleos(t)ides reducing the viral load by targeting the HBV reverse transcription from pgRNA to rcDNA [52]. Since nucleos(t)ide analogs are exclusively targeting viral transcripts, cccDNA remains unaffected hindering a complete cure [53].

In addition, such treatments can lead to the accumulation of mutations in the HBV genome. Lamivudine resistances are caused by M204V, M204I, L180M + M204V and combinations of those mutations. Telbivudine is not inhibited by the M204V mutation, but apart from that is akin affected as lamivudine. Entecavir has a reduced effectiveness to the mentioned single mutations, but copes with HBV resistance when the mutations occur in combination and with certain other point mutations. Adefovir resistance is caused especially by A181A/T and N236T, while tenofovir is so far the only one of the nucleos(t)ides not being completely undermined but those mutations [54].

3.1.8 Scope of Genotypes Study

Since vaccines and treatment options have been developed in the western countries, they are biased towards genotypes prevalent in those regions. A missing overview of the global distribution of genotypes and a low availability of data from poorer countries, who suffer the most from HBV, makes some important genotypes not considered for more effective treatments or vaccines. As there is no comprehensive data available regarding the frequency of infections worldwide for each genotype, this makes it difficult to process, predict and handle the disease burden. Knowing the genotype distribution and clinically relevant mutations associated with them, would improve preparation, planning and initialization of effective treatments. Therefore, a comprehensive literature review addressing incidence, genotypes and the respective country they have been associated to, was performed being now able to estimate the global genotypes distribution by merging all publicly available data into an overview.

3.1.9 Scope of Mutations Study

Mutations and their clinical relevance are well known and reviewed in several publications. However, there is no comprehensive overview on a global scale addressing them on the full genome of HBV. These studies are mostly limited to certain regions and based on clinical data from specific populations. Further the results mostly do not cover all available genotypes [55-68]. Understanding their impact on a global scale could contribute to a better understanding of the HBV phenotypes and burden and might help to establish improved diagnostics and antiviral therapies. Therefore, a computational analysis was performed on all HBV sequences that could be retrieved from publicly available databases.

3.2 Hepatitis D virus

An estimated 15 – 20 million humans are chronically infected with Hepatitis D virus (HDV) which causes the most severe form of viral hepatitis, often leading to liver cirrhosis and HCC [69]. Despite being relatively rare compared to HBV, in 70% of the cases a liver cirrhosis arises, which is three times more likely than for HBV [70]. HDV is a satellite virus of HBV and relays on the expression of HBs proteins to form new viral particles. Following this rational it solely occurs as a co-infection or superinfection with HBV [71]. Interestingly, both viruses do not share any similarity in terms of sequence.
Recently a functional block of the extracellular route was found, using Myrcludex B as an entry inhibitor [72].

3.2.1 Classification

HDV is a small (35-37 nm of diameter), enveloped, hepatotropic, circular negative sense single-stranded RNA virus [73]. It is the only member of the *Deltaviridae* genus.
Its lipid envelope incorporates the HBs proteins.

3.2.2 Genomic Organization and Proteins

The HDV genome is approximately 1.67 kb [74]. Its complementary strand contains only one ORF encoding for two isoforms of the hepatitis delta antigen (HDAg), S-HDAg and L-HDAg. L-HDAg is the larger protein which is 19 amino acids longer at its C-terminal side, compared to S-HDAg [75]. The switch from S-HDAg to L-HDAg is a result of the activity of the cellular enzyme adenosine deaminase (ADAR-1) that causes a change of the UAG stop codon of S-HDAg to UGG, resulting in a longer translation, becoming L-HDAg [76].
Both proteins contain domains for dimerization, nuclear-localization and RNA-binding. L-HDAg has an additional domain responsible for packaging [77], which makes it essential for that part, while S-HDAg is the crucial protein for HDV replication [78].

3.2.3 Genotypes

There are 8 known genotypes of HDV, termed 1 – 8. They genomic sequences differ at the lowest between HDV-5 and HDV-2 by 10%, and the highest 28.44% when comparing HDV-3 against all other genotypes [79]. Further analysis revealed sub-genotypes for HDV-1, HDV-2 and HDV-4. While sub-genotypes HDV-2a and HDV-2b, and HDV-4a and HDV-4b show a distinct geographic separation, HDV-1a – HDV-1e are more widely distributed and are evolving worldwide [79].

3.2.4 Identification of Genotypes

There is no publicly available database dedicated to HDV. The need for a fast, effective and reliable identification of HDV genotypes is growing, furthermore an overview of the available full genomes, coding sequences (CDS), proteins and even partial proteins is essential for a demanding scientific workflow. Assigning and comparing HDV sequences to countries for the evaluation of their origin and evolution is important for further studies. Therefore, HDVdb was established to provide a database for those tasks. Additionally, the newly established database offers the needed tools for on point establishment of alignments, primers and phylogenetic trees.

4 Materials and Methods

4.1 Software List

4.1.1 Data Processing

4.1.1.1 BLAST

Basic Local Alignment Search Tool (BLAST) [80] version 2.9.0+ was used for nucleotide and protein reference database generation and evaluation of genotypes and protein types of the acquired HBV sequences.

4.1.1.2 Clustal Omega

Clustal Omega [81] version 1.2.3 is used for the alignment of sequences in the HDVdb.

4.1.1.3 Excel

Microsoft Excel 2016 was used for several basic data arrangements and preparations for further visualizations. https://www.microsoft.com/de-de/microsoft-365/microsoft-office

4.1.1.4 EMBOSS

The European Molecular Biology Open Software Suite (EMBOSS) [82] was used for nucleotide to protein conversion of HBV sequences.

4.1.1.5 MUSCLE

MUSCLE [83] version 3.8.1551 was used for all alignments performed on HBV genomic and proteomic alignments.

4.1.1.6 Primer3

Primer3 [84] version 2.3.7 is used for primer design in HDVdb.

4.1.1.7 PhyML

PhyML [85] version 3.696 is used for maximum likelihood phylogenetic tree generation in HDVdb.

4.1.1.8 RAxML-NG

RAxML-NG [86] version 0.9.0 was used for phylogenetic tree generation of HBV sub-genotypes used as reference in the BLAST database.

4.1.1.9 Ruby

The custom scripts for evaluation of HBV und partially HDV data were written in Ruby 2.7. They were used for data aggregation, processing, cross-matching, evaluation and conversion of formats (i.e., GenBank to JSON, JSON to FASTA). https://www.ruby-lang.org

4.1.2 Visualization

4.1.2.1 FigTree

FigTree v1.4.4 was used for phylogenetic tree visualization of HBV and is integrated into the HDVdb workflow. http://tree.bio.ed.ac.uk/software/figtree/.

4.1.2.2 Powerpoint

Microsoft Powerpoint 2016 was used for several tables and figures arrangements. https://www.microsoft.com/de-de/microsoft-365/microsoft-office

4.1.2.3 Prism

Graphpad Prism 8.4.3 and later was used for all graphs shown in figures except the world maps. https://www.graphpad.com/scientific-software/ prism/

4.1.2.4 QGIS

Quantum Geographic Information System (QGIS) was used for generation of all world maps. http://www.qgis.org/en/site/

World Borders Dataset from Thematic Mapping was the data source for defining the country borders in the maps and their naming. http://thematicmapping.org/downloads/world_borders.php

The plugin MMQGIS was used to attach the evaluated country data to the corresponding countries. https://plugins.qgis.org/plugins/mmqgis/

4.1.3 Server-Side Applications & Services

4.1.3.1 Apache

The Apache HTTP Server Project is the daemon running the web services for the HDVdb. https://httpd.apache.org

4.1.3.2 Bash

Bourne-Again Shell (Bash) is used for basic file management on the HDVdb backend. https://www.gnu.org/software/bash/

4.1.3.3 Google Scholar

Google Scholar was used for literature search and research concerning HBV genotypes. https://scholar.google.de

4.1.3.4 Java

Java is used for the automatic annotation of the HDV genotypes. Additionally, it handles the queries and management of HDVdb. https://www.java.com/

4.1.3.5 Laravel

The Laravel framework is running on PHP and is responsible for the dynamic rendering of the HDVdb web page. It is the connection between Java, Bash and the utilized software packages BLAST, PhyML, Clustal Omega, Primer3 for the evaluation of the input data. https://laravel.com

4.1.3.6 NCBI

Nucleotide database of the National Center for Biotechnology Information (NCBI) [87] was used the retrieve all HBV sequences.

4.2 Data Aggregation for Estimation of the World-Wide HBV Genotype Distribution

4.2.1 HBV Prevalence

The prevalence data was extracted from Schweitzer *et al.* who had performed a systematic review on manuscripts describing epidemiological data on chronic HBV infections from 1st January 1965 until 23rd October 2013. The databases used by them for retrieval of publications were Medline, Embase, CAB Abstracts (Global health), Popline and Web of Science [28].

Schweitzer *et al.* screened 17,029 records for the incidence of HBV infection and provided from 1,800 reports the prevalence of HBs in 161 countries [28].

4.2.2 HBV Genotype Distribution

4.2.2.1 Identification of Records

Google Scholar was used for the acquirement of literature relevant to the genotypes. A search was performed on 23rd January 2018 with the following key words to acquire the

most relevant and available data: [Country Name] and [“HBV” or “Hepatitis B Virus”] and [“genotype”]

The country names used were those from the 161 countries that had a prevalence available from Schweitzer *et. al* [28]. For each country a separate search enquiry was performed. If no results were found, there were major cities addressed as a search option instead of the corresponding country. Additionally, references retrieved from the publications were screened when the genotyping information was not completely shown. Different languages used in the publications didn’t represent a major obstacle, as English, Spanish, French, German and Russian were readable and understandable. Missing manuscripts or limited access to such were replaced with corresponding data from other sources, and if not present, the information of their abstracts was used. The cutoff for the data used was chosen to be the year 2000, as more reliable results were expected with newer data**.**

4.2.2.2 Processing of Records

The total initial number of identified records through the search was 1650 hits. Additionally, 35 could be identified by reference screening. A filter was applied and records not reporting HBV genotyping data, using different than the mentioned languages, publication data before 2000 and duplicate records, were removed.

The remaining 913 articles were deeply evaluated and records with no primary data, exclusively testing non-representative populations like minorities and redundant data were excluded. 213 studies were left for the determination of the HBV genotype distribution.

4.2.2.3 Acquired Data

The obtained data for further evaluation was the year and type of publication, the year of the sample collection – if no information was available – two years before the publication was assumed, the date of analysis – if no information was available – one year before publication was assumed, location where the sample was collected, criteria for selecting the participants of the study, sex and age of the tested population, genotyping method, available samples and the identified genotype.

4.2.2.4 Establishment of a Scoring System

The acquired data needed to be addressed based on its origin. Especially this was the case concerning the provided genotypes. As most of the sequences used for its determination were not provided, a quality and weighting system needed to be applied.

The scoring system was developed after carefully analyzing all the acquired data to provide a scale of reliability and meaningfulness.

4.2.2.4.1 Study Quality Score

The first part of the scoring addressed the study itself by the study quality score. It consists of equally weighted genotyping and generalizability score. A better score means better quality. The genotyping score was rated based on the years of the analysis, giving a score of 1 for older articles, before 2010, and a higher one – 2, for the ones after that. This is based on the assumption that the technology and prices for sequencing and the tests used for identification have not only improved but generally have extended their scope of the genotypes being recognized, leading to the methodology score. The ability of the method used to identify the genotype was rated here. It was divided into 3 scores – 1 for non-sequencing methods, 3 for partially sequenced HBV genomes and 5 for whole viral genome sequencing. The non-sequencing-based methods bare a lower score as they have a high risk of misidentifying the right genotype as of their nature of not being able to detect all genotypes or recombinant viruses. These are probe-/PCR-/restriction fragment length polymorphisms and enzyme immunoassays. The whole genome sequencing is the 'gold standard' and therefore received the highest score – since it allows the distinction between known and *de novo* genotypes as well as the identification of their sub-genotypes. A median score was used for partially sequenced HBV genomes. That sequences are used for the identification of a protein coding ORF, which is cheaper and mostly effective, but can unfortunately lead to a false classification as it could be misinterpreted as a recombinant variant. Additionally, no sub-genotyping is possible as of the missing genetic information.

Both of those genotyping scores – years of samples analysis and ability of the method to recognize the correct genotype, are weighted with 30% / 70 % for that section's final score, which represents 50% of the final study score.

The generalizability score is split into three parts – representation of the country based on the location of the study, representation of the HBV infected population for the country and the year of sample collection. They are weighted respectively with 40%, 40% and 20%. The location of the study is used for samples allocation. The broader the spectrum of collection sites, the more accurate the data is representing that country. Single towns and regions have been rated with a score of 1, while several regions or nationwide studies have been rated with a score of 2. How representative the HBV-infected population is, is playing a major role. An important aspect is if only specific age groups or favored risk factors are covered– score of 1, or if studies were collecting samples from

non-predefined patients with an HBV infection– score of 2. Lastly, we are incorporating the year of sample collection into that equation. Allowing to bare samples available before the year 2000, those studies are scored with 1, between 2000 and 2010 – score of 2 and 2011 or later with 3. The rationale behind this is similar to the rating for the sample analysis year – better technology was available to take, store and transport that samples, causing them to be more reliable.

4.2.2.4.2 Country Quality Score

The second part of the scoring system is addressing the meaningfulness of the data for a country. The calculation was based on the weighted average of all studies in a country. The more samples a study provided, the higher was its contribution to the countries total score. To receive the final country quality score, the countries total score was multiplied by the score of the total amount of samples for that country. Countries with less than 100 samples received a score of 1, between 100 and 999 – score of 2 and 1000 or more – score of 3. To simplify the visualization, the country quality score was fitted to a 1 – 10 scale.

4.2.3 Estimation of HBV Genotypes Distribution

Population data was used to calculate the global HBV genotype distribution for the year 2017. The United Nations World Population Prospect from the year 2017 was used, as it was the most reliable and up to date data available at the time of the evaluation and manuscript preparation [88]. The available prevalence data from Schweitzer *et al.* [28] was multiplied with the genotype frequency acquired from the genotype distribution data from the 213 publications and resulted in genotyping results for 121 countries allowing to estimate the absolute number per genotype per country and word wide.

4.3 Data Aggregation of Clinically Relevant HBV Variants

4.3.1 Source of HBV Sequences

The analysis was performed on publicly available sequences. Those were retrieved from the Nucleotide database of the National Center for Biotechnology Information (NCBI) [87] on 1st February 2020. The selection criteria was the taxon ID 10407 that is representing HBV in the NCBI taxonomy browser. As the Nucleotide database is accessing GenBank, which is part of the International Nucleotide Sequence Database Collaboration consisting additionally of the DNA DataBank of Japan (DDBJ) and the European

Nucleotide Archive (ENA), that are exchanging data daily [89], the data received is the most up to date publicly available.
The taxon ID 10407 contains not only human HBV. This makes it possible to retrieve all sequences that might have been misclassified for certain reasons, reclassify them correctly and use them for the analysis. The output format in which the sequences were retrieved was the GenBank format [89].

4.3.2 Reference Sequences for Genotyping

4.3.2.1 Human HBV Sequences

Human HBV sequences were used for genotype references based on Pourkarim *et al.* [49]. Enumerated by accession number and sub-genotype:
JN182318: A1; HE576989: A2; AB194951: A3; AY934764: A4; FJ692613: A5; GQ331047: A6; FN545833: A7; AB642091: B1; FJ899779: B2; GQ924617: B3; GQ924626: B4; GQ924640: B5; JN792893: B6; GQ358137: B7; GQ358147: B8; GQ358149: B9; AB697490: C1; GQ358158: C2; DQ089801: C3; HM011493: C4; EU410080: C5; EU670263: C6; GU721029: C7; AP011106: C8; AP011108: C9; AB540583: C10; AB554019: C11; AB554025: C12; AB644280: C13; AB644284: C14; AB644286: C15; AB644287: C16; GU456636: D1; GQ477452: D2; EU594434: D3; GQ922003: D4; GQ205377: D5; KF170740: D6; FJ904442: D7; FN594770: D8; JN664942: D9; FN594748: E; FJ709464: F1b; DQ899146: F2b; AY090459: F1a; DQ899142: F2a; AB036920: F3; AF223965: F4; GU563556: G; AB516393: H; FJ023659: I1; FJ023664: I2 ; AB486012: J.

4.3.2.2 Non-Human HBV Sequences

Non-human HBV reference sequences were manually chosen and evaluated to be not matching to any human HBV genotype and represent internally the major non-human HBV sub-branches of the taxon ID 10407. As the naming of the non-human genotypes was not of interest, they were named arbitrary. Enumerated by accession number and non-human user defined genotype naming:
K02715: GSHV; U29144: ASHV; AF193864: OGHBV; AJ251935: STHBV; AY226578: WMHBV; AY628097: WCHBV; JQ664503: GOHBV; JQ664509: CHHBV; KY962705: BATHBV; KC790373: BATHBV1; KC790374: BATHBV2; KC790375: BATHBV3; KC790376: BATHBV4; KC790377: BATHBV5; KC790378: BATHBV6; KC790379: BATHBV7; KC790380: BATHBV8; KC790381: BATHBV9; AB823662: GIHBV; KT893897: SGIHBV; KT345708: STHBV2; KY703886: CMHBV; MF471768: DHBV.

4.3.2.3 BLAST Database Creation

The HBV genotype references were converted to the FASTA format [90]. A reference nucleotide database for BLAST [80] was created based on the reference genomes provided above with the default *makeblastdb* arguments for *dbtype nucl*.

4.3.3 Processing of Sequences

The processing of the sequences was done in one step using a custom Ruby script.
It evaluated at once the following steps using the GenBank formatted data that was retrieved from NCBI and stored it into the JSON format (https://www.ecma-international.org/publications-and-standards/standards/ecma-404/) with custom data. The data included all relevant information of the processing, concerning the future purpose those were the most relevant entries: accession number, size of the sequence, year of submission and sample collection, city / country / region / subregion, quality score, genotype, unverified and nonfunctional, host.
Preprocessing was performed by checking the starting position of each sequence. In case it was not the EcoRI cutting side, it was corrected by iterating over the whole genome to locate it.

4.3.3.1 Quality Check of Sequences

Sequences were checked for their integrity. All meta data entries in the GenBank file of the sequence were scanned for the *unverified* and *non-functional* value and sequences where at least one of both triggers could be found were discarded from the future procession. Sequences including unclear nuclear calling marked by *n* were excluded too - this was reflected in the custom quality score.

4.3.3.2 Regional Allocation of Sequences

The GenBank format includes the *FEATURES* entry where the submitter can provide further details about the sequence. One of those keys is called *country*. This key was used to perform regional allocations of the sequences by grouping countries into regions. This data was then stored in the city / country / region / subregion of our JSON. For future procession only countries were used with available sequences.

4.3.3.3 Regions Definition

Enumeration of regions and their assigned countries based on the available sequences: *Eastern Africa*: Ethiopia, Kenya, Madagascar, Malawi, Mauritius, Rwanda, Somalia, Tanzania - United Republic of, Uganda, Zimbabwe; *Middle Africa*: Angola, Cameroon, Central African Republic, Congo - The Democratic Republic Of, Gabon; Northern Africa: Egypt, Sudan, Tunisia; *Southern Africa*: Botswana, Namibia, South Africa *Western Africa:* Benin, Burkina Faso, Cape Verde, Gambia, Ghana, Guinea, Liberia, Mali, Niger, Nigeria; *Caribbean*: Cuba, Haiti, Martinique; *Central America*: Costa Rica, El Salvador, Mexico, Nicaragua, Panama; *Northern America*: Canada, Greenland, United States; *South America*: Argentina, Bolivia - Plurinational State of, Brazil, Chile, Colombia, Peru, Uruguay, Venezuela - Bolivarian Republic of; *Central Asia:* Kazakhstan, Tajikistan, Uzbekistan; Eastern Asia: China, Hong Kong, Japan, Korea - Republic of, Mongolia, Taiwan - Republic of China; *South-Eastern Asia:* Cambodia, Indonesia, Lao People's Democratic Republic, Malaysia, Myanmar, Philippines, Thailand, Vietnam; *Southern Asia:* Bangladesh, India, Iran - Islamic Republic Of, Nepal, Pakistan; *Western Asia:* Saudi Arabia, Syrian Arab Republic, Turkey, United Arab Emirates; *Eastern Europe:* Belarus, Poland, Russian Federation; *Northern Europe:* Estonia, Ireland, Latvia, Sweden, United Kingdom; *Southern Europe:* Italy, Serbia, Spain; *Western Europe:* Belgium, France, Germany, Netherlands; *Oceania:* Australia, New Zealand, Fiji, New Caledonia, Papua New Guinea, Vanuatu, Kiribati, Samoa, Tonga.

4.3.3.4 Genotyping

Genotyping was performed using the established BLAST database. The sequence's sub-genotype retrieved is based on the highest blast score. It was stored in the corresponding genotype field and the host was applied for non-human HBV sequencies, where possible. For the future processing only sequences with human genotypes were used.

4.3.3.5 Constructs

The initial preprocessing fixes the wrongly positioned sequences but doesn't address the size of the construct. After evaluating the initial results, the methodology of filtering for minimum and maximum length was applied. Minimum length was set to 3150, maximum to 3275. This was based on the known genotype specific full genome lengths [37].

4.3.4 Translation into HBV Proteins

The corresponding HBV proteins were retrieved by translating the nucleotide sequences. First a BLAST database was established using the longest protein of each ORF as a reference – preS1, preC, P and X, with the default *makeblastdb* arguments for *dbtype prot*.

There was no differentiation in-between the genotypes done and proteins from genotype A were used. This was doable as only the ambiguous proteins generated by computational translation by EMBOSS sixpack [82] needed to be removed and additionally the genotypes were already assigned based on the nucleotide sequence. Sixpack was used with the options *orfminsize 100* and *mstart*, which resulted in proteins with a minimum length of 100 amino acids (aa) for each ORF starting with a methionine. The input sequence was processed as twice the same sequence concatenated in respect of the circular genome. Further the resulting proteins were filtered by size, to obtain only valid results: ≥330 aa for preS1/preS2/S, ≥140 aa for preCore/core, ≥700 aa for polymerase, and ≥100 aa for X.

4.3.5 Unifying

The different HBV genomes of the genotypes vary in size and additionally insertions or deletions may have occurred in each of their sequences. Unifying the length of the genomic sequences is therefore essential to address the same positions, especially in-between of genotypes. This was performed after several rounds of alignment with MUSCLE [83] and manual adjusting after evaluation of the results. First this was performed genotype wise and afterwards globally over all genotypes. The resulting size for each genomic sequence was extended to 3257 bp. The adjustment was performed by adding gaps in the appropriate places. The following sub-chapters deal with positions, it is notable, that per definition, the base and starting positions are beginning their count by zero and not one.

4.3.5.1 Serotyping and Unifying

Serotyping of the genomic sequences was performed by evaluating the HBsAg on the already described amino acid positions by Norder *et al.* [91]. Positions 122, 127, 134, 159, 160, 177 and 178 were used where applicable. They were adjusted by unifying to be applicable to every genotype, based on the L surface protein's position after alignment.

L surface protein base: 122 = 299, 127 = 304, 134 = 311, 159 = 336, 160 = 337, 177 = 355, 178 = 356. Added to L protein base position depending on genotype: A no

modification; B positions 122, 127, 134, 159, 160 -1 and 177, 178 -2 positions; C positions 122, 127, 134, 159, 160, +33 and 177, 178 +32 positions; D positions 122, 127, 134, 159, 160, +29 and 177, 178 +28 positions; E positions 122, 127, 134, 159, 160, +8 and 177, 178 +7 positions; F positions 122, 127, 134, 159, 160, +1 position and 177, 178 unmodified; G positions 122, 127, 134, 159, 160, -2 and 177, 178 -3 positions; H and I positions 122, 127, 134, 159, 160, -3 and 177, 178 -4 positions; J positions 122, 127, 134, 159, 160, -14 and 177, 178 -15 positions.

D: -11 amino acids for each position; E and G: -1 amino acid for each position. Additional positions adjusted to the L protein genotypes were 122, 127, 134, 160. For genotypes C and H +45; D -11; E and G -1 amino acids.

Not corresponding to the following definitions of the serotypes were classified as undefined:

ayw1: R122, P127, F134,160K; R122, P127, A159, 160K; ayw2: R122, P127, 160K; ayw3: R122, T127, K160; ayw4: R122, 127L, 160K; ayr: R122, R160; adw2: K122, P127, K160; adw3: K122, T127, K160; adw4q-: K122, L127, K169, Q178; adrq+: K122, R160, V177, P178; adrq-: K122, R160, A177.

Amino acids: A = Alanine, F = Phenylalanine, K = Lysine, L = Leucine, P = Proline, Q = Glutamine, R = Arginine, T = Threonine, V = Valine.

4.3.5.2 Mutations und Unifying

The evaluation of the mutations required a unifying of the HBsAg, reverse transcriptase (RT) region of the P protein and the basal core promoter (BCP) genomic region until close after the epsilon stem loop. The kind and position of the mutations were retrieved from Lazarevic *et al.* [92].

The following starting position per genotype on the L surface protein for the HBsAg region was defined: A 177, B 176, C 204, D 203, E 185, F 174, G 175, H 174, I 174, J 163. The evaluated mutations were located following the mentioned starting positions at: 100, 101, 105, 115, 116, 119, 120, 122, 123, 124, 126, 127, 129, 133, 134, 136, 139, 140, 141, 142, 143, 144, 145, 167, 169, 174, 175, 177, 182. Those needed further adjustments for the genotypes as follows: A, B, E, G, H, I, J no adjustments; C except positions 100, 101, 105 all others +6; D except positions 100, 101, 105 all others +3; F except positions 100, 101, 105 all others +4.

The following starting position per genotype on the P protein for the RT region was defined: A 353, B 355, C 390, D 385, E 345, F 346, G 348, H 346, I 346, J 335. The evaluated mutations were located following the mentioned starting positions at: 80, 169, 173, 177, 180, 181, 184, 194, 202, 204, 215, 233, 236, 249, 250. Those needed further

adjustments for the genotypes as follows: B, E, G, H, I, J no adjustments; A position 80 unmodified, positions 233, 236, 249, 250 +2, all other +1; C except position 80 all others +10; D except position 80 all others +3; F except position 80 all others +4.
The following starting position per genotype on the genomic sequence for the BCP region was defined: A 21, B 161, C 114, D 80, E 0, F 17, G 0, H 7, I 1, J 0. The evaluated mutations were located following the mentioned starting positions at: 1653, 1753, 1762, 1764, 1766, 1768, 1862, 1896, 1899. Those needed further adjustments for the genotypes as follows: G, J no adjustments; A position 1768 +3, positions 1862, 1896, 1899 +21, all others unmodified; B positions 1653, 1753 unmodified, positions 1762, 1764 +3, position 1766 +5, position 1768 +8, position 1862 +26, positions 1896, 1899 +27; C position 1653 unmodified, position 1753 +1, positions 1762, 1764 +2, position 1766 +4, position 1768 +7, position 1862 +78, positions 1896, 1899 +79; D positions 1862, 1896, 1899 +126, all others unmodified; E position 1653 unmodified, position 1753 +9, positions 1762, 1764, 1766, 1768 +13, positions 1862, 1896, 1899 +14; F position 1653 unmodified, all others +13; H position 1653 unmodified, positions 1753, 1762, 1764, 1766, 1768 +3, position 1862 +6, positions 1896, 1899 +7; I position 1766 +2, positions 1768, 1862, 1896, 1899 +5, all others unmodified.

4.3.6 Assigning of Clinically Relevant Mutations

Processing the data respectively following the unifying methods for each sequence, two separate processes were run. One sorting the results based on genotypes and one based on regions. The data generated showed in both cases the amino acids for each of the positions known to include a clinically relevant mutation. After being exported to Excel, the results were manual revised and the frequency of occurrence of each mutation was calculated. Amino acids differing from the expected normal state, but not shown to be clinically relevant, were not addressed.

4.4 HDVdb

4.4.1 Source of HDV Sequences

The construction of the database was performed with publicly available sequences. Those were retrieved from NCBI [87]. The selection criteria was the taxon ID 12475 that is representing HDV in the NCBI taxonomy browser. Like already described for HBV, this data is descending from GenBank and is covering all publicly available sequences. At the time of the 24th May 2020, there were 2621 hepatitis delta virus nucleotide sequences deposited. Those were covering full and partial genomic sequences coding for L-HDAg and S-HDAg. Sequences shorter than 90 bp were discarded from the

dataset. 152 sequences with missing genotype data were assigned one by using BLAST and a reference dataset shown in Karimzadeh *et al.* [79].

4.4.2 Structure

BLAST databases were established containing all retrieved and processed sequences sorted by type – proteins and nucleotides. The sequences were named by country of origin and genotype. Queries via a Laravel framework are evaluating the user's input in FASTA format and are processing it via Java and Bash on the server side. The available services are similarity search, primer design, multiple sequence alignments (MSA) and phylogenetic analysis with tree visualization. The BLAST databases are exclusively used for the similarity search and are providing a genotype prediction based on BLAST scores. The three other services are not exclusively to HDV, but are providing a convenient way for evaluation any data using the provided server.

5 Results

5.1 The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data

Velkov, S., *et al.* (2018). "The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data." *Genes* **9**(10).

5.1.1 Authors

Stoyan Velkov, Jördis J. Ott, Ulrike Protzer and Thomas Michler

5.1.2 Short Summary and Contributions

The study constitutes a comprehensive literature review, allowing to perform the first approximation of the worldwide distribution of HBV genotypes. By combining published genotyping data and merging it with available data on HBV prevalence around the world, we added the missing link for the global burden assigned to the individual genotypes. Additionally, the absolute number of infections with each genotype per country, world region or globally, could be estimated after taking into account WHO population data. As there were vast differences regarding the quality and amount of data for the different countries, a scoring system was implemented. The score allowed to evaluate and obtain an overview of the quality of available data for each country. Low numbers of sequencing data from certain regions potentially caused inaccuracy of the determined genotype distribution in these areas and need to be addressed further.
The resulting data is a valuable addon for future HBV studies related not only to genotypes, but also towards human migration and viral co-migration.

S.V. contributed to**:** conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing of review and editing.

5.2 Global Occurrence of Clinically Relevant Hepatitis B Virus Variants as Found by Analysis of Publicly Available Sequencing Data

Velkov, S., *et al*. (2020). "Global Occurrence of Clinically Relevant Hepatitis B Virus Variants as Found by Analysis of Publicly Available Sequencing Data." *Viruses* **12**(11).

5.2.1 Authors

Stoyan Velkov, Ulrike Protzer and Thomas Michler

5.2.2 Short Summary and Contributions

HBV infection has a diverse profile of disease progression. Studies are coping with different mutations and their outcomes. Associating genotypes to those outcomes makes it possible to address the clinical implications by customizing therapies and diagnostic tests, addressing the expected viral evolution. Accumulation of mutations in distinct genotypes might demand for different treatment options in future and should ease the choice of an applicable method. Unifying the data emerged the possibility to compare between genotypes and, especially on the genomic level, makes them easier to be analyzed. Considering the region dependent distribution, one can observe a trend that high income countries, in which a high number of chronically infected patients are receiving antiviral therapy, show a higher frequency of resistance mutations. Furthermore, regions are disclosed where some of those therapeutics are still usable for treatment due to the low occurrence of resistance mutations.

Severity of disease progression in populations can potentially be estimated not only based on the genotype, but also on the region of acquirement. Those approaches should still be considered with care, as the data evaluation is based on publicly available sequences, where the source and possible bias is mostly not addressed.

S.V. contributed to: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft preparation and visualization.

5.3 HDVdb: A Comprehensive Hepatitis D Virus Database

Usman, Z., Velkov, S., *et al.* (2020). "HDVdb: A Comprehensive Hepatitis D Virus Database." *Viruses* **12**(5).

5.3.1 Authors

Zainab Usman †, **Stoyan Velkov** †, Ulrike Protzer, Michael Roggendorf, Dmitrij Frishman and Hadi Karimzadeh

† These authors contributed equally to this work.

5.3.2 Short Summary and Contributions

Considering the burden of HDV, a designated database was in demand. The evolving studies on the HDV genotypes and their mostly distinct geographical allocations are an important and urging point that needs to be addressed.

A comprehensive coverage of the publicly available genomic and protein sequences is provided by HDVdb. Further, those with missing genotype information were defined. HDVdb is aiming to become the online spot for scientific data evaluation related to HDV by giving its users the possibility to do all needed analysis in one place. Its fast, simple and user-friendly structure, backed up by a powerful server backbone, eases performing tasks.

S.V. contributed to: implementation of software services and writing the manuscript.

6 Discussion

6.1 Global Distribution of HBV Genotypes

Information about the global distribution of HBV genotypes is scarce. The aim of this study was to close this gap. Evaluation of the studies reporting HBV genotyping data showed that there are strong variations in the number and quality of available studies for the different world regions. 96% of the global chronic HBV infections can be accounted to the genotypes A – E, only 2% of the infections can be related to the genotype F – I.

There is a strong discrepancy in the amount of available genotyping data for countries. It is not exclusive to countries with a lower income, but also the western countries are partially missing sufficient data.

All regions contain countries with good to poor quality data, and some are without any data at all. It is advisable to stress this finding and in future aim to improve the data situation in all countries. This can be achieved by introducing a better HBV surveillance and overall perform more genotyping studies in regions and their respective countries.

The results suggest that geographic boundaries like oceans or the Sahara Desert play an important role but also ethnicity can be related to certain genotypes. In Sub-Saharan Africa almost exclusively found are genotype A and E, while in Northern Africa and the enclosed regions – West Asia and Southern Europe, genotype D [93].

Regions with high migration background show the corresponding genotypes of their descendants, like Northern America with populations from Europe and Asia where genotype A, B, C and D are mostly persistent, and the Caribbean with genotype A and D descending probably from the Sub-Saharan and Northern Africa. The global burden of the genotypes is therefore reflected in high populated regions and endemicity including distinct genotypes, like B and C being predominant in Southeast and East Asia.

The score implemented for evaluation of the trustworthiness of the genotype data underlines the limitations and problems of analyzing different sources of data. Genotyping applied only on partial sequences, non-sequencing-based methods and genotyping references at the time of sample classification, may had led to wrong classification. The best example is genotype I that was firstly defined in 2010, before that timepoint it might have been classified as a genotype A, C or G sequence or a recombinant of those [94]. The reason for that is not only the limited usage of full genome sequencing, that is a matter of cost, but relying on references for genotyping that are outdated as of today's classification. Additionally, the methods which did not sequence the virus, but identify genotypes via probes, have evolved over time and didn't cover all available genotypes at the time they were applied to the samples.

The extrapolation of a study population to a whole country's population bares additional uncertainty factors. E.g., the analyzed sequences might have derived from non-representative populations, minorities or cities and regions not reflecting the ethnicity or habitat of the whole country. The data shows that only one third of the studies were performed nationwide and mostly big countries like the United States, Brazil and Russia are not represented in a satisfactory way.

Distinct genotypes could had been favored in 40% of studies. The performed genotyping in studies dedicated to patients with advanced decease progression resulting in fibrosis, cirrhosis and HCC are expected to be selective for genotypes that are known to inherit a greater risk for these outcomes.

The exclusion of studies focusing only on minorities or foreigners to the country, as they are not representative for the country they immigrate to, as well as small samples size, bare the risk of missing rare genotypes, which might make them underrepresented in the results. Not including those patients at all in representative studies ads up to the bias.

Sample size was addressed as an important bias to the results. Countries having only a small sample size or even only single studies, could have led to an inaccurate estimation of the genotype distribution. To overcome this, where possible and available, samples were pooled from all studies for the respective country. Never the less, samples size was identified as the crucial factor and therefore the scoring system was primary based on it.

The different time points from data collection, HBsAg prevalence and population data taken together, introduce additional uncertainty. Although age is an important factor considering the chronicity and disease course caused by HBV, missing data made it impossible to weight the data according to age.

The estimation covered genotyping data for 125 countries, being representative for 96% of the world's population in 2015. The missing 4% are due to the lack of studies covering the remaining countries. It can't be excluded that studies were missed during the process of literature evaluation; however, the expected impact should be relatively small. The estimation of the worldwide genotyping distribution was based on 26,000 samples. That represents only 0.01% of the chronic HBV infections, which compromises a high risk of uncertainties, and describes the actual data situation.

Never the less, our study presents a broader up-to-date insight on the genotypes, their distribution, contribution and importance, especially for certain regions of the world. It specifies the global burden of chronic HBV infection and identifies countries where further and deeper diagnostics need to be performed. Especially genotype E seems to be undervalued in today's experimental models, even it represents an estimate of 17,6%

of the cases. A broader focus for drug discovery and therapies should be applied to genotypes A – E, as they represent over 96% of the worldwide genotypes.

6.2 Global Occurrence of Clinically Relevant HBV Mutations

The disease progression of chronic hepatitis B varies depending on the HBV genotype, the patient's immune condition and environmental factors. The study analyzed the publicly available full-length HBV sequences to present an overview of the clinically relevant HBV mutations based on their frequency within geographic regions and the individual HBV genotypes. Notably the results should not be over interpreted and used only as a hint, as there might be several internal biases. The sequences available might tend to represent more virus variants which presented with interesting phenotypes, and had therefore been sequenced. Additionally, the relevance of certain mutations might vary across genotypes, and it needs to be proven that they indeed carry an increased risk for a certain phenotype, as they have only been partially observed on selected genotypes. The available sequences don't represent the estimated worldwide distribution of the genotypes and their regions. This leads to different confidence in the statements about their occurrence and significance. For genotypes A – D, between 1004 and 2700 sequences could be obtained. While this finding is in line with the wide distribution of these genotypes, genotype E was underrepresented related to its significance [1]. It appears in the list of the lower represented ones, E – I, with sequences counting from 44 to 312. The reasons for this fluctuation, which corelates to the high- and low-income regions, is presumably related to higher costs for full genome sequencing in comparison to partial viral sequencing or antigenic binding tests. This corelates with the data, as in genotypes for which less sequences were available, less mutations were found. The same is true when comparing the available sequences per regions – as regions with lower available sequences tended to have less mutations found.

The quality of the sequences and the method used to derive them is another big factor contributing to the results. Low quality sequences – which were already marked as such by meta data using the keywords *undefined* or *non-functional*, and including at least one undefined base marked by *n*, were excluded. Never the less it can't be excluded that the good quality samples are biased in some way, including sequencing, analyzing and publishing. The sequencing method itself poses another variable. Sanger sequencing shows only predominant virus variants in the sample, that need to be above a threshold of 20% [95]. Next-generation sequencing (NGS) produces a better resolution, but it faces the challenge of corelating the short reads to their original sequence. Longer NGS reads,

covering the whole HBV genome, would be the best applicable method, that has been seldom used, as that technology had only recently become available [96]. An additional bias might be artefacts caused by any method used. Therefore, mutations with a low frequency should be handled with care in regard of their clinical relevance.

The serotype prediction was performed *in-silico*, however there is only limited experimental evidence of reactivity of expected antibodies towards HBs of included sequences. There is a lower bias expected by increased sequencing of certain serotypes, as there is no direct association to a disease outcome. Never the less, distinct genotypes are associated with specific serotypes, as already shown in literature [97]. The global distribution shows that *adw2* and *adrq+* each contribute to around 1/3rd of all cases, while the other 7 serotypes account together for the remaining 1/3rd. The importance of serotypes for HBV has been undermined by the definition of genotypes, still they might play an important role in establishment of therapeutic vaccines and antibodies.

The data suggests that P127H/L, that is associated with occult HBV infection, is the most common mutation in the HBsAg. Interestingly that mutation represents the wild-type for genotype E, F and H, and is also one of the determinants for the serotype definition. In Oceania, the V177A mutation is present above average, despite being most evenly distributed among the genotypes. That position is also used in the serotype definition, and is partially encoding for *adrq*. The diagnosis of occult hepatitis, for which this mutation is known, might be related not only to a biological reason, like ethnicity or selection pressure in that region, but to diagnostic tests deficiency in recognizing it. R122P, another suggested mutation for occult hepatitis, couldn't be identified in more than 7000 sequences, which questions its clinical significance.

The most common mutations of the reverse transcriptase are M204V/I and L180M. Both are resistance mutations against lamivudine, telbivudine and entecavir treatment. The geographic regions with high prevalence of these mutations are mostly high-income areas with broad access to antiviral therapies such as Europe, Northern America, Eastern and Southern Asia. In contrast, in low-income regions, like Africa, lower frequencies of those mutations were found. This might implicate that the resistances occurred because of selection pressure, based on broader treatments, in the respective regions.

P177G and F249A have not been found in any of the screened sequences. According to the literature [98], those variants were created *in vitro,* inducing a tenofovir resistance. Based on the results they do not seem to be clinically relevant, probably as the caused resistance was outweighed by fitness losses for HBV.

The mutations on the HBV genome for the basal core promotor and pre-core region of genotype G were the most frequent ones. The G1896A mutation, which causes hepatocellular carcinoma, HBeAg-negativity and fulminant hepatitis, could be observed in 100% of the 89 sequences, and should be considered as wild-type for genotype G. Additionally, C1653, T1753C, A1762T, G1764A, which can also contribute to the disease progression, were found in more than 93.3% of the sequences.

The HBeAg-negativity for genotype G is well described in literature [99,100]. Genotype G is mostly found as a co-infection with genotype A [101] and therefore no reliable data concerning hepatocellular carcinoma and fulminant hepatitis, describing outcomes of infections with only genotype G, is available. A distinct geographical enrichment can be found in the Caribbean with 60% of sequences baring the G1862T mutation, and Western Asia and Southern Europe with 55 – 66% of sequences harboring the G189A mutation. Considering the relative low count of sequences for those regions, Caribbean – 80, Western Asia – 146, Southern Europe – 69, further studies need to be conducted to confirm the results.

Overall, the study presents an overview over the frequency of their clinically relevant mutations. However, the results should be handled with care considering the limited availability of sequences and the fact that this was an *in-situ* evaluation that needs further *in vitro* or *in vivo* testing to be confirmed. Still, the results give a hint about conducting future studies for improving the diagnostics and therapies.

6.3 HDVdb

The emerging insights into HDV over the past years, in part due to the introduction of new technologies such as NGS, revealed many interesting findings on HDV biology [72]. As a result, the focus of the research community was shifted to the role of HDV genotypes and demanded for a dedicated database that is able to provide and define the genotypes for the acquired sequences. HDVdb was initiated with 512 complete genome sequences, 1066 L-HDAg and S-HDAg protein sequences, and 1281 partially coding sequences.

The backbone provides all the data in FASTA format and is therefore the most convenient way to retrieve all available genomic or proteomic sequences, in bulk, for a genotype or protein. A smart solution is the definition of the genotypes, as it relies not on single reference sequences, but uses the whole dataset as reference. In that manner, the genotyping describes additionally the closest relative to the query sequence and the country of origin. Comparative sequence analysis can be performed using various bioinformatic services provided to the user.

An important element of HDVdb are the provided updates, they allow to react to new genotype definitions and provide up to date data, not only by adding new sequences, but expanding and revising their classification. The webservices and the dynamic structure allows to expand the analysis tools selection, if needed, and keep those up to date when new functions arise. All together HDVdb is a webpage covering the comprehensive demands for HDV bioinformatic analysis.

7 References

1. Velkov, S.; Ott, J.J.; Protzer, U.; Michler, T. The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. *Genes (Basel)* **2018**, *9*, doi:10.3390/genes9100495.
2. WHO. Global Hepatitis Report. Availabe online: https://apps.who.int/iris/bitstream/handle/10665/255016/9789241565455-eng.pdf?sequence=1 (accessed on 01.02.2020).
3. Meireles, L.C.; Marinho, R.T.; Van Damme, P. Three decades of hepatitis B control with vaccination. *World J Hepatol* **2015**, *7*, 2127-2132, doi:10.4254/wjh.v7.i18.2127.
4. Ott, J.J.; Stevens, G.A.; Groeger, J.; Wiersma, S.T. Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. *Vaccine* **2012**, *30*, 2212-2219, doi:10.1016/j.vaccine.2011.12.116.
5. Schaefer, S. Hepatitis B virus taxonomy and hepatitis B virus genotypes. *World J Gastroenterol* **2007**, *13*, 14-21, doi:10.3748/wjg.v13.i1.14.
6. Barker, L.F.; Maynard, J.E.; Purcell, R.H.; Hoofnagle, J.H.; Berquist, K.R.; London, W.T.; Gerety, R.J.; Krushak, D.H. Hepatitis B Virus Infection in Chimpanzees: Titration of Subtypes. *The Journal of Infectious Diseases* **1975**, *132*, 451-458, doi:10.1093/infdis/132.4.451.
7. Walter, E.; Keist, R.; Niederöst, B.; Pult, I.; Blum, H.E. Hepatitis B virus infection of tupaia hepatocytes in vitro and in vivo. *Hepatology* **1996**, *24*, 1-5, doi:https://doi.org/10.1002/hep.510240101.
8. Tiollais, P.; Pourcel, C.; Dejean, A. The hepatitis B virus. *Nature* **1985**, *317*, 489-495, doi:10.1038/317489a0.
9. Crowther, R.A.; Kiselev, N.A.; Böttcher, B.; Berriman, J.A.; Borisova, G.P.; Ose, V.; Pumpens, P. Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell* **1994**, *77*, 943-950, doi:10.1016/0092-8674(94)90142-2.
10. Seeger, C.; Mason, W.S. Molecular biology of hepatitis B virus infection. *Virology* **2015**, *479-480*, 672-686, doi:10.1016/j.virol.2015.02.031.
11. Nassal, M. HBV cccDNA: viral persistence reservoir and key obstacle for a cure of chronic hepatitis B. *Gut* **2015**, *64*, 1972, doi:10.1136/gutjnl-2015-309809.
12. Chen, M.; Sällberg, M.; Hughes, J.; Jones, J.; Guidotti, L.G.; Chisari, F.V.; Billaud, J.N.; Milich, D.R. Immune tolerance split between hepatitis B virus precore and core proteins. *J Virol* **2005**, *79*, 3016-3027, doi:10.1128/jvi.79.5.3016-3027.2005.
13. Lucifora, J.; Arzberger, S.; Durantel, D.; Belloni, L.; Strubin, M.; Levrero, M.; Zoulim, F.; Hantz, O.; Protzer, U. Hepatitis B virus X protein is essential to initiate and maintain virus replication after infection. *Journal of hepatology* **2011**, *55*, 996-1003, doi:10.1016/j.jhep.2011.02.015.
14. Bell, T.G.; Kramvis, A. The Study of Hepatitis B Virus Using Bioinformatics. In *Bioinformatics - Updated Features and Applications*, InTech: 2016; 10.5772/63076.
15. Dane, D.S.; Cameron, C.H.; Briggs, M. Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. *Lancet* **1970**, *1*, 695-698, doi:10.1016/s0140-6736(70)90926-8.
16. Schulze, A.; Gripon, P.; Urban, S. Hepatitis B virus infection initiates with a large surface protein-dependent binding to heparan sulfate proteoglycans. *Hepatology* **2007**, *46*, 1759-1768, doi:10.1002/hep.21896.
17. Yan, H.; Zhong, G.; Xu, G.; He, W.; Jing, Z.; Gao, Z.; Huang, Y.; Qi, Y.; Peng, B.; Wang, H., et al. Sodium taurocholate cotransporting polypeptide is a functional

receptor for human hepatitis B and D virus. *eLife* **2012**, *1*, doi:10.7554/eLife.00049.
18. Appelman, M.D.; Chakraborty, A.; Protzer, U.; McKeating, J.A.; van de Graaf, S.F.J. N-Glycosylation of the Na+-Taurocholate Cotransporting Polypeptide (NTCP) Determines Its Trafficking and Stability and Is Required for Hepatitis B Virus Infection. *PLOS ONE* **2017**, *12*, e0170419, doi:10.1371/journal.pone.0170419.
19. Huang, H.C.; Chen, C.C.; Chang, W.C.; Tao, M.H.; Huang, C. Entry of hepatitis B virus into immortalized human primary hepatocytes by clathrin-dependent endocytosis. *J Virol* **2012**, *86*, 9443-9453, doi:10.1128/jvi.00873-12.
20. Macovei, A.; Radulescu, C.; Lazar, C.; Petrescu, S.; Durantel, D.; Dwek, R.A.; Zitzmann, N.; Nichita, N.B. Hepatitis B virus requires intact caveolin-1 function for productive infection in HepaRG cells. *J Virol* **2010**, *84*, 243-253, doi:10.1128/jvi.01207-09.
21. Schmitz, A.; Schwarz, A.; Foss, M.; Zhou, L.; Rabe, B.; Hoellenriegel, J.; Stoeber, M.; Panté, N.; Kann, M. Nucleoporin 153 arrests the nuclear import of hepatitis B virus capsids in the nuclear basket. *PLoS pathogens* **2010**, *6*, e1000741, doi:10.1371/journal.ppat.1000741.
22. Schreiner, S.; Nassal, M. A Role for the Host DNA Damage Response in Hepatitis B Virus cccDNA Formation-and Beyond? *Viruses* **2017**, *9*, doi:10.3390/v9050125.
23. Sohn, J.A.; Litwin, S.; Seeger, C. Mechanism for CCC DNA Synthesis in Hepadnaviruses. *PLOS ONE* **2009**, *4*, e8093, doi:10.1371/journal.pone.0008093.
24. Nassal, M. HBV cccDNA: viral persistence reservoir and key obstacle for a cure of chronic hepatitis B. *Gut* **2015**, *64*, 1972-1984, doi:10.1136/gutjnl-2015-309809.
25. Jiang, B.; Himmelsbach, K.; Ren, H.; Boller, K.; Hildt, E. Subviral Hepatitis B Virus Filaments, like Infectious Viral Particles, Are Released via Multivesicular Bodies. *J Virol* **2015**, *90*, 3330-3341, doi:10.1128/jvi.03109-15.
26. Ko, C.; Michler, T.; Protzer, U. Novel viral and host targets to cure hepatitis B. *Curr Opin Virol* **2017**, *24*, 38-45, doi:10.1016/j.coviro.2017.03.019.
27. Stevens, C.E.; Toy, P.; Kamili, S.; Taylor, P.E.; Tong, M.J.; Xia, G.L.; Vyas, G.N. Eradicating hepatitis B virus: The critical role of preventing perinatal transmission. *Biologicals* **2017**, *50*, 3-19, doi:10.1016/j.biologicals.2017.08.008.
28. Schweitzer, A.; Horn, J.; Mikolajczyk, R.T.; Krause, G.; Ott, J.J. Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *The Lancet* **2015**, *386*, 1546-1555, doi:10.1016/s0140-6736(15)61412-x.
29. Lok, A.S.; McMahon, B.J. Chronic hepatitis B: update 2009. *Hepatology* **2009**, *50*, 661-662, doi:10.1002/hep.23190.
30. Ganesan, M.; Eikenberry, A.; Poluektova, L.Y.; Kharbanda, K.K.; Osna, N.A. Role of alcohol in pathogenesis of hepatitis B virus infection. *World journal of gastroenterology* **2020**, *26*, 883-903, doi:10.3748/wjg.v26.i9.883.
31. Chu, Y.J.; Yang, H.I.; Wu, H.C.; Liu, J.; Wang, L.Y.; Lu, S.N.; Lee, M.H.; Jen, C.L.; You, S.L.; Santella, R.M., et al. Aflatoxin B(1) exposure increases the risk of cirrhosis and hepatocellular carcinoma in chronic hepatitis B virus carriers. *Int J Cancer* **2017**, *141*, 711-720, doi:10.1002/ijc.30782.
32. Wang, L.; Zou, Z.-Q.; Wang, K. Clinical Relevance of HLA Gene Variants in HBV Infection. *Journal of Immunology Research* **2016**, *2016*, 9069375, doi:10.1155/2016/9069375.
33. Koc, O.M.; Robaeys, G.; Yildirim, B.; Posthouwer, D.; Hens, N.; Koek, G.H. The influence of ethnicity on disease outcome in patients with chronic hepatitis B infection. *J Med Virol* **2019**, *91*, 623-629, doi:10.1002/jmv.25353.

34. Asabe, S.; Wieland, S.F.; Chattopadhyay, P.K.; Roederer, M.; Engle, R.E.; Purcell, R.H.; Chisari, F.V. The size of the viral inoculum contributes to the outcome of hepatitis B virus infection. *Journal of virology* **2009**, *83*, 9652-9662, doi:10.1128/JVI.00867-09.
35. Revill, P.A.; Tu, T.; Netter, H.J.; Yuen, L.K.W.; Locarnini, S.A.; Littlejohn, M. The evolution and clinical impact of hepatitis B virus genome diversity. *Nature Reviews Gastroenterology & Hepatology* **2020**, 10.1038/s41575-020-0296-6, doi:10.1038/s41575-020-0296-6.
36. Wong, G.L.-H.; Chan, H.L.-Y.; Yiu, K.K.-L.; Lai, J.W.-Y.; Chan, V.K.-K.; Cheung, K.K.-C.; Wong, E.W.-N.; Wong, V.W.-S. Meta-analysis: the association of hepatitis B virus genotypes and hepatocellular carcinoma. *Alimentary Pharmacology & Therapeutics* **2013**, *37*, 517-526, doi:10.1111/apt.12207.
37. Kramvis, A. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **2014**, *57*, 141-150, doi:10.1159/000360947.
38. Komatsu, H.; Inui, A.; Fujisawa, T.; Takano, T.; Tajiri, H.; Murakami, J.; Suzuki, M. Transmission route and genotype of chronic hepatitis B virus infection in children in Japan between 1976 and 2010: A retrospective, multicenter study. *Hepatol Res* **2015**, *45*, 629-637, doi:10.1111/hepr.12396.
39. Krekulova, L.; Rehak, V.; da Silva Filho, H.P.; Zavoral, M.; Riley, L.W. Genotypic distribution of hepatitis B virus in the Czech Republic: a possible association with modes of transmission and clinical outcome. *Eur J Gastroenterol Hepatol* **2003**, *15*, 1183-1188, doi:10.1097/00042737-200311000-00006.
40. Liu, C.J.; Kao, J.H. Global perspective on the natural history of chronic hepatitis B: role of hepatitis B virus genotypes A to J. *Semin Liver Dis* **2013**, *33*, 97-102, doi:10.1055/s-0033-1345716.
41. Lin, C.L.; Kao, J.H. Hepatitis B virus genotypes and variants. *Cold Spring Harb Perspect Med* **2015**, *5*, a021436, doi:10.1101/cshperspect.a021436.
42. Okamoto, H.; Tsuda, F.; Sakugawa, H.; Sastrosoewignjo, R.I.; Imai, M.; Miyakawa, Y.; Mayumi, M. Typing Hepatitis B Virus by Homology in Nucleotide Sequence: Comparison of Surface Antigen Subtypes. *Journal of General Virology* **1988**, *69*, 2575-2583, doi:https://doi.org/10.1099/0022-1317-69-10-2575.
43. Le Bouvier, G.L.; McCollum, R.W.; Hierholzer, W.J., Jr.; Irwin, G.R.; Krugman, S.; Giles, J.P. Subtypes of Australia Antigen and Hepatitis-B Virus. *JAMA* **1972**, *222*, 928-930, doi:10.1001/jama.1972.03210080020005.
44. Couroucé-Pauty, A.M.; Lemaire, J.M.; Roux, J.F. New hepatitis B surface antigen subtypes inside the ad category. *Vox Sang* **1978**, *35*, 304-308, doi:10.1111/j.1423-0410.1978.tb02939.x.
45. Tatematsu, K.; Tanaka, Y.; Kurbanov, F.; Sugauchi, F.; Mano, S.; Maeshiro, T.; Nakayoshi, T.; Wakuta, M.; Miyakawa, Y.; Mizokami, M. 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*, 10538-10547, doi:10.1128/jvi.00462-09.
46. Locarnini, S.; Littlejohn, M.; Aziz, M.N.; Yuen, L. Possible origins and evolution of the hepatitis B virus (HBV). In Proceedings of Seminars in cancer biology; pp. 561-575.
47. Tran, T.T.; Trinh, T.N.; Abe, K. New complex recombinant genotype of hepatitis B virus identified in Vietnam. *J Virol* **2008**, *82*, 5657-5663, doi:10.1128/jvi.02556-07.
48. Norder, H.; Couroucé, A.M.; Coursaget, P.; Echevarria, J.M.; Lee, S.D.; Mushahwar, I.K.; Robertson, B.H.; Locarnini, S.; Magnius, L.O. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology* **2004**, *47*, 289-309, doi:10.1159/000080872.

49. Pourkarim, M.R.; Amini-Bavil-Olyaee, S.; Kurbanov, F.; Van Ranst, M.; Tacke, F. Molecular identification of hepatitis B virus genotypes/subgenotypes: revised classification hurdles and updated resolutions. *World J Gastroenterol* **2014**, *20*, 7152-7168, doi:10.3748/wjg.v20.i23.7152.
50. Cacopardo, B.; Benanti, F.; Pinzone, M.R.; Nunnari, G. Rheumatoid arthritis following PEG-interferon-alfa-2a plus ribavirin treatment for chronic hepatitis C: a case report and review of the literature. *BMC Research Notes* **2013**, *6*, 437, doi:10.1186/1756-0500-6-437.
51. Zoulim, F.; Lebossé, F.; Levrero, M. Current treatments for chronic hepatitis B virus infections. *Curr Opin Virol* **2016**, *18*, 109-116, doi:10.1016/j.coviro.2016.06.004.
52. Rajbhandari, R.; Chung, R.T. Treatment of Hepatitis B: A Concise Review. *Clin Transl Gastroenterol* **2016**, *7*, e190-e190, doi:10.1038/ctg.2016.46.
53. Boni, C.; Barili, V.; Acerbi, G.; Rossi, M.; Vecchi, A.; Laccabue, D.; Penna, A.; Missale, G.; Ferrari, C.; Fisicaro, P. HBV Immune-Therapy: From Molecular Mechanisms to Clinical Applications. *International journal of molecular sciences* **2019**, *20*, 2754, doi:10.3390/ijms20112754.
54. Caligiuri, P.; Cerruti, R.; Icardi, G.; Bruzzone, B. Overview of hepatitis B virus mutations and their implications in the management of infection. *World journal of gastroenterology* **2016**, *22*, 145-154, doi:10.3748/wjg.v22.i1.145.
55. Bahar, M.; Pervez, M.T.; Ali, A.; Babar, M.E. In Silico Analysis of Hepatitis B Virus Genotype D Subgenotype D1 Circulating in Pakistan, China, and India. *Evolutionary Bioinformatics* **2019**, *15*, 1176934319861337, doi:10.1177/1176934319861337.
56. Colagrossi, L.; Hermans, L.E.; Salpini, R.; Di Carlo, D.; Pas, S.D.; Alvarez, M.; Ben-Ari, Z.; Boland, G.; Bruzzone, B.; Coppola, N., et al. Immune-escape mutations and stop-codons in HBsAg develop in a large proportion of patients with chronic HBV infection exposed to anti-HBV drugs in Europe. *BMC infectious diseases* **2018**, *18*, 251-251, doi:10.1186/s12879-018-3161-2.
57. Gao, S.; Duan, Z.-P.; Coffin, C.S. Clinical relevance of hepatitis B virus variants. *World journal of hepatology* **2015**, *7*, 1086-1096, doi:10.4254/wjh.v7.i8.1086.
58. Gupta, N.; Goyal, M.; Wu, C.H.; Wu, G.Y. The Molecular and Structural Basis of HBV-resistance to Nucleos(t)ide Analogs. *J Clin Transl Hepatol* **2014**, *2*, 202-211, doi:10.14218/JCTH.2014.00021.
59. Hao, R.; Xiang, K.; Shi, Y.; Zhao, D.; Tian, H.; Xu, B.; Zhu, Y.; Dong, H.; Ding, H.; Zhuang, H., et al. Naturally Occurring Mutations within HBV Surface Promoter II Sequences Affect Transcription Activity, HBsAg and HBV DNA Levels in HBeAg-Positive Chronic Hepatitis B Patients. *Viruses* **2019**, *11*, doi:10.3390/v11010078.
60. Ismail, A.M.; Sharma, O.P.; Kumar, M.S.; Kannangai, R.; Abraham, P. Impact of rtI233V mutation in hepatitis B virus polymerase protein and adefovir efficacy: Homology modeling and molecular docking studies. *Bioinformation* **2013**, *9*, 121-125, doi:10.6026/97320630009121.
61. Meier-Stephenson, V.; Bremner, W.T.R.; Dalton, C.S.; van Marle, G.; Coffin, C.S.; Patel, T.R. Comprehensive Analysis of Hepatitis B Virus Promoter Region Mutations. *Viruses* **2018**, *10*, doi:10.3390/v10110603.
62. Neumann-Fraune, M.; Beggel, B.; Pfister, H.; Kaiser, R.; Verheyen, J. High frequency of complex mutational patterns in lamivudine resistant hepatitis B virus isolates. *Journal of Medical Virology* **2013**, *85*, 775-779, doi:10.1002/jmv.23530.
63. Pacheco, S.R.; Dos Santos, M.I.M.A.; Stocker, A.; Zarife, M.A.S.A.; Schinoni, M.I.; Paraná, R.; Dos Reis, M.G.; Silva, L.K. Genotyping of HBV and tracking of resistance mutations in treatment-naïve patients with chronic hepatitis B. *Infect Drug Resist* **2017**, *10*, 201-207, doi:10.2147/IDR.S135420.

64. Phan, N.M.H.; Faddy, H.; Flower, R.; Spann, K.; Roulis, E. In silico Analysis of Genetic Diversity of Human Hepatitis B Virus in Southeast Asia, Australia and New Zealand. *Viruses* **2020**, *12*, doi:10.3390/v12040427.
65. Tong, S.; Revill, P. Overview of hepatitis B viral replication and genetic variability. *Journal of hepatology* **2016**, *64*, S4-S16.
66. Wagner, J.; Yuen, L.; Littlejohn, M.; Sozzi, V.; Jackson, K.; Suri, V.; Tan, S.; Feierbach, B.; Gaggar, A.; Marcellin, P., et al. Analysis of Hepatitis B virus haplotype diversity detects striking sequence conservation across genotypes and chronic disease phase. *Hepatology* **2020**, *n/a*, doi:10.1002/hep.31516.
67. Yano, Y.; Azuma, T.; Hayashi, Y. Variations and mutations in the hepatitis B virus genome and their associations with clinical characteristics. *World journal of hepatology* **2015**, *7*, 583-592, doi:10.4254/wjh.v7.i3.583.
68. Zhang, X.; Chen, X.; Wei, M.; Zhang, C.; Xu, T.; Liu, L.; Xu, Z. Potential resistant mutations within HBV reverse transcriptase sequences in nucleos(t)ide analogues-experienced patients with hepatitis B virus infection. *Scientific reports* **2019**, *9*, 8078-8078, doi:10.1038/s41598-019-44604-6.
69. Chen, H.Y.; Shen, D.T.; Ji, D.Z.; Han, P.C.; Zhang, W.M.; Ma, J.F.; Chen, W.S.; Goyal, H.; Pan, S.; Xu, H.G. Prevalence and burden of hepatitis D virus infection in the global population: a systematic review and meta-analysis. *Gut* **2019**, *68*, 512-521, doi:10.1136/gutjnl-2018-316601.
70. Farci, P.; Niro, G.A. Clinical features of hepatitis D. *Semin Liver Dis* **2012**, *32*, 228-236, doi:10.1055/s-0032-1323628.
71. Noureddin, M.; Gish, R. Hepatitis delta: epidemiology, diagnosis and management 36 years after discovery. *Curr Gastroenterol Rep* **2014**, *16*, 365, doi:10.1007/s11894-013-0365-x.
72. Zhang, Z.; Urban, S. Interplay between Hepatitis D Virus and the Interferon Response. *Viruses* **2020**, *12*, 1334.
73. Lai, M.M. The molecular biology of hepatitis delta virus. *Annu Rev Biochem* **1995**, *64*, 259-286, doi:10.1146/annurev.bi.64.070195.001355.
74. Wang, K.S.; Choo, Q.L.; Weiner, A.J.; Ou, J.H.; Najarian, R.C.; Thayer, R.M.; Mullenbach, G.T.; Denniston, K.J.; Gerin, J.L.; Houghton, M. Structure, sequence and expression of the hepatitis delta (delta) viral genome. *Nature* **1986**, *323*, 508-514, doi:10.1038/323508a0.
75. Chao, M.; Hsieh, S.Y.; Taylor, J. Role of two forms of hepatitis delta virus antigen: evidence for a mechanism of self-limiting genome replication. *J Virol* **1990**, *64*, 5066-5069, doi:10.1128/jvi.64.10.5066-5069.1990.
76. Casey, J.L. RNA editing in hepatitis delta virus. *Curr Top Microbiol Immunol* **2006**, *307*, 67-89, doi:10.1007/3-540-29802-9_4.
77. Moraleda, G.; Dingle, K.; Biswas, P.; Chang, J.; Zuccola, H.; Hogle, J.; Taylor, J. Interactions between hepatitis delta virus proteins. *J Virol* **2000**, *74*, 5509-5515, doi:10.1128/jvi.74.12.5509-5515.2000.
78. Taylor, J.M. Hepatitis D Virus Replication. *Cold Spring Harb Perspect Med* **2015**, *5*, doi:10.1101/cshperspect.a021568.
79. Karimzadeh, H.; Usman, Z.; Frishman, D.; Roggendorf, M. Genetic diversity of hepatitis D virus genotype-1 in Europe allows classification into subtypes. *Journal of Viral Hepatitis* **2019**, *26*, 900-910, doi:https://doi.org/10.1111/jvh.13086.
80. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **1990**, *215*, 403-410, doi:10.1016/S0022-2836(05)80360-2.
81. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **2011**, *7*, 539, doi:10.1038/msb.2011.75.

82. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **2000**, *16*, 276-277, doi:10.1016/s0168-9525(00)02024-2.
83. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**, *32*, 1792-1797, doi:10.1093/nar/gkh340.
84. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3--new capabilities and interfaces. *Nucleic Acids Res* **2012**, *40*, e115, doi:10.1093/nar/gks596.
85. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **2010**, *59*, 307-321, doi:10.1093/sysbio/syq010.
86. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **2019**, *35*, 4453-4455, doi:10.1093/bioinformatics/btz305.
87. Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **2016**, *44*, D7-D19, doi:10.1093/nar/gkv1290.
88. United Nations, D.o.E.a.S.A.P.D. World Population Prospects: The 2017 Revision. 2017.
89. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res* **2013**, *41*, D36-42, doi:10.1093/nar/gks1195.
90. Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435-1441, doi:10.1126/science.2983426.
91. Norder, H.; Courouce, A.M.; Magnius, L.O. Molecular basis of hepatitis B virus serotype variations within the four major subtypes. *J Gen Virol* **1992**, *73 (Pt 12)*, 3141-3145, doi:10.1099/0022-1317-73-12-3141.
92. Lazarevic, I. Clinical implications of hepatitis B virus mutations: recent advances. *World J Gastroenterol* **2014**, *20*, 7653-7664, doi:10.3748/wjg.v20.i24.7653.
93. Reed, F.A.; Tishkoff, S.A. African human diversity, origins and migrations. *Curr Opin Genet Dev* **2006**, *16*, 597-605, doi:10.1016/j.gde.2006.10.008.
94. Ismail, A.M.; Goel, A.; Kannangai, R.; Abraham, P. Further evidence of hepatitis B virus genotype I circulation in Northeast India. *Infect Genet Evol* **2013**, *18*, 60-65, doi:10.1016/j.meegid.2013.04.033.
95. Fu, Y.; Zeng, Y.; Chen, T.; Chen, H.; Lin, N.; Lin, J.; Liu, X.; Huang, E.; Wu, S.; Wu, S., et al. Characterization and Clinical Significance of Natural Variability in Hepatitis B Virus Reverse Transcriptase in Treatment-Naive Chinese Patients by Sanger Sequencing and Next-Generation Sequencing. *Journal of Clinical Microbiology* **2019**, *57*, e00119-00119, doi:doi:10.1128/JCM.00119-19.
96. McNaughton, A.L.; Roberts, H.E.; Bonsall, D.; de Cesare, M.; Mokaya, J.; Lumley, S.F.; Golubchik, T.; Piazza, P.; Martin, J.B.; de Lara, C., et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Scientific Reports* **2019**, *9*, 7081, doi:10.1038/s41598-019-43524-9.
97. Liu, C.J.; Kao, J.H.; Chen, P.J.; Lai, M.Y.; Chen, D.S. Molecular epidemiology of hepatitis B viral serotypes and genotypes in taiwan. *Journal of Biomedical Science* **2002**, *9*, 166-170, doi:10.1007/BF02256028.
98. Chan, S.L.; Wong, V.W.S.; Qin, S.; Chan, H.L.Y. Infection and Cancer: The Case of Hepatitis B. *Journal of Clinical Oncology* **2016**, *34*, 83-90, doi:10.1200/jco.2015.61.5724.
99. Stuyver, L.; De Gendt, S.; Van Geyt, C.; Zoulim, F.; Fried, M.; Schinazi, R.F.; Rossau, R. A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. *J Gen Virol* **2000**, *81*, 67-74, doi:10.1099/0022-1317-81-1-67.

100. Zaaijer, H.L.; Boot, H.J.; van Swieten, P.; Koppelman, M.H.G.M.; Cuypers, H.T.M. HBsAg-negative mono-infection with hepatitis B virus genotype G. *Journal of Viral Hepatitis* **2011**, *18*, 815-819, doi:https://doi.org/10.1111/j.1365-2893.2010.01397.x.
101. Osiowy, C.; Gordon, D.; Borlang, J.; Giles, E.; Villeneuve, J.P. Hepatitis B virus genotype G epidemiology and co-infection with genotype A in Canada. *J Gen Virol* **2008**, *89*, 3009-3015, doi:10.1099/vir.0.2008/005124-0.

Appendix – Reprint Permissions

MDPI journals (*Genes, Viruses*)

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.

List of Publications

Velkov, S.; Ott, J.J.; Protzer, U.; Michler, T. The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. *Genes (Basel)* **2018**, *9*, doi:10.3390/genes9100495.

Velkov, S.; Protzer, U.; Michler, T. Global Occurrence of Clinically Relevant Hepatitis B Virus Variants as Found by Analysis of Publicly Available Sequencing Data. *Viruses* **2020**, *12*, 1344.

Usman, Z.*; **Velkov, S.***; Protzer, U.; Roggendorf, M.; Frishman, D.; Karimzadeh, H. HDVdb: A Comprehensive Hepatitis D Virus Database. *Viruses* **2020**, *12*, doi:10.3390/v12050538.

Afridi, S.Q.; Usman, Z.; Donakonda, S.; Wettengel, J.M.; **Velkov, S**.; Beck, R.; Gerhard, M.; Knolle, P.; Frishman, D.; Protzer, U., et al. Prolonged norovirus infections correlate to quasispecies evolution resulting in structural changes of surface-exposed epitopes. *iScience* **2021**, *24*.

* These authors contributed equally to this work.

Acknowledgements

First of all, I want to thank Prof. Dr. Ulrike Protzer for giving me the opportunity to work on various demanding and fascinating topics resulting in several papers and furthermore broadening my knowledge in various virological aspects. Her expertise and the surrounding discussions were always of high value to me and extremely helpful for undergoing the next steps of my research and thesis.

I want to thank Prof. Dr. Dmitrij Frishman, my second advisor, for valuable feedback related to bioinformatics procedures, data approach and evaluation of the results.

Dr. Fabiana Perocchi, my mentor, was of great help concerning my species studies related to HBV. She provided me with the tools and information on how to address and evaluate big data, looking for susceptibility factors.

Dr. Thomas Michler played an important role not only concerning my thesis, but my overall career development. He was in charge of me during my master thesis and further was always an essential part of my thesis studies. He was very supportive to me and had always time to help and support me. I'm lucky to have found such a great friend in him.

Dr. Andreas Oswald became more than a colleague to me, discussing not several, but close to all topics related to work and additionally evolving to a very good friend. He had always time and knowledge to support and help me, no matter which project I was addressing.

Dr. Anindita Chakraborty, Martin Kächele, Samuel Jeske and Till Bunse were always there for me and we spent a lot of time together. Working on projects, problems and solutions. They provided me always with support, in the lab and especially in everyday life.

I had a lot of great colleagues which I'm thankful for, and hereby I want to say a thank you to all of the lab members that have crossed my way.

I'm most thankful to my wonderful wife for being always there for me, before, during, and after the thesis. I'm very lucky having her support in everything, without her this thesis would have never been finished, probably ;) A big thank you goes out to my lovely son, providing me with joy and motivation every single day! Last but not least, without my parents supporting me from day one, nothing of this would have ever been possible. Thank you for always having trust in me!

genes

MDPI

Article

The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data

Stoyan Velkov [1], Jördis J. Ott [2], Ulrike Protzer [1,3] and Thomas Michler [1,3,*]

1 Institute of Virology, Technische Universität München/Helmholtz Zentrum München, Trogerstrasse 30, D-81675 München, Germany; stoyan.velkov@tum.de (S.V.); protzer@tum.de (U.P.)
2 Department of Epidemiology, Helmholtz Centre for Infection Research, D-38124 Braunschweig, Germany; Hannover Medical School, D-30625 Hannover, Germany; Joerdis.Ott@helmholtz-hzi.de
3 German Center for Infection Research (DZIF), Munich Partner Site
* Correspondence: thomas.michler@tum.de; Tel.: +49-(0)89-4140-6814

Received: 7 September 2018; Accepted: 10 October 2018; Published: 15 October 2018

Abstract: Hepatitis B virus (HBV) is divided into nine genotypes, A to I. Currently, it remains unclear how the individual genotypes contribute to the estimated 250 million chronic HBV infections. We performed a literature search on HBV genotyping data throughout the world. Over 900 publications were assessed and data were extracted from 213 records covering 125 countries. Using previously published HBV prevalence, and population data, we approximated the number of infections with each HBV genotype per country and the genotype distribution among global chronic HBV infections. We estimated that 96% of chronic HBV infections worldwide are caused by five of the nine genotypes: genotype C is most common (26%), followed by genotype D (22%), E (18%), A (17%) and B (14%). Genotypes F to I together cause less than 2% of global chronic HBV infections. Our work provides an up-to-date analysis of global HBV genotyping data and an initial approach to estimate how genotypes contribute to the global burden of chronic HBV infection. Results highlight the need to provide HBV cell culture and animal models that cover at least genotypes A to E and represent the vast majority of global HBV infections to test novel treatment strategies.

Keywords: Hepatitis B virus; chronic hepatitis B; genotype; sequencing; molecular epidemiology

1. Introduction

An estimated 250 million humans [1] are chronically infected with hepatitis B virus (HBV), causing an estimated 887,000 annual deaths, mostly due to the long-term sequelae liver cirrhosis and hepatocellular carcinoma (HCC) [2]. The viral population can be divided into nine genotypes (A to I) [3,4] which differ in more than 7.5% of their nucleotide sequences [3,4] and which are further subdivided into subgenotypes with a nucleotide divergence greater than 4% [3,4]. While genotypes A to H have long been accepted as individual genotypes, two new genotypes (I and J) were proposed more recently [5,6]. Genotype I was first described in 2008 after isolation from a Vietnamese patient and constitutes a recombination of genotypes A, C and G [5]. Since the nucleotide divergence, especially compared to genotype C, is relatively small, it was long debated if this strain should be considered a new genotype [7]. Finally, the identification of similar HBV strains in Laos, North India and China in isolated native populations, indicating that the virus strains have circulated in these populations for a longer time, led to acceptance as an independent genotype [8–10]. Another strain, previously proposed to be a new genotype "J", was isolated from a Japanese patient who had lived on Borneo island [6]. Phylogenetic analysis revealed that the strain rather resembles gibbon than human HBV, and may result from recombinations with human genotype C [11]. Since further reports of human infections with this strain are lacking, it has so far not been recognized as a relevant HBV genotype [11].

Sequencing and genotyping of HBV isolates is not routinely done and rarely reported in, e.g., epidemiological studies. Hepatitis B virus genotypes, however, vary in their clinical consequences including the natural course of infection, disease progression and treatment response (reviewed in [3,12,13]): While genotypes B, C and I are associated with a more frequent vertical transmission from mother to child, a higher transmission rate during sexual contact or injecting drug use has been reported for genotypes A, D and G [3,14–16]. A higher chronification rate after infections with genotypes A and C, compared to genotypes B and D, has also been reported [14], but may also be due to the transmission route. Among chronic HBV carriers, a lower rate of seroconversion to HBV-e-antigen antibodies (anti-HBe) was proposed in genotype C and D infections [14]. Also, a faster disease progressions to liver cirrhosis and HCC are associated with infections with genotypes C, D and F [14]. While all genotypes similarly respond to treatment with reverse transcriptase inhibitors, under interferon-α treatment, genotypes A and B show an increased virological response and higher anti-HBe seroconversion than other HBV genotypes [14,17].

The separation of the HBV population into different genotypes can be dated back 30 million years ago, when the ancestors of modern *Homo sapiens* dispersed across Africa and Eurasia [18]. The distinct appearances of genotypes and subgenotypes in certain geographical regions and ethnic groups [3,4,12] allow for the confirmation of prehistoric human migrations by the transfer of HBV genotypes and subtypes from one continent to another [3,19]. While previous studies described differences between local genotype appearances, the number of infections with each HBV genotype and the genotype distribution within global chronic HBV infections have not been studied. Given the specific characteristics of each HBV genotype, this information could, however, be valuable for precising the HBV disease burden and informing health policy, for example, by means of predicting and estimating prevention and treatment needs and relevance of specific treatments in a country. A further adaptation and rational application of diagnostic tests, as well as the development of new broadly applicable therapies, may also be stimulated by scoping HBV genotypes worldwide. We approached this need and performed a literature search for HBV genotyping studies and applied previously published HBV surface antigen (HBsAg) prevalence estimates [1] and United Nations (UN) population data [20] to approximate the number of HBV infections by each genotype per country, world region and globally.

2. Materials and Methods

2.1. Literature Search for Hepatitis B Virus Genotyping Data

A literature search was performed up to 23rd January 2018 using Google Scholar and the search terms: [Country Name] and ["HBV" or "Hepatitis B Virus"] and ["genotype"]. Countries entered were these for which HBsAg prevalence data were reported by Schweitzer et al. [1]. If no publications were found for a certain country, the search was repeated with names of major cities within the respective country instead of the country name. Additional publications were added after manually screening references in identified records. Publications were subjected to a two-step review. First, titles and abstracts in English, Spanish, French, German or Russian languages were screened with regard to the provision of HBV genotyping results. If a manuscript was not accessible, data were extracted from its abstract if possible or taken from a secondary publication. Publications before the year 2000 and duplicate records were excluded (Figure 1).

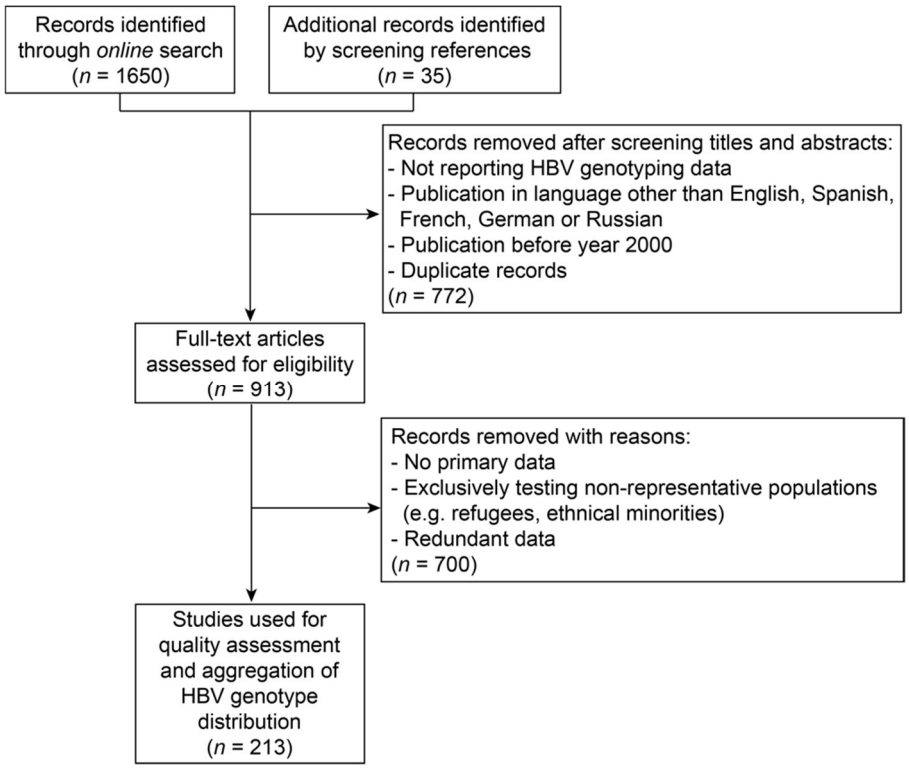

Figure 1. Flow chart on selection of records describing primary hepatitis B virus (HBV) genotyping data.

Other exclusion criteria applied during full text screening were:

- Study population was based exclusively on individuals of foreign origin (e.g., refugees) or an ethnical minority. If information on country of origin was available, data were allocated to the respective country;
- Secondary data;
- Redundant data which were also described in another record.

2.2. Extracted Variables and Assumptions

From manuscripts, which were deemed relevant during the full text review, the following information was extracted (for variable characteristics, see Table S1 in Supplementary Materials):

- Year of publication;
- Publication type;
- Year of sample collection (if no year was available, two years prior to publication was assumed);
- Date of analysis (if no information was available, one year prior to publication was assumed);
- Location of sample collection;
- Selection criteria of study participants;
- Sex and age of tested population;
- Method of genotyping;
- Number of samples and result of genotyping.

2.3. Qualitative Assessment of Data

A scoring system was developed for a descriptive assessment of the quality of genotyping data that were deemed relevant during full text screening.

1. A study quality score was calculated with equal weighting from two different scores, the genotyping and the generalizability score (lowest score implying the lowest quality):

   (a) Genotyping score: Reliability of genotype information. A weighted average was calculated from scores for:

   - Year of sample analysis in studies (30% weighting) (median was used in case of year range provided):
     - Before 2010 (before genotype I was described): Score 1;
     - 2010 or later: Score 2.
   - Ability of the method to correctly identify genotype, including recombinant viruses (e.g., genotype I) (70% weighting):
     - A. Non-sequencing-based methods like probe-/PCR-/restriction fragment length polymorphism (RFLP), or enzyme immunoassay (EIA) (partly not capable to detect all genotypes/high risk to misclassify recombinant viruses): Score 1;
     - B. Region of viral genome sequenced (capable to identify all genotypes but with medium risk to misclassify recombinant viruses): Score 3;
     - C. Whole viral genome sequenced (capable to identify all genotypes including recombinant viruses): Score 5.

   (b) Generalizability score: Potential for generalizability of genotype information to chronic HBV infections in a country in the year 2015. A weighted average was calculated from scores for:

   - Representation of country by the study location (40% weighting):
     - Samples derived from a single town or region: Score 1;
     - Samples collected from several regions or nationwide: Score 2.
   - Representation of HBV-infected population by the study population (40% weighting):
     - Favored selection (e.g., individuals of specific age group, with a specific risk factor to acquire HBV infection or showing a specific sequala): Score 1;
     - Non-favored selection (e.g., all HBV-DNA positive individuals): Score 2.
   - Year of sample collection (20% weighting). Median was used in case of year range provided:
     - Before 2000: Score 1;
     - 2000–2010: Score 2;
     - 2011 or later: Score 3.

2. A country quality score was calculated as the weighted average of study quality scores of all studies providing results for a certain country. Each study was weighted by the proportion of samples it contributed. The result was multiplied by a score for the sum of genotyped samples from all studies for the country:

   - <100 samples: Score 1;
   - 100–999 samples: Score 2;
   - 1000 or more samples: Score 3.

   The resulting score was fitted to a 1–10 scale.

2.4. Aggregation of Genotyping Data

If several sources provided genotyping results for a country, data from studies were aggregated by weighting each study by the proportion of genotyped samples it contributed. The other quality parameters were not included at this step. Inter-genotype recombinant viruses were not allocated to a specific genotype but reported separately together with co-infections with more than one genotype and samples, from which the genotype could not be determined (Table S1, Supplementary Materials).

2.5. Approximation of Number of Infections with Each Hepatitis B Virus Genotype

Prevalence of HBsAg as a marker of chronic HBV infection was retrieved from Schweitzer et al. [1]. Population data were extracted from the United Nations World Population Prospect (UN WPP) [20] for the year 2015 to calculate the number of HBV infections per country. To approximate the number of infections with a certain genotype, genotype frequency was multiplied by the number of HBV infections for each country. Data aggregation was performed in an automated way using custom coded scripts.

2.6. Creation of Maps and Additional Software

Maps were created using the Quantum Geographic Information System (QGIS) (http://www.qgis.org/en/site/) with the map file World Borders Dataset from Thematic Mapping (http://thematicmapping.org/downloads/world_borders.php) and the plugin MMQGIS (https://plugins.qgis.org/plugins/mmqgis/). The quality score was calculated in Microsoft Excel for Mac Version 16.13.1. Diagrams were created with Prism 6.0f for Mac. Custom scripts were coded in the Ruby programming language.

3. Results

3.1. Description of Hepatitis B Virus Genotyping Data

One thousand, six hundred and fifty records were identified through online search, and 35 additionally by screening references in identified records. Studies reporting HBV genotyping data, which were published between 1 January 2000 and 23 January 2018, were subjected to a selection process, as described in Figure 1 and 'Materials and Methods'.

Two hundred and thirteen publications were included in the data aggregation (Table S1, File S1, Supplementary Materials), composed of 95% peer-reviewed and 5% non-peer-reviewed (e.g., posters) publications. The majority of samples were collected between the years 1996 to 2015, with oldest samples derived from 1984 (Figure S1A, Supplementary Materials). Publication dates were equally distributed across years 2000 to 2018 (Figure S1B, Supplementary Materials). In 62% of studies, the sampling region covered small towns or areas of a country, whereas 38% of studies collected samples from several regions or nationwide. The majority (75%) of studies sequenced the viral genome to determine the genotype, with 18% of studies sequencing the whole and 57% parts of the genome. Twenty-five percent of studies performed alternative, mostly probe- or PCR-based assays to determine the genotype. Overall, included records encompass genotyping results from 26,319 HBV-infected individuals from 125 countries. The availability and quality of data varied between countries (Table S1, Supplementary Materials, and Figure 2). According to the scoring system applied in this study, most relevant data were available for Denmark, China and Cuba, for which in recent years, large numbers of samples from HBV-infected individuals from several regions of the respective country were sequenced. In contrast, quality of data was assessed to be lowest for records from Sweden, Namibia, Ireland, Angola, Bolivia, Eritrea, Liberia and Sri Lanka. This was mostly due to a small number of samples derived from a single study site, which were analyzed using non-sequencing-based methods. When relating to the number of HBV infections, about $1/3^{rd}$ of total HBV infections were represented by low (Scores 1–3), medium (Scores 4–7) or high (Scores 8–10) quality genotyping data. Quality was homogenously distributed across genotypes but rated best for genotypes B, C and I,

mainly because of their frequent appearance in China for which good quality data were available (Figure S1C,D, Supplementary Materials).

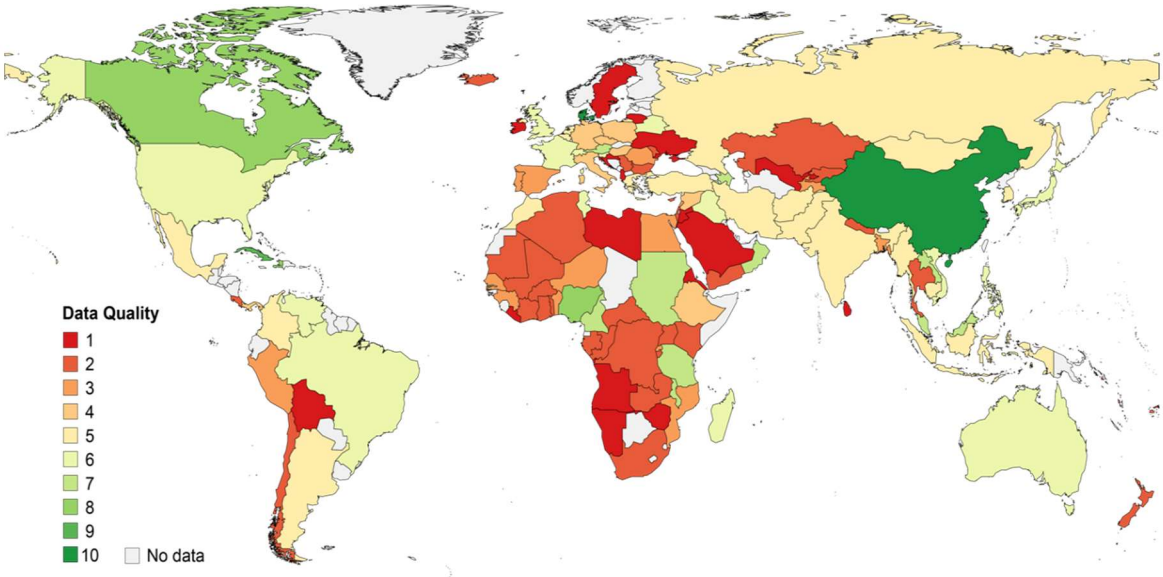

Figure 2. Quality of acquired HBV genotyping data per country. One (red) indicates the lowest and ten (green) represents the highest data quality. The scoring system was based on the sampling region, study population, sampling year, genotyping method, year of analysis, and number of analyzed samples.

3.2. World-Wide Hepatitis B Virus Genotype Appearance

In most European countries, genotypes A and D were found to be most common, with an increasing appearance of genotype A towards Northwestern Europe (Figure 3). In Eastern Europe and Western-, Central-, North- and South Asia as well as Northern Africa, genotype D was predominant without further genotypes present in a significant number. On the British Isles and Denmark, genotypes B and C were also common. In Eastern and Southeastern Asia, as well as Australasia and Oceania, genotype C was the most frequent genotype. Besides genotype C, genotype B also constituted a relative high proportion in China, Southeast Asia, and Australia, whereas genotype D was more frequent in Australasia and Oceania. One exception was the Philippines, where genotype A constituted about half of infections. In Sub-Saharan Africa (comprising Eastern Africa, Middle Africa, Southern Africa, and Western Africa), a distinct genotype distribution was found, with genotype E predominating in the western part, and decreasing proportions towards Eastern Africa, where, with the exception of Madagascar (genotype E), mainly genotype A was found. In Latin America, three genotypes (F, G and H) were found that are rare in other parts of the world. Genotype F was predominant in most Latin American countries, for which HBV sequencing data were identified. Brazil and Mexico differed from the other countries in the region. In Brazil, genotypes A and D dominated, and uniquely in Mexico, genotypes G (10.2%) and H (63.3%) were found in relevant numbers. Northern America and the Caribbean differed in their genotype distribution from Latin America. In Northern America, genotype A, B, C and D were predominant, whereas in the Caribbean, mainly genotype A, and, to a lesser extent, genotype D were found (for details, see Table S1, Supplementary Materials). Taken together, the HBV genotype distribution showed similar patterns between countries of the same world region but strongly varied between different parts of the world.

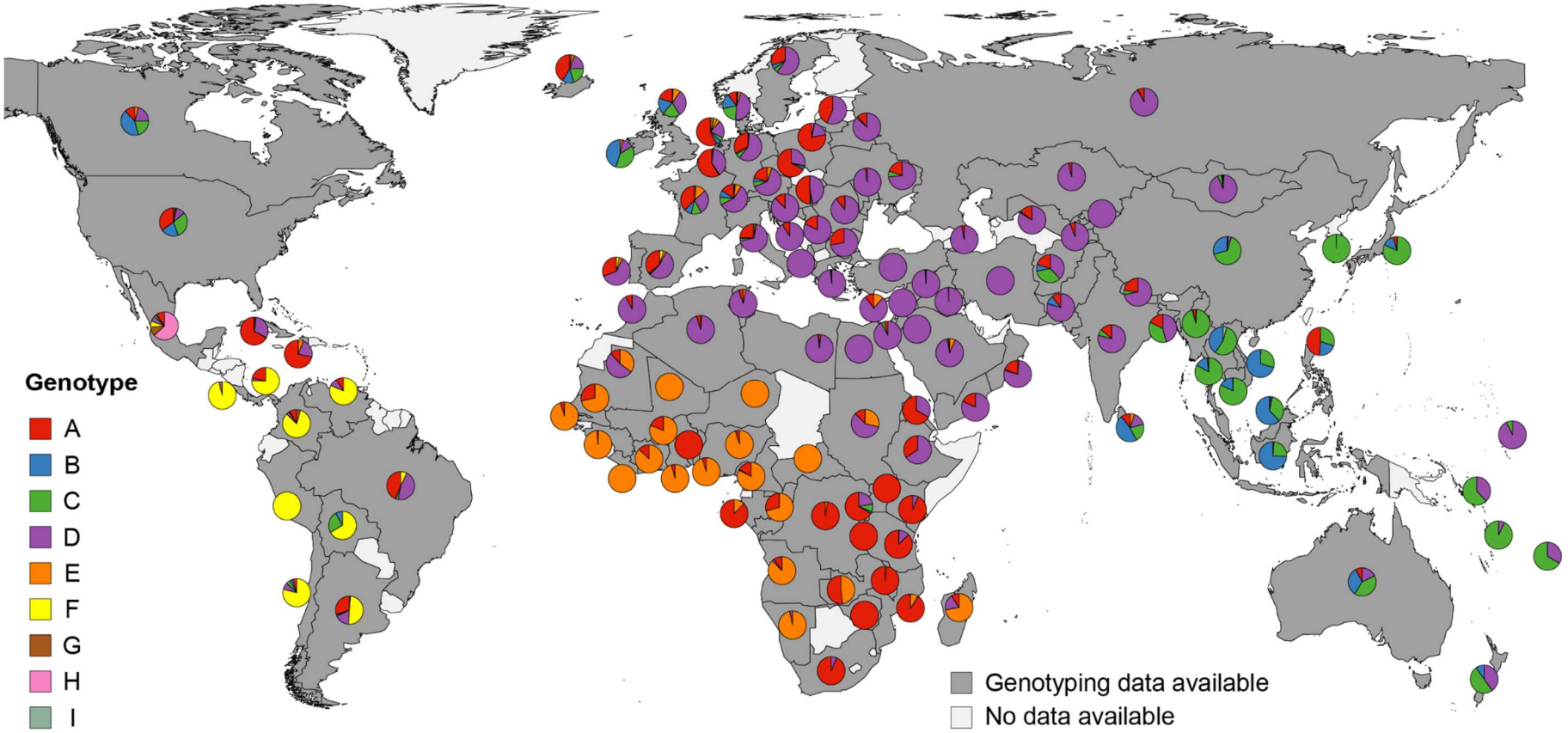

Figure 3. Distribution of HBV genotypes by country. Pie charts indicate proportional HBV genotype distributions in the respective countries. Genotype distributions within samples with successful genotyping are presented, excluding inter-genotype recombinant viruses, co-infections with more than one HBV genotype or undefined infections. Underlying literature sources and number of sequenced isolates are given in Table S1 (Supplementary Materials).

3.3. Approximation of Number of Chronic Hepatitis B Virus Infections with Each Hepatitis B Virus Genotype

We approximated the number of infections with a certain HBV genotype by extrapolating the genotype distribution found in study populations to all HBV-infected individuals in a respective country, using HBsAg prevalence estimates by Schweitzer et al. [1] and UN Population data for the year 2015 [20] (Table S2, Supplementary Materials). The distinct differences between the number of HBV-infected individuals in different world regions, together with variations in genotype frequencies, determined the total number of infections with each genotype in countries (Figure 4; Table S3, Supplementary Materials): This was most apparent for genotypes with high frequencies in regions harboring large HBV-infected populations, such as Sub-Saharan Africa (32.3% of global HBV infections, mostly genotypes A and E) or Eastern and Southeast Asia (together 36.6% of global HBV infections, mostly genotypes B and C) (Figures 4 and 5A; Table S3, Supplementary Materials). Overall, genotype C infected 26.1% of HBsAg-positive individuals worldwide and thus caused the highest number of HBV infections (Figures 4 and 5B), with 98.6% of genotype C infections occurring in Asia (Table S3, Supplementary Materials). Genotype D was found in 22.1% of all HBV-infected individuals, of which 61.9% were found in Asia with 22.0% in Africa and 13.5% in Europe. Genotype E caused 17.6% of all HBV infections globally, of which 97.0% occurred in Sub-Saharan Africa. Genotype A caused 16.9% of HBV infections worldwide, of which a majority of 72.2% was found in Sub-Saharan Africa followed by 17.2% in Asia. Globally, 13.5% of chronic HBV infections were caused by genotype B, of which 98.8% occurred in Asia. In summary, genotypes A to E were estimated to together cause 96.2% of global chronic HBV infections. The other 4 genotypes F to I together accounted only for 1.3% of all infections and occurred mostly in Latin America (genotypes F to H) or Eastern Asia (genotype I); 2.5% of global infections were described as infections with inter-genotype recombinants or mixed infections with more than one genotype or could not be defined (Figure 5B).

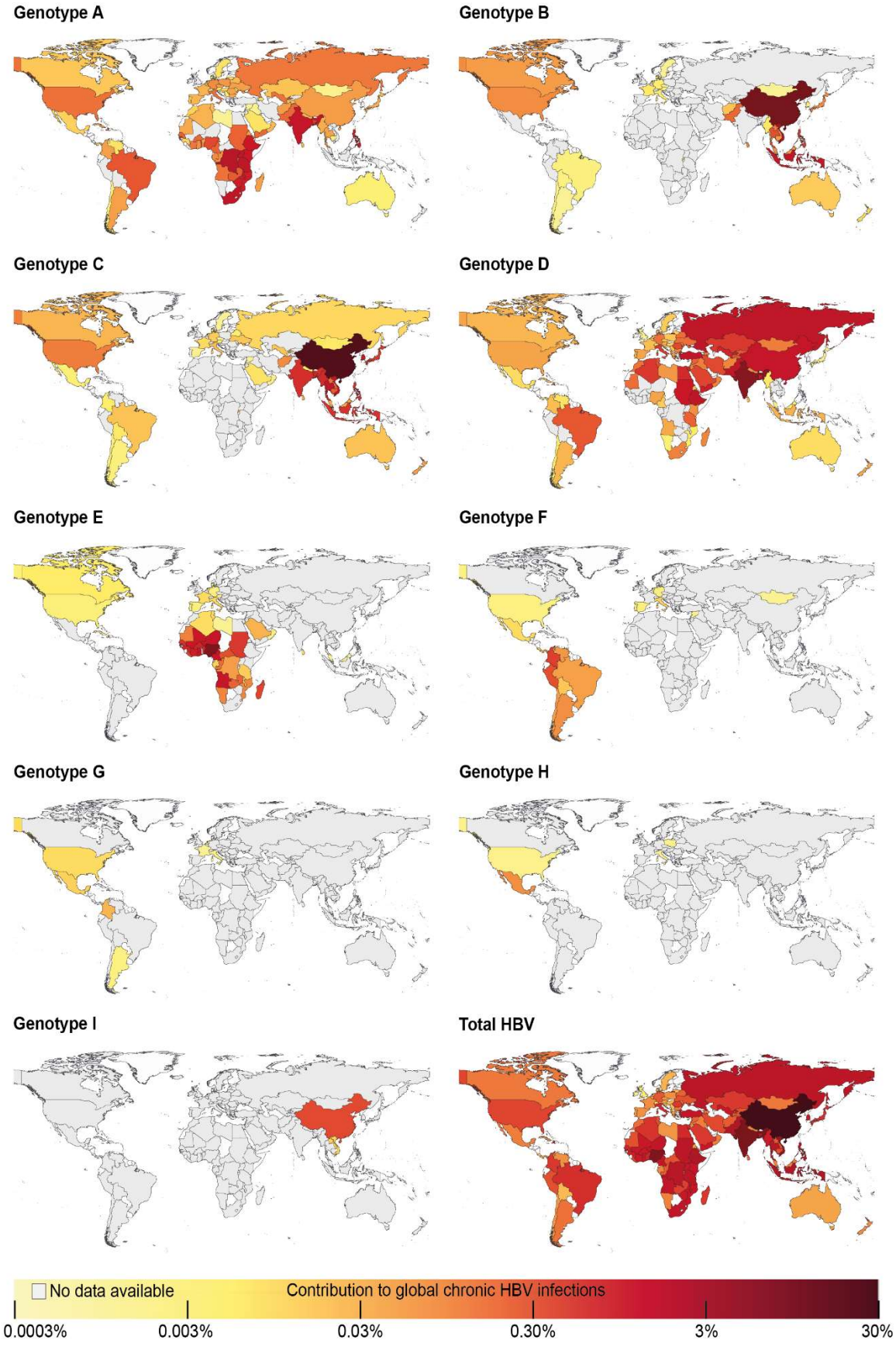

Figure 4. Contribution of genotypes to global chronic HBV infections. The number of infections with each genotype in a respective country is illustrated as percentage of global chronic HBV infections.

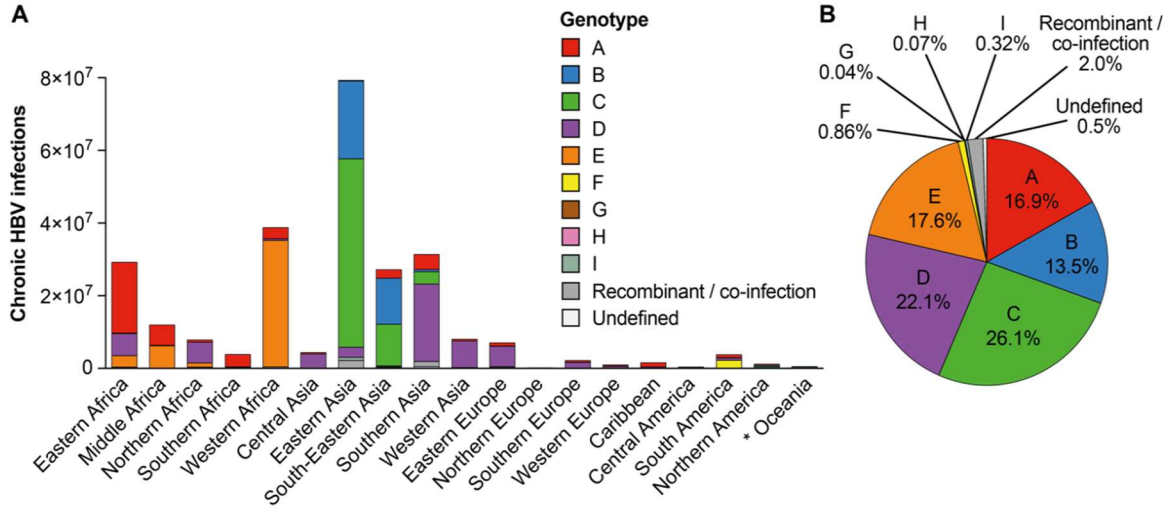

Figure 5. Approximation of the contribution of HBV genotypes to global burden of chronic HBV infection. (**A**) Estimation of the number of chronic infections with each genotype per world region. *: Oceania includes pooled data of Australia/New Zealand, Melanesia, Micronesia, and Polynesia. (**B**) Approximation of the genotype distribution within global chronic HBV infections. Values <2% are given with two decimals to prevent distortion of genotype distribution. Recombinant/co-infection: infection with an inter-genotype recombinant or with more than one HBV genotype; Undefined: genotype allocation not possible.

4. Discussion

Information on the contributions of each HBV genotype to the global burden of HBV infection is missing. This scoping review on HBV genotyping results and the combination with HBsAg prevalence and population estimates approaches this gap. We found that the different distributions of HBV genotypes, together with varying numbers of chronically HBV-infected individuals in world regions, result in strong variations in the global number of infections caused by each genotype: genotypes A to E were estimated to cause the vast majority of infections, although each in different parts of the world accounting together for 96% of chronic HBV infections worldwide. In contrast, the remaining genotypes F to I each were estimated to cause a significantly lower number of infections, together causing less than 2% of chronic HBV infections in the world.

The number and quality of data identified by our scoping review strongly varied between countries even within highly developed regions of the world. In each world region, there were countries with good, but also with poor quality data or no data at all. The distribution of HBV genotypes showed an interrelation with geographic boundaries (e.g., oceans or the Sahara Desert) or the dissemination of ethnic groups. For instances, while HBV-infections in Sub-Saharan Africa were primarily caused by genotypes A and E, this was not the case in Northern Africa. Here, the population, which rather resembles the populations of West Asia [21], showed also a similar HBV genotype distribution as this region (mainly genotype D). Additionally, world areas, in which large proportions of the population descend from migrants from other parts of the world, showed a genotype distribution reflecting the areas, from which migrants originated. This is illustrated by, for example, relative high frequencies of genotypes A, B, C and D in Northern America referring to migrants from Europe and Asia. Another example constitutes the Caribbean, where mostly genotypes A and D were found, correlating with higher proportions of migrants originating from the African continent. As a consequence of different HBV endemicity levels and population sizes, genotypes with a high occurrence in regions, such as Southeast and East Asia (genotypes B and C) and Sub-Saharan Africa (genotypes A and E), caused a large proportion of worldwide infections. In contrast, genotype F, which was dominant in Latin America, was estimated to cause less than 1% of chronic HBV infections globally, due to a relatively low HBV endemicity in this area.

Our results need to be interpreted with caution, as several technical limitations are inherent in the underlying data and the method of extrapolating the genotype distribution from study populations to global HBV infections. A potential cause of bias is the genotyping method applied: the majority of studies only sequenced parts of the viral genome or used non-sequencing-based methods. This can lead to wrong classifications, especially for recombinant viruses (including genotype I which constitutes a recombination of genotype A, C and G) [22]. In some instances, tests were used (e.g., line probe assays) with inability to detect all genotypes. Another impacting factor is the analysis year of samples: almost half of studies were published before 2010, i.e., before genotype I was described or recognized as an independent genotype. While current data confirm that genotype I is rare and only found in Southeast Asia, there is a potential that we underestimated its frequency, because infections with genotype I were missed in earlier studies.

Regarding the precision of extrapolating the genotype distribution found in a study population to all HBV-infected individuals of the respective country, some limitations need to be mentioned. Two-thirds of studies were based on a single town or region in a country. This region may not reflect HBV-infected individuals in the whole country, especially in cases, where ethnic groups that potentially carry different HBV genotypes live in separate geographic areas. However, a significant number of studies which sampled more than one region did not do this in a manner that would represent the whole country in a satisfying way, which was especially a problem for large countries (e.g., Russia).

In 40% of studies included, we identified a risk to potentially favor selection of certain genotypes. This, for example, applies to studies which exclusively included patients with advanced fibrosis/cirrhosis or HCC which could favor the selection of genotypes that are associated with a faster disease progression. Additionally, often genotypes were not detected at all in included studies for a country, but studies testing specific minorities or non-national immigrants in the country (which we excluded due to our exclusion criteria) proofed the presence of these genotypes at least at low level. This suggests that rare genotypes were often not detected, most likely because of small sample sizes or because certain minorities were not represented in study populations.

Low numbers of genotyped samples could also result in an imprecise estimation of the genotype distribution in several countries. For many countries, the sample size could be enlarged by pooling samples from several studies, but this was not possible for all countries. Further compromising the precision of the global genotype distribution, the other parameters used—HBsAg prevalence and population data—also only constitute estimates and in most cases did not derive from the same time point and might have changed over time. Moreover, some genotyping studies included samples which were either collected several years ago or during a large period, questioning representativeness for the year 2015. We furthermore could only calculate the number of genotype infections if all three parameters were available for a country. While we were able to retrieve data for 125 countries which, according to the UN WPP, represented 96% of the world population in 2015, we could not calculate genotype infections for the remaining countries. Future studies should focus on countries with highest need to perform HBV genotyping, including countries for which no data could be retrieved or countries for which we assessed available data to have poor quality (Table S1, Supplementary Materials).

Importantly, while we assessed the quality of studies using a scoring system, this served only descriptive purposes. Due to limitation of data, we did not adjust for factors besides the sample number in our aggregation analyses. We also did not include age as a factor which is of note due to the interrelation of transmission routes and genotypes which could lead to varying genotype frequencies between age groups. We omitted these factors from our analysis, as for many countries, only a single source was available and information on the age of individuals infected with a certain genotype was mostly missing. The scarcity of genotyping data, when compared to seroprevalence studies, probably results from the fact that genotyping, at least when performed by sequencing, is labor and cost-intensive, which constitutes a limitation, especially for resource-poor countries. However, we also cannot exclude that available genotyping data were missed by our literature search. As a consequence of the scarcity of HBV genotyping data, the extrapolation of genotype distribution was based on only

26,000 genotyped samples, which constitutes only around 0.01% of chronic HBV infections. Thus, the global HBV genotype distribution calculated in our study carries the risk of high uncertainties, which we did not define due to limited data points. The global genotype distribution calculated in this study should, therefore, be regarded as a first approximation, but a more precise estimation is warranted, by taking other factors and co-variates into consideration and by including additional data that may come up.

We chose not to extend our study to include subgenotypes for several reasons: (i) Only few publications included information on subgenotypes; (ii) the majority of studies used genotyping methods, which are not suitable to determine subgenotypes, including sequencing of only parts of the viral genome [4]; and (iii) the definition of subgenotypes has been subject to changes during the phase, from which genotyping studies were selected [4,23–25]. Thus, we believe the quantity and quality of available data were not sufficient to allow estimation of the subgenotype distribution with sufficient precision.

Despite the limitations described, this study provides an up-to-date insight into the worldwide HBV genotyping data from recent years and is an important initial approach to quantifying how genotypes contribute to the global burden of chronic HBV infection. Our study identified countries with no or only low quality genotyping data urgently requiring further studies. The wide distribution of HBV genotypes around the world underscores the need to ensure that the applied diagnostic tests and therapeutic approaches address the variety of HBV genotypes. Most experimental cell culture and animal models currently used to study HBV biology and new treatment options are based on genotype D or A, which we estimated, account for only 1/5th and 1/6th of infections worldwide, respectively. This reflects the need to expand experimental models to the other HBV genotypes. While efforts are ongoing to establish models for genotypes B and C, genotype E seems to be disregarded in this respect, although it seems more important than genotypes A and B. Experimental models for drug development should be expanded to at least cover genotypes A to E to represent the vast majority of HBV infections worldwide.

Supplementary Materials: The following are supplied with the manuscript and are available online at http://www.mdpi.com/2073-4425/9/10/495/s1: Table S1: Results from literature review for HBV genotyping data including quality score File S1: List of references from which genotyping data was extracted; Figure S1: Quality of included genotyping data; Table S1: Results from literature review for HBV genotyping data including the quality score; Table S2: Extrapolation of genotype distribution to all HBV-infected individuals of the country; Table S3. Geographical dispersal of HBV infections and genotypes.

Author Contributions: Conceptualization, S.V., J.J.O. and T.M.; data curation, S.V.; formal analysis, S.V.; funding acquisition, U.P. and T.M.; investigation, S.V. and T.M.; methodology, S.V., J.J.O. and T.M.; project administration, T.M.; resources, U.P.; software, S.V.; supervision, T.M.; validation, T.M.; visualization, S.V. and T.M.; writing of the original draft, T.M.; writing of review and editing, S.V., J.J.O., U.P. and T.M.

Funding: T.M. is supported by a Clinical Leave stipend of the Else Kröner Forschungskolleg "Antimikrobielle Trigger als Auslöser von Krankheiten" at Technical University Munich. The study was in part supported by the German research Foundation (DFG) via TRR179.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Schweitzer, A.; Horn, J.; Mikolajczyk, R.T.; Krause, G.; Ott, J.J. Estimations of worldwide prevalence of chronic hepatitis B virus infection: A systematic review of data published between 1965 and 2013. *Lancet* **2015**, *386*, 1546–1555. [CrossRef]
2. Word Health Organisation. *Global Hepatitis Report 2017*; World Health Organisation: Geneva, Switzerland, 2017.
3. Kramvis, A. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **2014**, *57*, 141–150. [CrossRef] [PubMed]

4. Pourkarim, M.R.; Amini-Bavil-Olyaee, S.; Kurbanov, F.; Van Ranst, M.; Tacke, F. Molecular identification of hepatitis B virus genotypes/subgenotypes: Revised classification hurdles and updated resolutions. *World J. Gastroenterol.* **2014**, *20*, 7152–7168. [CrossRef] [PubMed]
5. Tran, T.T.; Trinh, T.N.; Abe, K. New complex recombinant genotype of hepatitis B virus identified in Vietnam. *J. Virol.* **2008**, *82*, 5657–5663. [PubMed]
6. Tatematsu, K.; Tanaka, Y.; Kurbanov, F.; Sugauchi, F.; Mano, S.; Maeshiro, T.; Nakayoshi, T.; Wakuta, M.; Miyakawa, Y.; Mizokami, M. 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*, 10538–10547. [CrossRef] [PubMed]
7. Kurbanov, F.; Tanaka, Y.; Kramvis, A.; Simmonds, P.; Mizokami, M. When should "I" consider a new hepatitis B virus genotype? *J. Virol.* **2008**, *82*, 8241–8242. [CrossRef] [PubMed]
8. Yu, H.; Yuan, Q.; Ge, S.X.; Wang, H.Y.; Zhang, Y.L.; Chen, Q.R.; Zhang, J.; Chen, P.J.; Xia, N.S. Molecular and phylogenetic analyses suggest an additional hepatitis B virus genotype "I". *PLoS ONE* **2010**, *5*, e9297. [CrossRef] [PubMed]
9. Arankalle, V.A.; Gandhe, S.S.; Borkakoty, B.J.; Walimbe, A.M.; Biswas, D.; Mahanta, J. A novel HBV recombinant (genotype I) similar to Vietnam/Laos in a primitive tribe in Eastern India. *J. Viral Hepat.* **2010**, *17*, 501–510. [CrossRef] [PubMed]
10. 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*, 942–948. [CrossRef] [PubMed]
11. Locarnini, S.; Littlejohn, M.; Aziz, M.N.; Yuen, L. Possible origins and evolution of the hepatitis B virus (HBV). *Semin. Cancer Biol.* **2013**, *23*, 561–575. [CrossRef] [PubMed]
12. Sunbul, M. Hepatitis B virus genotypes: Global distribution and clinical importance. *World J. Gastroenterol.* **2014**, *20*, 5427–5434. [CrossRef] [PubMed]
13. Kramvis, A. The clinical implications of hepatitis B virus genotypes and HBeAg in pediatrics. *Rev. Med. Virol.* **2016**, *26*, 285–303. [CrossRef] [PubMed]
14. Liu, C.J.; Kao, J.H. Global perspective on the natural history of chronic hepatitis B: Role of hepatitis B virus genotypes A to J. *Semin. Liver Dis.* **2013**, *33*, 97–102. [CrossRef] [PubMed]
15. Krekulova, L.; Rehak, V.; da Silva Filho, H.P.; Zavoral, M.; Riley, L.W. Genotypic distribution of hepatitis B virus in the Czech Republic: A possible association with modes of transmission and clinical outcome. *Eur. J. Gastroenterol. Hepatol.* **2003**, *15*, 1183–1188. [CrossRef] [PubMed]
16. Komatsu, H.; Inui, A.; Fujisawa, T.; Takano, T.; Tajiri, H.; Murakami, J.; Suzuki, M. Transmission route and genotype of chronic hepatitis B virus infection in children in Japan between 1976 and 2010: A retrospective, multicenter study. *Hepatol. Res.* **2015**, *45*, 629–637. [CrossRef] [PubMed]
17. Lin, C.L.; Kao, J.H. Hepatitis B virus genotypes and variants. *Cold Spring Harb. Perspect. Med.* **2015**, *5*, a021436. [CrossRef] [PubMed]
18. Lauber, C.; Seitz, S.; Mattei, S.; Suh, A.; Beck, J.; Herstein, J.; Borold, J.; Salzburger, W.; Kaderali, L.; Briggs, J.A.G.; et al. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host Microbe* **2017**, *22*, 387–399. [CrossRef] [PubMed]
19. Paraskevis, D.; Magiorkinis, G.; Magiorkinis, E.; Ho, S.Y.; Belshaw, R.; Allain, J.P.; Hatzakis, A. Dating the origin and dispersal of hepatitis B virus infection in humans and primates. *Hepatology* **2013**, *57*, 908–916. [CrossRef] [PubMed]
20. United Nations, Department of Economic and Social Affairs Population Division. *World Population Prospects: The 2017 Revision*, Dvd ed.; United Nations: New York, NY, USA, 2017.
21. Reed, F.A.; Tishkoff, S.A. African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* **2006**, *16*, 597–605. [CrossRef] [PubMed]
22. Ismail, A.M.; Goel, A.; Kannangai, R.; Abraham, P. Further evidence of hepatitis B virus genotype I circulation in Northeast India. *Infect. Genet. Evol.* **2013**, *18*, 60–65. [CrossRef] [PubMed]
23. Shi, W.; Zhu, C.; Zheng, W.; Carr, M.J.; Higgins, D.G.; Zhang, Z. Subgenotype reclassification of genotype B hepatitis B virus. *BMC Gastroenterol.* **2012**, *12*, 116. [CrossRef] [PubMed]

24. Shi, W.; Zhu, C.; Zheng, W.; Zheng, W.; Ling, C.; Carr, M.J.; Higgins, D.G.; Zhang, Z. Subgenotyping of genotype C hepatitis B virus: Correcting misclassifications and identifying a novel subgenotype. *PLoS ONE* **2012**, *7*, e47271. [CrossRef] [PubMed]
25. Yousif, M.; Kramvis, A. Genotype D of hepatitis B virus and its subgenotypes: An update. *Hepatol. Res.* **2013**, *43*, 355–364. [CrossRef] [PubMed]

viruses

MDPI

Article

Global Occurrence of Clinically Relevant Hepatitis B Virus Variants as Found by Analysis of Publicly Available Sequencing Data

Stoyan Velkov [1], Ulrike Protzer [1,2] and Thomas Michler [1,2,*]

[1] Institute of Virology, Technical University of Munich/Helmholtz Zentrum München, Trogerstrasse 30, D-81675 München, Germany; stoyan.velkov@tum.de (S.V.); protzer@tum.de (U.P.)

[2] German Center for Infection Research (DZIF), Munich Partner Site, D-81675 Munich, Germany

* Correspondence: thomas.michler@tum.de; Tel.: +49-89-4140-6814

Received: 30 September 2020; Accepted: 20 November 2020; Published: 23 November 2020

Abstract: Several viral factors impact the natural course of hepatitis B virus (HBV) infection, the sensitivity of diagnostic tests, or treatment response to interferon-α and nucleos(t)ide analogues. These factors include the viral genotype and serotype but also mutations affecting the HBV surface antigen, basal core promoter/pre-core region, or reverse transcriptase. However, a comprehensive overview of the distribution of HBV variants between HBV genotypes or different geographical locations is lacking. To address this, we performed an in silico analysis of publicly available HBV full-length genome sequences. We found that not only the serotype frequency but also the majority of clinically relevant mutations are primarily associated with specific genotypes. Distinct mutations enriched in certain world regions are not explained by the local genotype distribution. Two HBV variants previously identified to confer resistance to the nucleotide analogue tenofovir in vitro were not identified, questioning their translational relevance. In summary, our work elucidates the differences in the clinical manifestation of HBV infection observed between genotypes and geographical locations and furthermore helps identify suitable diagnostic tests and therapies.

Keywords: hepatitis B virus; genotype; serotype; escape mutation; pre-core mutation; nucleoside resistance mutation

1. Introduction

Hepatitis B is a major global health burden with nearly a quarter of the human population exposed to infection with hepatitis B virus (HBV), which is the causative agent [1]. While acute infection is self-limiting, it can cause symptomatic hepatitis and, in some cases, liver failure and death. In contrast, patients who develop chronic infection run the risk of developing long-term sequalae such as liver cirrhosis or hepatocellular carcinoma (HCC). A total of 257 million, or 3.5% of the world's population, are estimated to be chronically infected [2]. HBV is a major cause of liver cirrhosis and one of the most prevalent carcinogens in the world [3,4]. While deaths due to other major pathogens such as human immunodeficiency virus (HIV), mycobacterium tuberculosis, and malaria are declining, HBV-related deaths—currently estimated at 887,000 per year [2]—are increasing, making HBV a leading cause of death attributed to infectious disease [5].

The spectrum of outcomes following HBV infection varies, and several factors are associated with certain phenotypes. These include age at the time of infection, immune status, human leukocyte antigen type [6], and ethnicity [7]. However, environmental factors, such as alcohol [8] or aflatoxin [9], also play a role in disease progression and incidence of HCC.

Additionally, several viral factors, including the viral inoculum size at infection [10], the viral genotype and serotype, as well as mutations ascribing the virus to a certain phenotype, have shown

clinical relevance (Reviewed in [11]). HBV is classified into at least nine genotypes (A–I) that impact the primary transmission route, rate of progression after infection, response to interferon alpha treatment, and incidence of HCC [12,13]. A putative 10th genotype 'J' has been proposed [14]; however, as it has only been isolated from a single patient and displays greater homology to gibbon HBV, it is not universally accepted as an independent genotype [15].

HBV variants cannot only be differentiated by their nucleotide sequence but also their reactivity toward reference antibodies, which defines the "serotype". Despite playing a relative minor role in current patient care, they can potentially impact the efficacy of vaccines or antibody-based therapies as well as the recognition by diagnostic tests. In addition to genotype and serotype, several mutations have been identified to alter the clinical outcome, diagnostics, and treatment response to HBV infection (reviewed in [16]). This includes mutations in the HBV surface antigen (HBsAg), which render the virus undetectable in diagnostic tests or evasion from vaccine-induced or therapeutic antibodies [17], and mutations in the reverse-transcriptase (RT) domain of the viral polymerase driving resistance to nucleoside treatment [18]. Finally, mutations in the basal core promoter (BCP)/pre-core region are associated with increased risk of fulminant hepatitis or HCC [19,20].

Several publications describe or review the clinical relevance of HBV variants [21–29]. However, whilst previous studies analyze the frequency of mutations in a certain genomic region, within selected populations or within specific genotypes [30–34], there is currently no comprehensive overview of the distribution of clinically relevant HBV variants between world regions or genotypes. However, this information could be useful to understand varying phenotypes of the different HBV genotypes and inform further optimization of diagnostic tests and treatment regimens. Therefore, we performed a computerized analysis to study the frequency of clinically relevant HBV variants within publicly available HBV sequences.

2. Materials and Methods

2.1. Retrieval of HBV Sequences

Hepatitis B virus (HBV) sequences were retrieved from the Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore) of the National Center for Biotechnology Information (NCBI). To identify the HBV sequences, the taxon ID 10407 that represents HBV in the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) was used. All available data for each entry were retrieved in GenBank format [35].

2.2. Allocation of Sequences to HBV Genotypes

Basic Local Alignment Search Tool (BLAST) [36] version 2.9.0+ was used to generate a reference nucleotide database to assign genotypes to the retrieved HBV sequences. The reference BLAST database was created from previously published sequences of full-length HBV genomes of each sub-genotype [37], which were combined with additional non-human HBV reference sequences. The non-human HBV references were added to exclude the possibility of mis-genotyping sequences that do not belong to human HBV. The default makeblastdb parameters for dbtype nucl were used with FASTA [38] format references retrieved from their GenBank entries for the BLAST database generation.

The list of accession numbers of human HBV genomes followed by sub-genotype used in the reference BLAST database are as follows: JN182318: A1; HE576989: A2; AB194951: A3; AY934764: A4; FJ692613: A5; GQ331047: A6; FN545833: A7; AB642091: B1; FJ899779: B2; GQ924617: B3; GQ924626: B4; GQ924640: B5; JN792893: B6; GQ358137: B7; GQ358147: B8; GQ358149: B9; AB697490: C1; GQ358158: C2; DQ089801: C3; HM011493: C4; EU410080: C5; EU670263: C6; GU721029: C7; AP011106: C8; AP011108: C9; AB540583: C10; AB554019: C11; AB554025: C12; AB644280: C13; AB644284: C14; AB644286: C15; AB644287: C16; GU456636: D1; GQ477452: D2; EU594434: D3; GQ922003: D4; GQ205377: D5; KF170740: D6; FJ904442: D7; FN594770: D8; JN664942: D9; FN594748: E; FJ709464:

F1b; DQ899146: F2b; AY090459: F1a; DQ899142: F2a; AB036920: F3; AF223965: F4; GU563556: G; AB516393: H; FJ023659: I1; FJ023664: I2; AB486012: J.

The list of accession numbers of non-human HBV genomes followed by naming used in the reference BLAST database are as follows: K02715: GSHV; U29144: ASHV; AF193864: OGHBV; AJ251935: STHBV; AY226578: WMHBV; AY628097: WCHBV; JQ664503: GOHBV; JQ664509: CHHBV; KY962705: BATHBV; KC790373: BATHBV1; KC790374: BATHBV2; KC790375: BATHBV3; KC790376: BATHBV4; KC790377: BATHBV5; KC790378: BATHBV6; KC790379: BATHBV7; KC790380: BATHBV8; KC790381: BATHBV9; AB823662: GIHBV; KT893897: SGIHBV; KT345708: STHBV2; KY703886: CMHBV; MF471768: DHBV.

2.3. Analysis of Regional Frequency of HBV Genotypes and Clinically Relevant Variants

HBV sequences and associated data were downloaded and the information under FEATURES in the GenBank entry was evaluated for the country key. Countries were grouped in world regions as follows (number of included full-length HBV sequences for each country in brackets): Eastern Africa: Ethiopia (13), Kenya (17), Madagascar (1), Malawi (2), Mauritius (1), Rwanda (14), Somalia (9), Tanzania—United Republic of (3), Uganda (2), Zimbabwe (4); Middle Africa: Angola (14), Cameroon (62), Central African Republic (30), Congo—The Democratic Republic Of (5), Gabon (6); Northern Africa: Egypt (6), Sudan (17), Tunisia (5); Southern Africa: Botswana (10), Namibia (6), South Africa (58); Western Africa: Benin (4), Burkina Faso (17), Cape Verde (10), Gambia (2), Ghana (14), Guinea (74), Liberia (6), Mali (1), Niger (24), Nigeria (27); Caribbean: Cuba (8), Haiti (49), Martinique (23); Central America: Costa Rica (2), El Salvador (2), Mexico (27), Nicaragua (4), Panama (40); Northern America: Canada (54), Greenland (15), United States (508); South America: Argentina (131), Bolivia—Plurinational State of (11), Brazil (72), Chile (32), Colombia (1), Peru (3), Uruguay (8), Venezuela—Bolivarian Republic of (34); Central Asia: Kazakhstan (2), Tajikistan (8), Uzbekistan (10); Eastern Asia: China (2294), Hong Kong (75), Japan (281), Korea—Republic of (91), Mongolia (13), Taiwan—Republic Of China (57); South-Eastern Asia: Cambodia (28), Indonesia (120), Lao People's Democratic Republic (43), Malaysia (195), Myanmar (18), Philippines (15), Thailand (107), Vietnam (145); Southern Asia: Bangladesh (81), India (330), Iran—Islamic Republic Of (53), Nepal (3), Pakistan (6); Western Asia: Saudi Arabia (4), Syrian Arab Republic (58), Turkey (83), United Arab Emirates (1); Eastern Europe: Belarus (8), Poland (33), Russian Federation (107); Northern Europe: Estonia (16), Ireland (1), Latvia (8), Sweden (15), United Kingdom (8); Southern Europe: Italy (47), Serbia (7), Spain (15); Western Europe: Belgium (116), France (12), Germany (12), Netherlands (6); Oceania: Australia (55), New Zealand (30), Fiji (5), New Caledonia (7), Papua New Guinea (16), Vanuatu (1), Kiribati (4), Samoa (1), Tonga (3).

2.4. Sequence Analysis

All sequences identified as non-human HBV with our BLAST database, as well as sequences classified as "unverified" or "non-functional" in the NCBI database were excluded. The occurrence of unclear bases in the sequence (residues labeled '*n*') was an additional reason for exclusion. To allow analysis of sequences from different genotypes that vary in length, blank insertions were inserted into shorter sequences to achieve a length of 3257 bp. Sequences that did not start at the EcoRI site, which is generally considered as the start point for annotation, were corrected to allow for alignment with other sequences.

2.5. Analysis of Amino Acid Sequences

To analyze the HBV proteins, nucleotide sequences were translated to amino acid sequences with sixpack (EMBOSS) [39] using the orfminsize 100 and mstart options. Only open reading frames (ORF) starting with a methionine and a minimum length of 100 amino acids were taken into account. The correct reading frame of the respective protein was identified by a BLAST search against a database containing reference sequences of the HBV proteins. After assignment to the correct protein, a further

size exclusion was performed by excluding implausible short proteins. Requirements for amino acid sequence length were ≥330 for pre-S1/pre-S2/S, ≥140 for pre-core/core, ≥700 for polymerase, and ≥100 for X. Proteins and nucleotide sequences were aligned in separate FASTA files using MUSCLE v3.8.1551 [40]. To account for the different lengths of the genotypes, sequences of each genotype were first aligned separately and, after including blank positions in the sequences of shorter genotypes, all sequences were combined into a single file.

2.6. Prediction of HBV Serotypes

Serotypes were predicted based on the amino acid variation at defined positions of HBsAg —122, 127, 134, 159, 160, 177, and 178—as previously described [41] and as outlined in Table 1. Sequences that could not be assigned to a specific serotype by the combinations of amino acids below were classified as undefined.

Table 1. Amino acid combinations within hepatitis B virus surface antigen (HBsAg) used to predict hepatitis B virus (HBV) serotypes. R = Arginine, K = Lysine, P = Proline, T = Threonine, L = Leucine, F = Phenylalanine, A = Alanine, V = Valine, Q = Glutamine.

Serotype	**Amino Acid Position of HBsAg**						
	122	**127**	**134**	**159**	**160**	**177**	**178**
ayw1 (option 1)	R	P	F	-	K	-	-
ayw1 (option 2)	R	P	-	A	K	-	-
ayw2	R	P	-	-	K	-	-
ayw3	R	T	-	-	K	-	-
ayw4	R	L	-	-	K	-	-
ayr	R	-	-	-	R	-	-
adw2	K	P	-	-	K	-	-
adw3	K	T	-	-	K	-	-
adw4q−	K	L	-	-	K	-	Q
adrq+	K	-	-	-	R	V	P
adrq−	K	-	-	-	R	A	-

2.7. Analysis of Amino Acid Conservation and Frequency of Clinically Relevant Mutations

To analyze the conservation of nucleotide and protein sequences, each aligned FASTA file containing the full-length HBV genome sequence or the individual protein sequence was analyzed using a custom coded inhouse script. Briefly, each position of the analyzed sequence was individually assessed and the frequency of each nucleotide or amino acid counted. Based on the total number of sequences, the nucleotide/amino acid distribution was calculated for each position. The consensus sequence was generated by taking the most frequent nucleotide/amino acid at each position. Clinically relevant mutations were analyzed based on the overview established by Lazarevic et al. [42].

2.8. Data Processing

All data processing, including the evaluation of nucleotide and amino acid sequence conservation, serotype prediction, and analysis of frequency of clinically relevant mutations was performed with custom inhouse scripts written in Ruby programming language (https://www.ruby-lang.org). Graphical representations were created in Graphpad Prism 8.4.3 (https://www.graphpad.com/scientific-software/prism/). The phylogenetic tree was generated using RAxML-NG v. 0.9.0 [43] from the aligned HBV reference from the BLAST database. Visualization of the tree was performed with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

3. Results

3.1. Sequence Acquisition and Processing

We retrieved 115,955 HBV nucleotide sequences (Figure 1A) from the NCBI database. Sequences that contained unspecified nucleotides or were marked as "non-functional" or "unverified" in the database entry were excluded. To allocate sequences to genotypes, a search was performed using BLAST against a database of reference *hepadnaviridae* genomes containing reference sequences of the different HBV genotypes and sub-genotypes as well as viruses with non-human hosts (see methods). After exclusion of non-human HBV sequences, 82,813 sequences were processed further.

Once the sequences were filtered by length, we selected only full-length genomes for further analysis. The size range was chosen to ensure that sequences of the shortest (genotype D with 3182 bp) as well as longest genotype (genotype A with 3221 bp) were represented. However, to account for the size distribution of available sequences and to allow the inclusion of variants with unusual length, the size range was enlarged from 3150 to 3275 bp (Figure 1B). This resulted in a panel of 7278 HBV full-length genome sequences for further analysis. Genotypes were differentially represented in our database. Genotype C had the most abundant full-length genome sequences (n = 2700) followed by genotype B (n = 1539), D (n = 1218), A (n = 1004), E (n = 312), F (n = 275), I (n = 96), G (n = 89), and genotype H (n = 44) (Figure 1C). When compared to the global genotype distribution as estimated by us previously [44], we found that genotype E sequences were strongly underrepresented (4.2% vs. 17.6%). Genotypes A and D were also slightly underrepresented (13.8% vs. 16.9%; 16.7% vs. 22.1%), whereas the remaining genotypes (B, C, F, G, H, I) were overrepresented in our database.

We also determined the origin of 6142 out of 7278 samples from the NCBI database. Sequences from countries within the same geographical area (for the definition of the geographical areas, see methods) were pooled for a better overview and to allow a more reliable calculation of the frequencies of clinically relevant HBV variants. The majority of samples originated from Eastern Asia (45.8% of samples, 34.4% alone from China), followed by South-Eastern Asia (10.9%), Northern America (9.4%), and Southern Asia (7.7%; Figure 1D). We noted a distinct distribution the HBV genotypes across the globe [44], with genotype B and C sequences mostly derived from East Asia and Southeast Asia, genotype D sequences from Southern Asia, Western Asia, and Europe, and genotype E sequences from Sub-Saharan Africa. Genotype A was widely distributed throughout world regions, with most sequences (302) originating in Northern America. In line with their geographical occurrence, genotype F, G, and H sequences were mainly identified in Southern America and genotype I sequences were mainly identified in South-Eastern Asia (Figure 1D). From the 7278 full-length nucleotide sequences, we predicted the amino acid sequences expressed from open reading frames. As some sequences did not contain start codons for certain ORFs, or coded for implausible short sequences due to premature stop codons, this resulted in 7211 pre-core/core, 7158 polymerase, 7157 pre-S1/pre-S2/S, and 7242 X protein sequences.

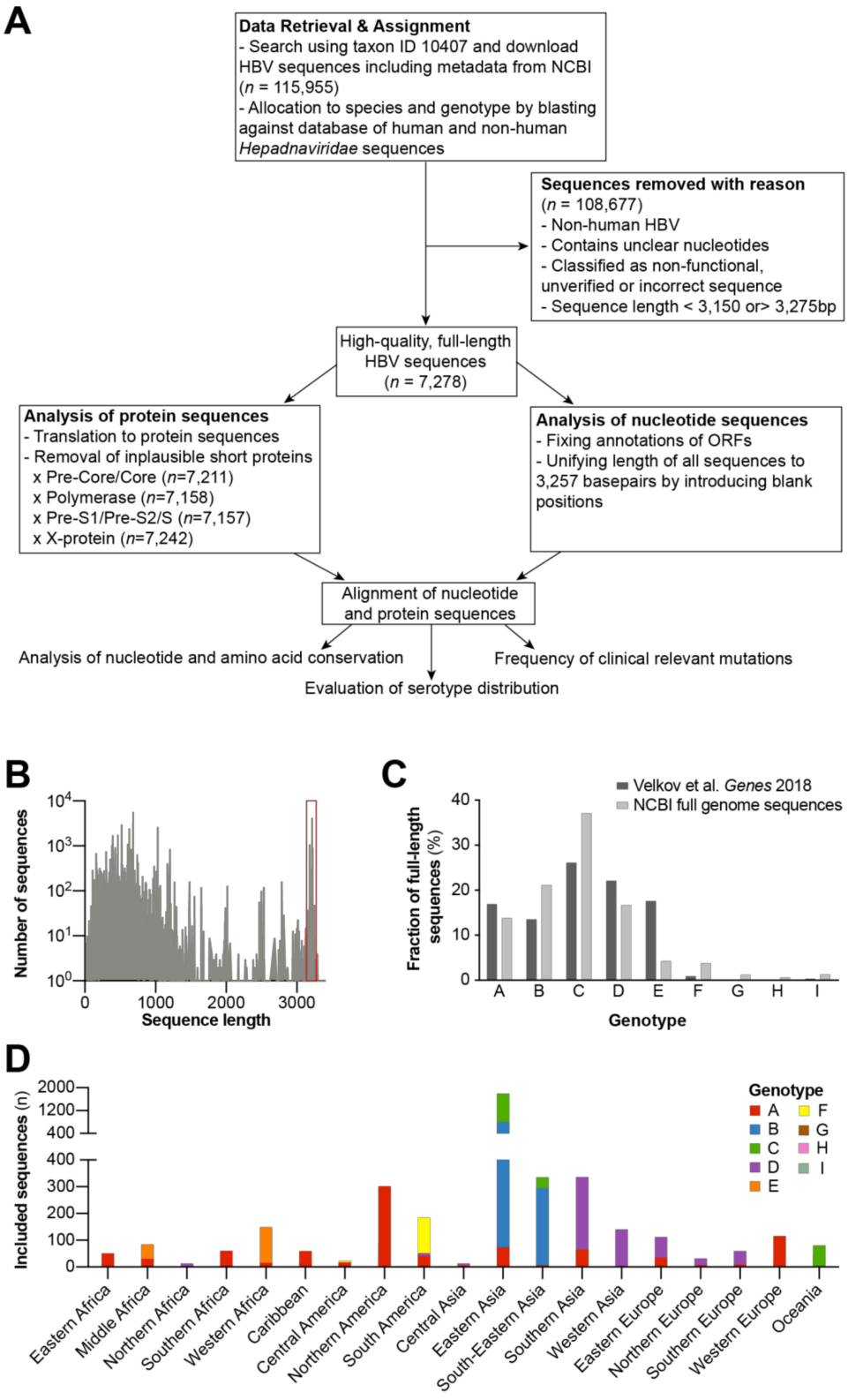

Figure 1. In silico analysis of publicly available full-length HBV sequences. (**A**) Work flow of study to estimate the frequency of clinically relevant HBV variants. (**B**) Retrieved HBV sequences were filtered by length, and only those between 3150 and 3275 base pairs (as indicated by the red box) were considered for further analysis. (**C**) Genotype distribution of included full-length HBV sequences compared to estimates of the global genotype distribution determined in Velkov et al. Genes 2018. (**D**) Overview of geographical origin and number of sequences for each genotype.

3.2. Distribution of HBV Serotypes

The structural basis that determines HBV serotype has been thoroughly studied, leading to the discovery that certain amino acids at positions 122, 127, 134, 159, 160, 177, and 178 of HBsAg (shown in Table 1) determine reactivity toward serological reference antibodies. This allows the serotype of an HBV variant to be predicted, depending on a known amino acid sequence [41]. Interestingly, when analyzing the conservation of HBsAg, we found significant variation especially at positions that determine the serotype (Figure 2A). The different amino acids found at these positions were mostly consistent with the algorithm determining the HBV serotypes (Table 1), as almost all sequences studied could be allocated to a certain serotype. Genotypes were each associated with a distinct serotype (Figure 2B). Most genotypes presented primarily with the adw serotype; however, the majority of the genotype A, I, G, and B sequences presented as adw2 and the genotypes H and F presented as adw4. In contrast, genotypes D and E presented with the ayw serotype, with genotype D almost equally divided into ayw2 and ayw3 serotypes, whereas genotype E was almost exclusively ayw4. In contrast, genotype C was the only genotype predicted to have the adrq+ serotype in significant quantities. When analyzing all HBV full-length sequences irrespective of genotype (pie chart labeled "Total", Figure 2B) the majority of sequences had an adrq+ (30.9%, which derived almost exclusively from genotype C sequences, Figure 2C) or adw2 serotype (33.9%, mainly genotype B and A, Figure 2C), followed by ayw2 (11.0%) and ayw3 serotypes (7.8%). Although representing almost all genotype H and F sequences, the adw4q- serotype constituted only 4.2% of total sequences, reflecting the low number of included genotype H and F sequences (Figure 2C), due to the relatively low global occurrence [44].

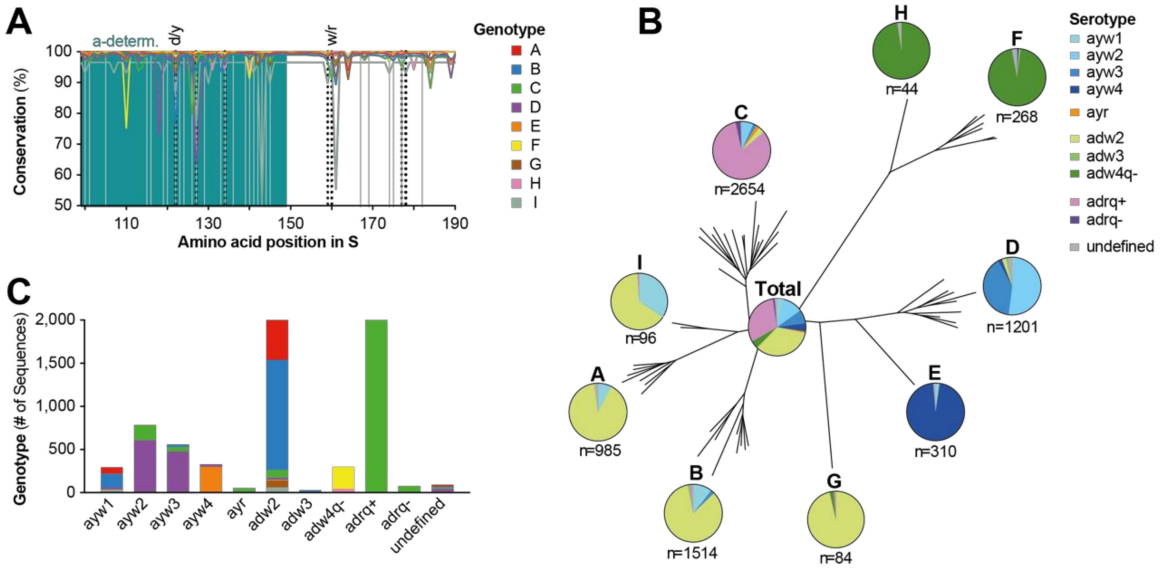

Figure 2. Distribution of HBV serotypes. (**A**) Amino acid conservation of HBsAg (starting at position 99 of S, marking the beginning of the major hydrophilic region) for each HBV genotype was determined by using the consensus sequence of each genotype as a reference. Relevant positions for serotype prediction are indicated by vertical dotted lines. Positions at which clinically relevant mutations occur (for details see Figure 3) are indicated by vertical gray lines. a-determ. = a-determinant. (**B**) Distribution of predicted serotypes within each HBV genotype. The serotype distribution of all global HBV sequences is shown in the middle. The phylogenetic tree was obtained using reference sequences for all HBV genotypes and sub-genotypes, which were employed to identify genotypes of sequences by Basic Local Alignment Search Tool (BLAST) search (see methods for details). (**C**) Total number of sequences with a certain serotype, subdivided by their genotype.

3.3. Clinically Relevant HBsAg Mutations

Then, we interrogated HBsAg coding sequences for previously described clinically relevant mutations, which have been shown to cause occult infection, false negative diagnostic tests, or escape from vaccine induced or passively administered antibodies. The overall frequencies of such mutations were low, and the different mutations were relatively evenly distributed between genotypes (Figure 3). The only exception constituted the P127H/L mutations, which are associated with occult HBV infection appearing to be the wild-type amino acid for genotypes E, F, and H (96.8%, 98.9% and 97.7%). In contrast, the R122P (associated with occult infection) was not identified in any sequence, questioning its clinical significance. S136P and C139R were also very infrequent (each 0.04%) and only identified in genotype C. When looking at the regional distribution of HBsAg mutations (Figure 3B), P127H/L showed a concentration in the African continent as well as Central and South America, where most sequences belonged to either genotype E, F, or H (Figure 1D). Interestingly, the V177A mutation did not show a significant enrichment in any genotypes; however, it was found in approximately 33% of sequences from Oceania (including Australia, New Zealand, Melanesia, and Polynesia).

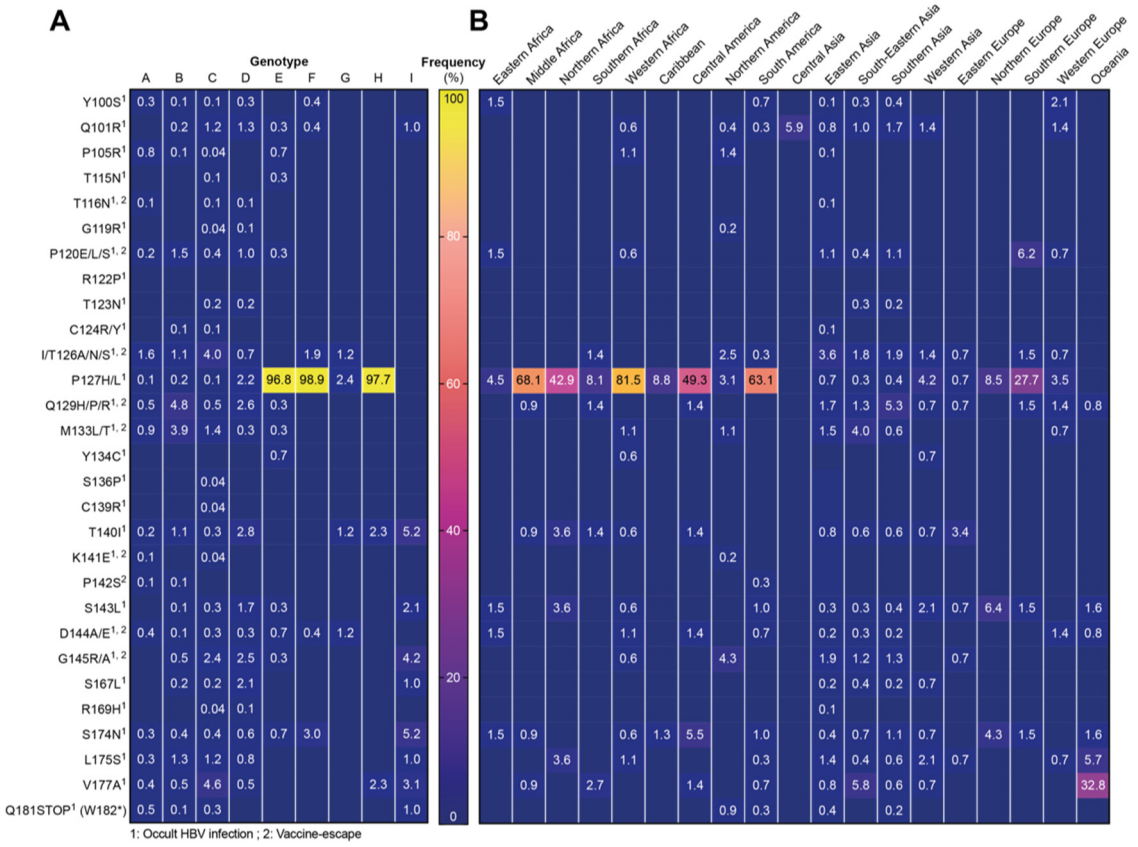

Figure 3. Frequency of clinically relevant HBsAg mutations. Numbers in the table indicate percentage of sequences containing the respective mutation within (**A**) each genotype or (**B**) different world regions. Blue fields without numbers represent a value of 0.0%. Superscript numbers indicate the clinical finding, which has been associated with the respective mutation.

3.4. Frequency of Resistance Mutations against Reverse-Transcriptase Inhibitors

Then, we analyzed the conservation of the reverse-transcriptase (RT) domain of the polymerase protein. As this region is targeted with standard-of-care therapy using nucleoside analogues, mutations rendering the virus resistant to such treatments are highly clinically relevant. The different domains (A–G) of the RT that are crucial for function were found to be highly conserved throughout all genotypes, whereas the other regions showed greater variation (Figure 4A).

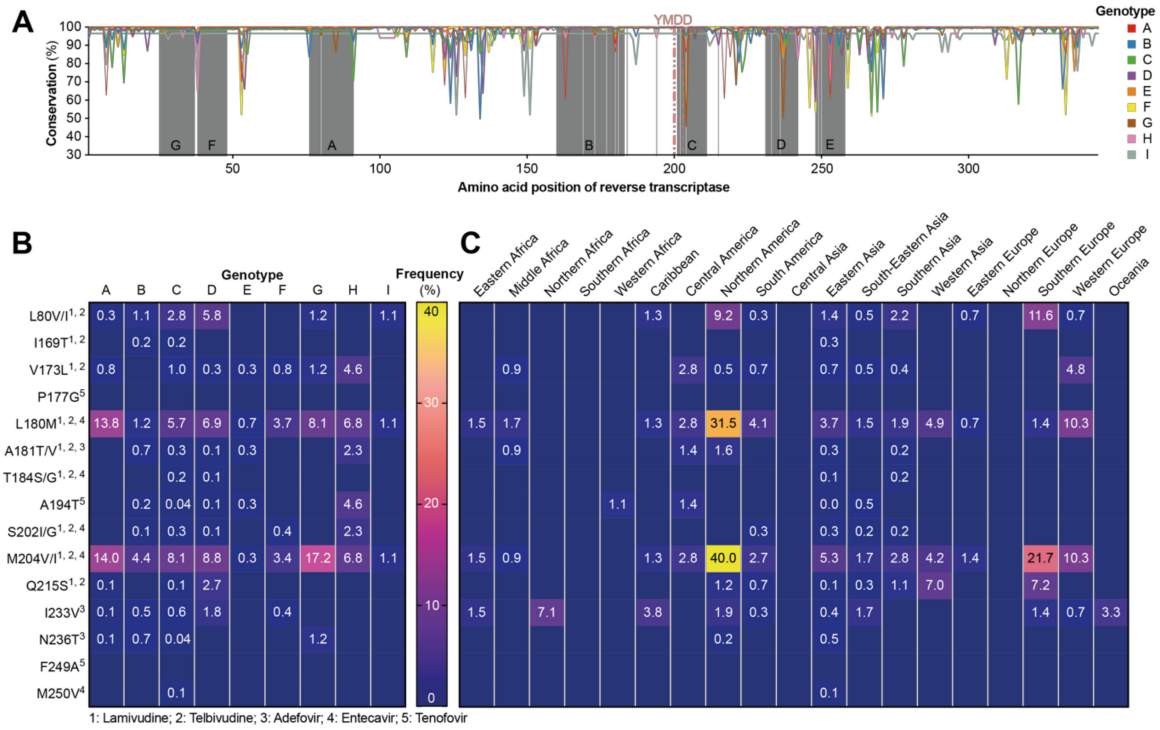

Figure 4. Frequency of mutations in the reverse transcriptase associated with resistance to nucleoside analogues. (**A**) Conservation of the reverse transcriptase (RT) part of the HBV polymerase within genotypes as determined using the consensus sequence of each genotype as a reference. The different domains (A–G) of the RT are highlighted by underlying gray color. Vertical light gray lines indicate positions at which resistance mutations occur. (**B**,**C**) Numbers in the table indicate percentage of sequences containing the respective mutation within (**B**) each genotype or (**C**) the different world regions. Blue fields without numbers represent a value of 0.0%. Superscript numbers indicate the clinical finding, which has been associated with the respective mutation.

In a next step, we determined the frequency of mutations that are known to cause resistance to nucleoside therapy. M204V/I and L180M were the most abundant mutations found that were present in sequences of all genotypes, and they are associated with resistance toward lamivudine, telbivudine, and entecavir. Genotype A showed the highest frequency of both of these mutations (L180M: 13.8% and M204V/I: 14.0%) followed by genotype G (8.1% and 17.2%). High frequencies (between 5 and 9%) were also found in genotypes C, D, and G but were found in a minority of genotype E and I sequences (<1.0% and 1.1%, respectively). A few mutations have been found to drive resistance toward tenofovir treatment. These include A194T, which was identified in HIV/HBV co-infected patients and was shown to mediate a partial resistance toward tenofovir [45]. Two other mutations have been associated with tenofovir resistance, P177G and F249A; however, these mutations were created by site-directed mutagenesis in vitro [46], and it is unclear if they are found in patient-derived circulating virus. Interestingly, none of the >7000 sequences analyzed by us contained either the P177G or F249A mutation, which questions their clinical relevance. In contrast, A194T was identified in several genotypes, albeit in very low numbers, with the exception of genotype H, where 4.6% of sequences harbored this mutation. While only six out of 15 resistance mutations we looked for were found in genotype H, all of these were present above 2%. In genotype E, similar mutations were found as in genotype H, but with overall lower frequencies (each at or below 0.7%).

When analyzing the geographical distribution of each mutation, we found the highest frequency, especially of the L180M and M204V/I mutations, in Northern America (mainly comprising genotype A sequences, Figure 1D), followed by Europe and Eastern/Southern Asia (Figure 4C). In contrast, lower numbers were found on the African and Asian continents, as well as Central and South America.

3.5. Distribution of Mutations in the Basal Core Promoter and Pre-Core Region

The basal core promoter (BCP) and pre-core regions of the HBV genome play an important role in determining the pathogenicity of HBV, and variants have been associated with an increased risk of fulminant hepatitis or HCC. Therefore, we analyzed the nucleotide sequence of this region and found the direct repeat 1 (DR1) region of the BCP and the epsilon stem loop to have the highest conservation across all genotypes (Figure 5A). As previously described, we found that genotype G sequences had an insertion at nucleotide position 1906–1941 (not shown), making it the longest of all genotypes (3248 bp). When analyzing the occurrence of clinically relevant mutations, we found BCP and pre-core mutations in relatively high abundance in almost all genotypes (Figure 5B). For genotype G, many "mutations" (C1653, T1753C, A1762T, G1764A and G1896A) were found with such high frequencies (93.3–100%) that they should be regarded as wild-type. Of all mutations, A1762T and G1764A were found at the highest frequencies in other genotypes, with decreasing rates in genotypes C, A, I, F, H, D, B, and E.

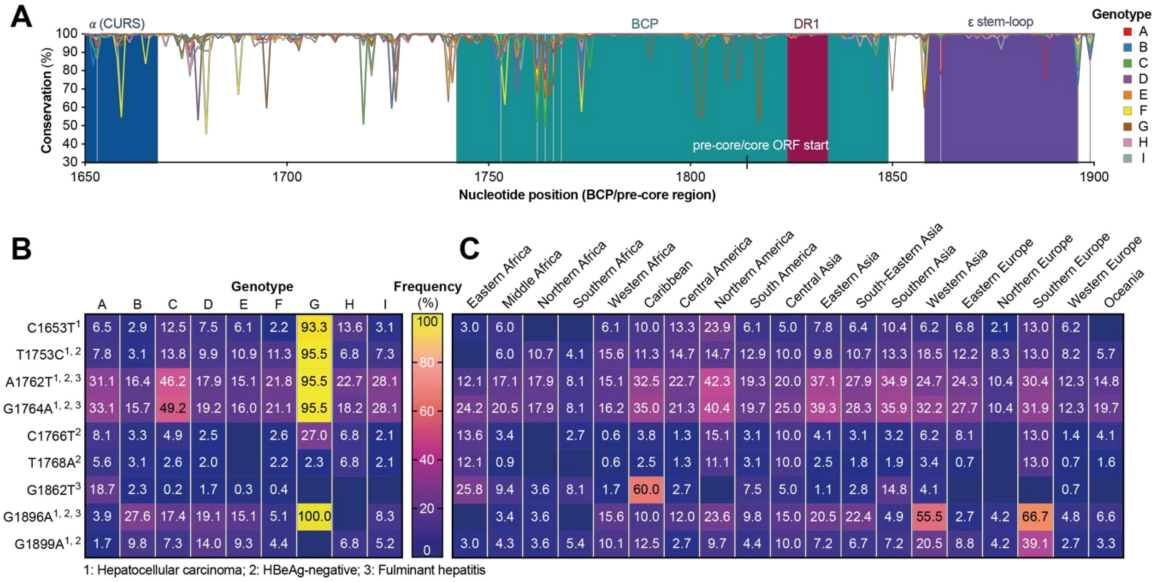

Figure 5. Frequency of clinically relevant mutations in basal core promoter or pre-core region of HBV. (**A**) Conservation of basal core promoter (BCP) and pre-core region of the HBV genome, vertical gray lines indicate positions as which clinically relevant mutations occur. Numbering according to convention starting at EcoRI site. α CURS = α core upstream regulatory sequence. (**B**,**C**) Frequency of clinically relevant mutations in the BCP/pre-core region within (**B**) each genotype or (**C**) the different world regions. Numbers in the table indicate percentage of sequences containing the respective mutation. Blue fields without numbers represent a value of 0.0%. Superscript numbers indicate the clinical finding, which has been associated with the respective mutation.

The majority of mutations did not show a distinct geographical enrichment, barring a few exceptions. G1862T, which was frequently found in the Caribbean, and G1896A, which was identified in more than 50% of sequences from Western Asia and Southern Europe, show distinct geographical enrichment that cannot be explained by genotype G, as it is rarely found in these regions (Figure 1D).

4. Discussion

HBV infection presents with a diverse disease profile with variation seen in the primary transmission mode, rate of disease progression, symptomatic disease, occurrence of sequalae, diagnostics, and treatment response. In addition to environmental and host factors, HBV genomic variation has been shown to be associated with a certain disease phenotype. In this study, we analyzed publicly available full-length HBV sequences to get an overview over the frequencies of clinically relevant HBV variants worldwide.

Importantly, the results of our study should be taken with caution, as several confounding factors potentially influenced the frequencies of mutations identified. Most importantly, sequences isolated from patients exhibiting an abnormal disease phenotype might be preferentially sequenced; thus, one can assume an enrichment of clinically relevant mutations compared to naturally circulating variants. Second, it is not clear if these mutations are associated with a specific disease phenotype in all genotypes; thus, harboring a certain mutation does not necessarily mean that there is an increased risk associated with this finding.

Furthermore, genotypes were differentially represented in our database with varying numbers of sequences. Genotypes A–D had between 1004 and 2700 sequences for each genotype; however, less sequences were available for the remaining genotypes (44 to 312). While this in large part reflects the global distribution, as previously reported [44], genotypes E–I sequences were underrepresented in our database. The differences in genotype frequency within sequenced isolates could be due to different prevalence in high vs. low income countries, as sequencing, at least when performed with modern technologies, is associated with significant cost. However, the relatively few sequences available for certain genotypes questions the precision of calculated frequencies. This is supported by the observation that in genotypes for which fewer sequences were available, many mutations were not found at all. Thus, low frequency mutations were possibly overlooked in these minority genotypes. This was even more relevant for the analysis of the regional distribution of HBV variants, as for some regions, relatively few sequences (between 20 and 2811) were available.

Another important factor to consider is that the majority of sequences were likely determined using classical Sanger sequencing of nucleic acids extracted from patient sera. The sensitivity of this approach is limited, with only variants occurring at a frequency of >20% within a viral population reliably detected [47]. Deciding which mutations are located on the same viral genome is much harder to achieve using next-generation techniques that traditionally have much shorter reads. Recent advances in next-generation sequencing (NGS) technology utilize longer reads and overlapping sequences, enabling the identification of mutations located on the same viral genome with high confidence [48]. NGS techniques are able to achieve a high sequence depth, readily detecting low frequency variants within a viral population. However, one needs to carefully evaluate the clinical relevance of these rare populations, which may simply be artefacts of high precision sequencing technologies.

A further possible reason for bias stems from difficulties in validating the quality of sequences included in our study. While information on sequence quality was in most cases not available, we tried to account for this by excluding sequences with database entries such as "unverified" or "non-functional" or which contained unspecified nucleotides (which could derive from poor sequencing quality). However, we cannot exclude other possible biases during sampling, sequencing, or database entry. Thus, conclusions based on low-frequency mutations should not be over interpreted.

For the prediction of HBV serotype, at least some confounding factors can be assumed to be less important, as it is less likely that the serotype influenced the likelihood of a certain variant to be sequenced. However, it is important to state that we only performed an in silico prediction and did not test the reactivity toward reference antibodies. While the relationship between amino acids at the relevant positions and the respective serotype is well established, we cannot exclude that some sequences would not show the predicted reactivity. Our finding that most genotypes have a distinct serotype confirms earlier observations in this regard [49]. We found that roughly 1/3rd of global HBV sequences showed either an adw2 or adrq+ serotype, with the other serotypes constituting the remaining 1/3rd of sequences. While serotypes are currently only playing a minor role in patient care, the information on occurrence of each serotype could still be helpful when designing future prophylactic or therapeutic vaccines or when developing antibody-based therapies (including antibodies to be passively administered but also bi-specific antibody constructs or chimeric antigen receptor T cells).

When analyzing the frequency of HBsAg mutants, we found that the overall frequency of such mutations was relatively low with no significant enrichment in any genotype. The only exception constituted mutations at position 127 (P to H or L), which are associated with occult HBV infection and

seemed to be the wild-type for genotypes E, F, and H. Importantly, amino acid position 127 of HBsAg is also one of the determinants for the HBV serotype, and variants with Leucine (L) at this position are found in the ayw4 and adw4 serotypes (see Table 1). We observed the V177A mutation which was quite evenly distributed throughout genotypes but showed a regional enrichment in Australia, New Zealand, Melanesia, Polynesia (together defined as Oceania). Interestingly, both the V177A and P127L mutations are determinants of the HBV serotype and together with positions 122 and 160 of HBsAg define the adrq-serotype. Thus, the finding that these "mutations" are associated with occult infection could also mean that diagnostic tests used in these studies had a deficiency in recognizing this serotype. However, we cannot disregard factors, such as ethnicity, in driving selection pressure leading to an enrichment of these variants in certain regions.

For nucleoside analogue resistance mutations, we found that M204V/I and L180M displayed the highest frequency by far, which is consistent with the fact that they mediate resistance to the majority of currently used nucleoside analogues for hepatitis B therapy, including lamivudine. The relatively low barrier to resistance associated with these mutations likely led to relatively high occurrences, especially in regions with broad access to antiviral therapy. Along this line, nucleoside analogue resistance mutations in general showed the highest frequency in high-income regions such as Northern America (especially L180M and M204V/I), Europe, and Eastern/Southern Asia. In contrast, lower frequencies were found on African continent, Central and South America and Asia. This could indicate that more patients had been under treatment in these regions, causing a selective pressure on the virus to generate resistance mutations.

Of further interest, three different mutations have been described to interfere with tenofovir susceptibility, yet only one of these was found in virus isolated from patients (A194T; [45]), whereas the other two were generated in vitro (P177G and F249A; [50]). While we were able to find the A194T mutation in several sequences of different genotypes, the P177G and F249A mutations were not identified in any of the more than 7000 sequences analyzed by us. Therefore, it is unlikely that they constitute clinically relevant variants, and we speculate that these mutations create fitness losses for the virus that outweigh the advantages of tenofovir resistance.

The most interesting observation regarding the BCP/pre-core region was that almost all genotype G sequences contained several mutations (C1653, T1753C, A1762T, G1764A, and G1896A) which are associated with HBeAg negativity and increased risk for fulminant hepatitis and/or HCC. Whilst the lack of HBeAg expression is well described for genotype G [51,52], the risk of HCC and fulminant hepatitis is not as clearly associated, and may be hard to elucidate, as genotype G is mostly found in co-infections with genotype A [53]. The finding that there was regional enrichment of BCP/pre-core mutations G1862T in the Caribbean and G1896A in Western Asia and Southern Europe was also of note, especially as few genotype G sequences were identified in these regions. However, relatively low number of sequences (Caribbean: 80; Western Asia: 146; Southern Europe: 69) were available for all three regions; thus, future studies should seek to investigate if the increased frequency of these mutations are genuine and more importantly whether an increased risk for fulminant hepatitis or HCC is conferred.

In summary, we present an overview of the frequency of clinically relevant HBV variants in publicly available HBV sequencing data. While the results should be interpreted with care as several confounding factors could potentially have influenced the frequency of individual mutations described here, our data should help to identify interesting research questions for future studies, as well as help to design and select suitable diagnostic tests and therapies.

Author Contributions: Conceptualization, S.V. and T.M.; methodology, S.V. and T.M.; software, S.V.; S.V. and T.M.; formal analysis, S.V.; investigation, S.V.; data curation, S.V.; writing—original draft preparation, S.V. and T.M.; writing—review and editing, T.M., U.P.; visualization, S.V. and T.M.; resources: U.P.; funding acquisition: U.P.; supervision, U.P., T.M. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by a clinical leave stipend by the Else-Kröner research college 'Microbial triggers as cause for disease' to T.M. and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project No. 272983813, via the Transregional Collaborative Research Center TRR179 which supplied funding to U.P.

Acknowledgments: We thank Peter Wing and James Harris from University of Oxford for proofreading the final version of the manuscript.

Conflicts of Interest: T.M. received a research grant by Gilead Sciences GmbH. T.M. and U.P. are ad hoc scientific advisors to VIR Biotechnology. U.P. is named as inventor on a patent application describing the therapeutic vaccination scheme TherVacB (PCT/EP2017/050553). U.P. and T.M. are named as inventors on combining siRNA with therapeutic vaccination (PCT/EP2018/028116). U.P. is a co-founder and shareholder of SCG Cell Therapy.

References

1. Lavanchy, D. Hepatitis B Virus Epidemiology, Disease Burden, Treatment, and Current and Emerging Prevention and Control Measures. *J. Viral Hepat.* **2004**, *11*, 97–107. [CrossRef] [PubMed]
2. WHO. Global Hepatitis Report. 2017. Available online: https://apps.who.int/iris/bitstream/handle/10665/255016/9789241565455-eng.pdf?sequence=1 (accessed on 1 February 2020).
3. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef] [PubMed]
4. Fako, V.; Wang, X.W. Molecular Carcinogenesis of HBV-Related HCC. In *Hepatitis B Virus and Liver Disease*; Kao, J.-H., Chen, D.-S., Eds.; Springer: Singapore, 2018; pp. 143–162.
5. Graber-Stiehl, I. The silent epidemic killing more people than HIV, malaria or TB. *Nature* **2018**, *564*, 24–26. [CrossRef] [PubMed]
6. Wang, L.; Zou, Z.-Q.; Wang, K. Clinical Relevance of HLA Gene Variants in HBV Infection. *J. Immunol. Res.* **2016**, *2016*, 9069375. [CrossRef] [PubMed]
7. Koc, O.M.; Robaeys, G.; Yildirim, B.; Posthouwer, D.; Hens, N.; Koek, G.H. The influence of ethnicity on disease outcome in patients with chronic hepatitis B infection. *J. Med. Virol.* **2019**, *91*, 623–629. [CrossRef]
8. Ganesan, M.; Eikenberry, A.; Poluektova, L.Y.; Kharbanda, K.K.; Osna, N.A. Role of alcohol in pathogenesis of hepatitis B virus infection. *World J. Gastroenterol.* **2020**, *26*, 883–903. [CrossRef]
9. Chu, Y.J.; Yang, H.I.; Wu, H.C.; Liu, J.; Wang, L.Y.; Lu, S.N.; Lee, M.H.; Jen, C.L.; You, S.L.; Santella, R.M.; et al. Aflatoxin B(1) exposure increases the risk of cirrhosis and hepatocellular carcinoma in chronic hepatitis B virus carriers. *Int. J. Cancer* **2017**, *141*, 711–720. [CrossRef]
10. Asabe, S.; Wieland, S.F.; Chattopadhyay, P.K.; Roederer, M.; Engle, R.E.; Purcell, R.H.; Chisari, F.V. The size of the viral inoculum contributes to the outcome of hepatitis B virus infection. *J. Virol.* **2009**, *83*, 9652–9662. [CrossRef]
11. Revill, P.A.; Tu, T.; Netter, H.J.; Yuen, L.K.W.; Locarnini, S.A.; Littlejohn, M. The Evolution and Clinical Impact of Hepatitis B Virus Genome Diversity. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 618–634. [CrossRef]
12. Wong, G.L.-H.; Chan, H.L.-Y.; Yiu, K.K.-L.; Lai, J.W.-Y.; Chan, V.K.-K.; Cheung, K.K.-C.; Wong, E.W.-N.; Wong, V.W.-S. Meta-analysis: The association of hepatitis B virus genotypes and hepatocellular carcinoma. *Aliment. Pharmacol. Ther.* **2013**, *37*, 517–526. [CrossRef]
13. Kramvis, A. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **2014**, *57*, 141–150. [CrossRef] [PubMed]
14. Tatematsu, K.; Tanaka, Y.; Kurbanov, F.; Sugauchi, F.; Mano, S.; Maeshiro, T.; Nakayoshi, T.; Wakuta, M.; Miyakawa, Y.; Mizokami, M. 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*, 10538–10547. [CrossRef] [PubMed]
15. Locarnini, S.; Littlejohn, M.; Aziz, M.N.; Yuen, L. Possible origins and evolution of the hepatitis B virus (HBV). *Semin. Cancer Biol.* **2013**, *23*, 561–575. [CrossRef] [PubMed]
16. Caligiuri, P.; Cerruti, R.; Icardi, G.; Bruzzone, B. Overview of hepatitis B virus mutations and their implications in the management of infection. *World J. Gastroenterol.* **2016**, *22*, 145–154. [CrossRef] [PubMed]
17. Leong, J.; Lin, D.; Nguyen, M.H. Hepatitis B surface antigen escape mutations: Indications for initiation of antiviral therapy revisited. *World J. Clin. Cases* **2016**, *4*, 71–75. [CrossRef]

18. Shaw, T.; Bartholomeusz, A.; Locarnini, S. HBV drug resistance: Mechanisms, detection and interpretation. *J. Hepatol.* **2006**, *44*, 593–606. [CrossRef]
19. Buckwold, V.E.; Xu, Z.; Chen, M.; Yen, T.S.; Ou, J.H. Effects of a naturally occurring mutation in the hepatitis B virus basal core promoter on precore gene expression and viral replication. *J. Virol.* **1996**, *70*, 5845–5851. [CrossRef]
20. Xu, Z.; Ren, X.; Liu, Y.; Li, X.; Bai, S.; Zhong, Y.; Wang, L.; Mao, P.; Wang, H.; Xin, S.; et al. Association of hepatitis B virus mutations in basal core promoter and precore regions with severity of liver disease: An investigation of 793 Chinese patients with mild and severe chronic hepatitis B and acute-on-chronic liver failure. *J. Gastroenterol.* **2011**, *46*, 391–400. [CrossRef]
21. Gao, S.; Duan, Z.-P.; Coffin, C.S. Clinical relevance of hepatitis B virus variants. *World J. Hepatol.* **2015**, *7*, 1086–1096. [CrossRef]
22. Yano, Y.; Azuma, T.; Hayashi, Y. Variations and mutations in the hepatitis B virus genome and their associations with clinical characteristics. *World J. Hepatol.* **2015**, *7*, 583–592. [CrossRef]
23. Tong, S.; Revill, P. Overview of hepatitis B viral replication and genetic variability. *J. Hepatol.* **2016**, *64*, S4–S16. [CrossRef] [PubMed]
24. Zhang, X.; Chen, X.; Wei, M.; Zhang, C.; Xu, T.; Liu, L.; Xu, Z. Potential resistant mutations within HBV reverse transcriptase sequences in nucleos(t)ide analogues-experienced patients with hepatitis B virus infection. *Sci. Rep.* **2019**, *9*, 8078. [CrossRef] [PubMed]
25. Gupta, N.; Goyal, M.; Wu, C.H.; Wu, G.Y. The Molecular and Structural Basis of HBV-resistance to Nucleos(t)ide Analogs. *J. Clin. Transl. Hepatol.* **2014**, *2*, 202–211.
26. Colagrossi, L.; Hermans, L.E.; Salpini, R.; Di Carlo, D.; Pas, S.D.; Alvarez, M.; Ben-Ari, Z.; Boland, G.; Bruzzone, B.; Coppola, N.; et al. Immune-escape mutations and stop-codons in HBsAg develop in a large proportion of patients with chronic HBV infection exposed to anti-HBV drugs in Europe. *BMC Infect. Dis.* **2018**, *18*, 251. [CrossRef] [PubMed]
27. Pacheco, S.R.; Dos Santos, M.I.M.A.; Stocker, A.; Zarife, M.A.S.A.; Schinoni, M.I.; Paraná, R.; Dos Reis, M.G.; Silva, L.K. Genotyping of HBV and tracking of resistance mutations in treatment-naïve patients with chronic hepatitis B. *Infect. Drug Resist.* **2017**, *10*, 201–207. [CrossRef] [PubMed]
28. Neumann-Fraune, M.; Beggel, B.; Pfister, H.; Kaiser, R.; Verheyen, J. High frequency of complex mutational patterns in lamivudine resistant hepatitis B virus isolates. *J. Med. Virol.* **2013**, *85*, 775–779. [CrossRef]
29. Ismail, A.M.; Sharma, O.P.; Kumar, M.S.; Kannangai, R.; Abraham, P. Impact of rtI233V mutation in hepatitis B virus polymerase protein and adefovir efficacy: Homology modeling and molecular docking studies. *Bioinformation* **2013**, *9*, 121–125. [CrossRef]
30. Meier-Stephenson, V.; Bremner, W.T.R.; Dalton, C.S.; van Marle, G.; Coffin, C.S.; Patel, T.R. Comprehensive Analysis of Hepatitis B Virus Promoter Region Mutations. *Viruses* **2018**, *10*, 603. [CrossRef]
31. Wagner, J.; Yuen, L.; Littlejohn, M.; Sozzi, V.; Jackson, K.; Suri, V.; Tan, S.; Feierbach, B.; Gaggar, A.; Marcellin, P.; et al. Analysis of Hepatitis B virus haplotype diversity detects striking sequence conservation across genotypes and chronic disease phase. *Hepatology* **2020**. [CrossRef]
32. Hao, R.; Xiang, K.; Shi, Y.; Zhao, D.; Tian, H.; Xu, B.; Zhu, Y.; Dong, H.; Ding, H.; Zhuang, H.; et al. Naturally Occurring Mutations within HBV Surface Promoter II Sequences Affect Transcription Activity, HBsAg and HBV DNA Levels in HBeAg-Positive Chronic Hepatitis B Patients. *Viruses* **2019**, *11*, 78. [CrossRef]
33. Phan, N.M.H.; Faddy, H.; Flower, R.; Spann, K.; Roulis, E. In silico Analysis of Genetic Diversity of Human Hepatitis B Virus in Southeast Asia, Australia and New Zealand. *Viruses* **2020**, *12*, 427. [CrossRef] [PubMed]
34. Bahar, M.; Pervez, M.T.; Ali, A.; Babar, M.E. In Silico Analysis of Hepatitis B Virus Genotype D Subgenotype D1 Circulating in Pakistan, China, and India. *Evol. Bioinform.* **2019**, *15*, 1176934319861337. [CrossRef] [PubMed]
35. Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2016**, *44*, D67–D72. [CrossRef] [PubMed]
36. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
37. Pourkarim, M.R.; Amini-Bavil-Olyaee, S.; Kurbanov, F.; Van Ranst, M.; Tacke, F. Molecular identification of hepatitis B virus genotypes/subgenotypes: Revised classification hurdles and updated resolutions. *World J. Gastroenterol.* **2014**, *20*, 7152–7168. [CrossRef] [PubMed]

38. Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441. [CrossRef] [PubMed]
39. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
40. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
41. Norder, H.; Courouce, A.M.; Magnius, L.O. Molecular basis of hepatitis B virus serotype variations within the four major subtypes. *J. Gen. Virol.* **1992**, *73*, 3141–3145. [CrossRef]
42. Lazarevic, I. Clinical implications of hepatitis B virus mutations: Recent advances. *World J. Gastroenterol.* **2014**, *20*, 7653–7664. [CrossRef]
43. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **2019**, *35*, 4453–4455. [CrossRef] [PubMed]
44. Velkov, S.; Ott, J.J.; Protzer, U.; Michler, T. The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. *Genes* **2018**, *9*, 495. [CrossRef] [PubMed]
45. Amini-Bavil-Olyaee, S.; Herbers, U.; Sheldon, J.; Luedde, T.; Trautwein, C.; Tacke, F. The rtA194T polymerase mutation impacts viral replication and susceptibility to tenofovir in hepatitis B e antigen-positive and hepatitis B e antigen-negative hepatitis B virus strains. *Hepatology* **2009**, *49*, 1158–1165. [CrossRef]
46. Qin, B.; Budeus, B.; Cao, L.; Wu, C.; Wang, Y.; Zhang, X.; Rayner, S.; Hoffmann, D.; Lu, M.; Chen, X. The amino acid substitutions rtP177G and rtF249A in the reverse transcriptase domain of hepatitis B virus polymerase reduce the susceptibility to tenofovir. *Antivir. Res.* **2013**, *97*, 93–100. [CrossRef] [PubMed]
47. Fu, Y.; Zeng, Y.; Chen, T.; Chen, H.; Lin, N.; Lin, J.; Liu, X.; Huang, E.; Wu, S.; Wu, S.; et al. Characterization and Clinical Significance of Natural Variability in Hepatitis B Virus Reverse Transcriptase in Treatment-Naive Chinese Patients by Sanger Sequencing and Next-Generation Sequencing. *J. Clin. Microbiol.* **2019**, *57*, e00119-19. [CrossRef]
48. McNaughton, A.L.; Roberts, H.E.; Bonsall, D.; de Cesare, M.; Mokaya, J.; Lumley, S.F.; Golubchik, T.; Piazza, P.; Martin, J.B.; de Lara, C.; et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci. Rep.* **2019**, *9*, 7081. [CrossRef]
49. Liu, C.J.; Kao, J.H.; Chen, P.J.; Lai, M.Y.; Chen, D.S. Molecular epidemiology of hepatitis B viral serotypes and genotypes in taiwan. *J. Biomed. Sci.* **2002**, *9*, 166–170. [CrossRef]
50. Chan, S.L.; Wong, V.W.; Qin, S.; Chan, H.L. Infection and cancer: The case of hepatitis B. *J. Clin. Oncol.* **2015**, *34*, 83–90. [CrossRef]
51. Stuyver, L.; De Gendt, S.; Van Geyt, C.; Zoulim, F.; Fried, M.; Schinazi, R.F.; Rossau, R. A new genotype of hepatitis B virus: Complete genome and phylogenetic relatedness. *J. Gen. Virol.* **2000**, *81*, 67–74. [CrossRef]
52. Zaaijer, H.L.; Boot, H.J.; van Swieten, P.; Koppelman, M.H.; Cuypers, H.T. HBsAg-negative mono-infection with hepatitis B virus genotype G. *J. Viral Hepat.* **2011**, *18*, 815–819. [CrossRef]
53. Osiowy, C.; Gordon, D.; Borlang, J.; Giles, E.; Villeneuve, J.P. Hepatitis B virus genotype G epidemiology and co-infection with genotype A in Canada. *J. Gen. Virol.* **2008**, *89*, 3009–3015. [CrossRef] [PubMed]

viruses

MDPI

Article

HDVdb: A Comprehensive Hepatitis D Virus Database

Zainab Usman [1,†], Stoyan Velkov [2,†], Ulrike Protzer [2], Michael Roggendorf [2], Dmitrij Frishman [1] and Hadi Karimzadeh [2,3,*]

1 Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany; zainabasrar@gmail.com (Z.U.); d.frishman@wzw.tum.de (D.F.)
2 Institute of Virology, Technische Universität München, 81675 Munich, Germany; stoyan.velkov@tum.de (S.V.); protzer@tum.de (U.P.); michael.roggendorf@tum.de (M.R.)
3 Division of Clinical Pharmacology, University Hospital, LMU Munich, 80337 Munich, Germany
* Correspondence: Hadi.Karimzadeh@med.uni-muenchen.de
† These authors contributed equally to this work.

Received: 26 March 2020; Accepted: 11 May 2020; Published: 14 May 2020

Abstract: Hepatitis D virus (HDV) causes the most severe form of viral hepatitis, which may rapidly progress to liver cirrhosis and hepatocellular carcinoma (HCC). It has been estimated that 15–20 million people worldwide are suffering from the chronic HDV infection. Currently, no effective therapies are available to treat acute or chronic HDV infection. The remarkable sequence variability of the HDV genome, particularly within the hypervariable region has resulted in the provisional classification of eight major genotypes and various subtypes. We have developed a specialized database, HDVdb, which contains a collection of partial and complete HDV genomic sequences obtained from the GenBank and from our own patient cohort. HDVdb enables the researchers to investigate the genetic variability of all available HDV sequences, correlation of genotypes to epidemiology and pathogenesis. Additionally, it will contribute in understanding the drug resistant mutations and develop effective vaccines against HDV infection. The database can be accessed through a web interface that allows for static and dynamic queries and offers integrated generic and specialized sequence analysis tools, such as annotation, genotyping, primer prediction, and phylogenetic analyses.

Keywords: hepatitis delta virus; database; genotyping; Webserver

1. Introduction

Hepatitis D virus (HDV) infection remains the most difficult-to-treat form of viral hepatitis, affecting 15–20 million patients worldwide with chronic hepatitis, liver cirrhosis and hepatocellular carcinoma (HCC) [1]. The HDV infection in humans occurs so far only together with hepatitis B virus (HBV) because HDV needs the envelope proteins from HBV to complete its life cycle. Therefore, two main forms of HDV infection have been described: (1) coinfection; with a high rate of viral clearance in adults similar to HBV mono-infection [2], or (2) super-infection in the presence of a pre-existing HBV infection. The latter results in a persistent chronic HDV infection in 70–90% of the cases and is associated with an early risk to develop cirrhosis and HCC [3]. The current anti-HDV therapy is mainly based on administration of interferon with a very low response rate in patients [4,5] and high chance of relapse upon discontinuation [6]. Nevertheless, efforts have been made recently to develop new anti-HDV drugs to treat chronic HDV infection, with promising results in the clinical trials [7–11].

HDV is a small, spherical virus of 35–37 nm in diameter, with an envelope containing the hepatitis B surface antigen (HBsAg), which surrounds the genomic RNA-nucleoprotein complex [12]. The genome is a negative sense single-stranded RNA (1.67 kb), whose complementary strand (antigenomic RNA)

contains one single functional open reading frame (ORF) encoding two isoforms of the hepatitis delta antigen (HDAg), the small (*S*-HDAg, 195 aa) and the large (*L*-HDAg, 214 aa) [13,14]. The sequence encoding these isoform proteins resides in the antigenomic RNA, which, as a result of the cellular editing activity of ADAR-1, modifies the amber stop codon (UAG) to (UGG) of *S*-HDAg, resulting in the extension of the amino acid sequences by 19–20 aa at the C terminus [15].

HDV RNA sequences identified so far have been classified into eight known genotypes (HDV-1–8) based both on the nucleotide and amino acid sequences of the coding region of HDAg [16]. These genotypes are distributed across different geographical regions. In our recent studies, we identified and introduced different subtypes for the genotype 1, genotype 2 and genotype 4 [17]. Subgrouping of the so far identified HDV genotypes into distinct clusters or subtypes has been also suggested by others [18,19]. These data provide a clearer picture of the geographical a global distribution of HDV isolates. However, it is not known how these subtypes correlate with the clinical manifestation and response to therapy.

HDV-1 is the most geographically widespread genotype distributed across major regions such as Europe, Middle East, East Asia, America and Africa; whereas all other genotypes (HDV-2–8) are associated with distinct geographical and ethnic regions. HDV-2 and HDV-4 are found in North Asia and East Asia, respectively [20–23]; HDV-3 is exclusively found in the north part of South America (Brazil, Peru, Colombia, Argentina, Ecuador and Venezuela) [24–28] and HDV-5 to HDV-8 were previously described to be found "only" in Africa [29,30], however, a recent study reported HDV-8 isolates from Northeast Brazil, which presumably crossed the ocean through slave trades in the 16–18th centuries [31]. In humans, HDV infections with different genotypes exhibit different clinical courses and outcomes. For instance, HDV-1 strains show a broad spectrum of virulent and pathogenic phenotypes [32], HDV-2 (and HDV-4) cause milder forms of liver disease [33], whereas HDV-3 isolates are associated with outbreaks of fulminant hepatitis in South America [24]. The pathogenic properties of HDV 5–8 isolates are not well characterized [34].

For decades, HDV has been thought to have evolved in humans in combination with HBV as a helper virus providing a viral envelope to form infectious particles. Recently, however, HDV like sequences have been identified in a variety of animals and insects [35–37]. Sequence analysis in ducks revealed an approximately 1700-nt circular RNA genome with self-complementary, unbranched rod-like structures, and coiled-coil domains [36]. The predicted HDV-like protein discovered in ducks shares 32% amino acid similarity with the small delta antigen (*S*-HDAg) of the human HDV (hHDV).

This discovery of an HDV-like agent in ducks was followed by the identification of a deltavirus in snakes (*Boa constrictor*), designated as snake HDV (sHDV) [37]. Sequence comparison of the snake delta antigen (sHDAg) showed that its aa sequence is 55% identical to its human counterparts. Anti-sera raised against a recombinant sHDAg was used in immunohistology studies. A broad viral target was demonstrated in different snake cells, including neurons, epithelial cells and leukocytes. The duck and snake viruses constitute divergent phylogenetic lineages as compared to the human HDV (hHDV), which so far seem quite distant related to the known human isolates.

Using additional meta-transcriptomic data, highly divergent HDV-like viruses were also found to be present in fish, amphibians and invertebrates. These newly identified viruses share human HDV-like genomic features such as a small genome size of 1.7 kb in length [35].

The identification of a much broader range of hosts as initially anticipated and the fact that the HDV RNA genome can efficiently replicate in different tissues and species, raise the possibility that HDV is able to be transmitted independently of HBV. Perez-Vargas J. et al. [38] have shown, that envelope glycoproteins (GPs) of unrelated viruses can act as helper viruses for HDV including vesiculovirus, flavivirus and hepacivirus. These GPs can package HDV RNPs, allowing efficient egress of HDV particles in the extracellular milieu of coinfected cells and subsequent entry into the cells expressing the corresponding receptors. In vivo studies in humanized mice indicate that HDV RNPs packaged into an HCV envelope can propagate HDV infection in the liver of coinfected mice [38].

In recent years, the amount of HDV genomic data has increased exponentially. Intensive sequencing efforts have resulted in approximately 2621 nucleotide HDV sequences (partial and full length) deposited into the DDBJ, EMBL and GenBank databases. The GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC), which comprises of the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA) and GenBank at NCBI. Those three organizations are synchronized and exchange data on a daily basis. Therefore, the sequences dataset can be retrieved using any of the platforms, i.e., the GenBank database. In order to exploit this large and growing collection of sequences efficiently and to facilitate sequence analysis we sought to develop a specialized database. Databases established for other types of viruses, in particular for HIV [39], HBV [40] and HCV [41] have proved to be very helpful for epidemiological and clinical studies, more importantly in characterizing resistance to direct anti-viral drugs. Here, we present the hepatitis delta virus database (HDVdb; http://hdvdb.bio.wzw.tum.de/). This comprehensive database collates HDV sequences and is mainly oriented towards the sequence analysis of HDV isolates, including the complete viral genomic sequences, large and small HDV antigen sequences (*L*-HDAg and *S*-HDAg, respectively). HDVdb provides a platform for genotyping and phylogenetic analyses including prediction of HDV genotypes for user-supplied HDV sequence entries. Moreover, the database will help in identifying the emerging variants related to immune escape from the B and T cell response as described recently [42,43] and in detecting therapy resistant variants across different HDV genotypes, which can be correlated with clinical studies.

2. Materials and Methods

The HDVdb building process began with the manual retrieval of all the HDV entries using the keyword: "hepatitis delta virus" from GenBank hosted at NCBI [44]. All results corresponding to taxon "Hepatitis delta virus" (taxid: 12475) were considered. Currently, a total of 2621 hepatitis delta virus nucleotide sequences are deposited into GenBank. These entries contain full sequence records of both HDV "complete genomic" sequences and subgenomic fragments (*S*-HDAg; 1–195 aa and *L*-HDAg; 1–214 aa) as well as partial cds sequences. GenBank entries containing complete HDV protein sequences were also incorporated. Majority of the sequences were retained to provide the maximum data information to our visitors, however, sequences shorter than 90 bases were not included into the dataset. The sequence dataset was than parsed by creating an automatic pipeline using Java programming language to extract essential information for each accession number such as strain name, genotype, country and date. In addition, 152 sequences lacking the genotype information were assigned to a genotype by performing similarity search using BLAST [45].

The HDVdb web interface is hosted using Apache HTTP server and runs on PHP Laravel framework. The HDVdb is updated on an annual basis. The software for the automatic annotation, as well as for the querying and the managing the database is implemented in Java and Bash programming languages. It makes use of all defined genotypes and their subgenotypes. The workflow of database construction is schematically demonstrated in Figure 1.

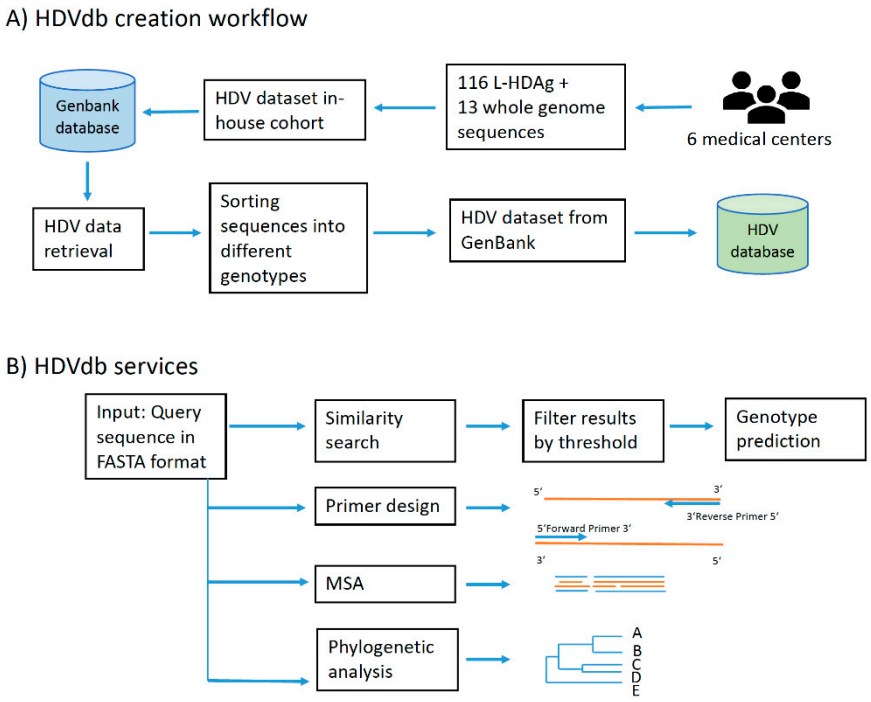

Figure 1. HDVdb construction and analysis workflow. (**A**) Building blocks of HDVdb based on publicly available and in-house isolates. (**B**) List of services available at the HDVdb. In Primer design, the orange lines represent template and blue lines represent the primers. In MSA (Multiple sequence alignment): blue and orange lines represent different aligned sequences, schematically.

3. Results and Discussion

The HDVdb is accessible online through the website: http://hdvdb.bio.wzw.tum.de/. HDVdb contains entries for human hepatitis delta virus sequences, with 512 complete genome sequences, as well as 1066 *L*-HDAg and *S*-HDAg and 1281 partial cds nucleotide sequences as well as protein sequences for *L*-HDAg and *S*-HDAg. These sequences can be directly downloaded from the database for any further analysis. Links to protein sequences for both *L*-HDAg and *S*-HDAg sequences are directly provided at the home page. Additionally, we included 13 complete genome (Accession MH457142-MH457154) and 116 *L*-HDAg sequences (Accession MF175257-MF175360, MH447633-MH447644) retrieved from six different medical centers of our European study cohort [17]. In this study, sequence conservation at each position across the entire length of the 322 multiply aligned genome sequences (i.e., genotype-1) was visualized. The multiple sequence alignments were performed using MUSCLE v3.8.5551 [46] whereas the evaluation was performed using customized Ruby scripts (Figure 2). We concluded that despite low conservation rate throughout the HDV genome, there were no significant differences on genotyping results using the whole genome or the *L*-HDAg encoding region.

The HDVdb is divided into a static and a dynamic part as demonstrated in Figure 3. The static part allows the user to access the general information about HDV. The homepage provides a data summary of updated number of *S*-HDAg, *L*-HDAg and complete genomes of all the eight known genotypes on the database. The user can retrieve pre-compiled protein and nucleotide datasets for complete genome, *L*-HDAg, and *S*-HDAg separately for each genotype, alternatively the user can also download a single FASTA file containing these datasets for all genotypes. In addition, the database also provides a tutorial to help the users with necessary technical information required to access tools available on the database.

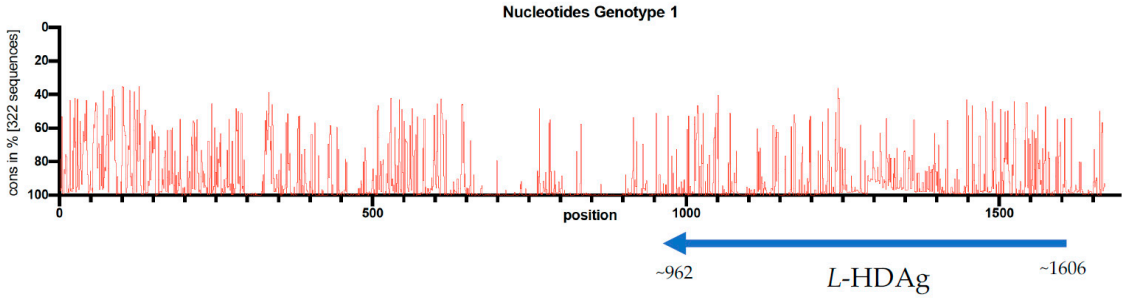

Figure 2. Conservation rate of 322 full genome nucleotide sequences of the HDV genotype 1. The single open reading frame of HDAg is located on antigenomic strand between position 962 and 1606 and indicated with an arrow under the genome.

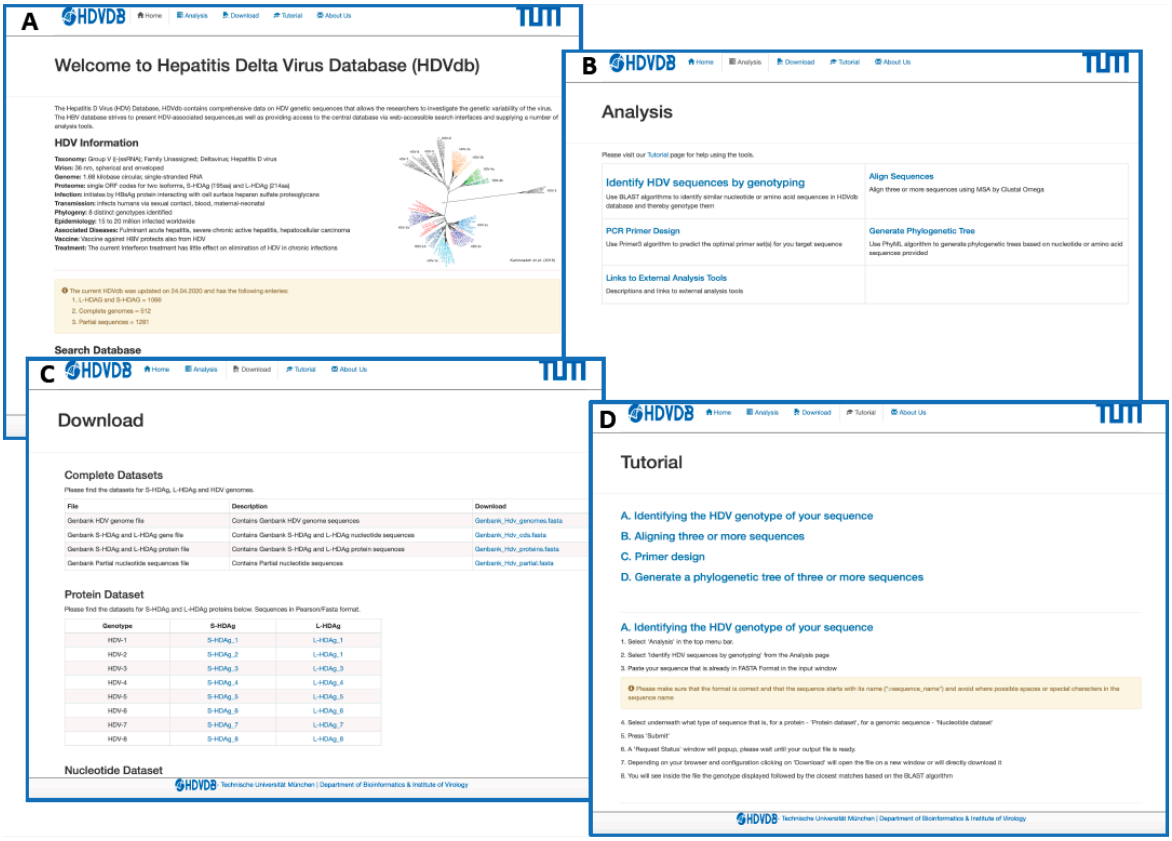

Figure 3. Web interface of the HDVdb. (**A**) The homepage, with the menu bar and all the menus repeated as buttons for ease-of-use. The page summarizes the characteristics and statistics of HDV. The nucleotide as well as protein datasets can be directly accessed through the homepage. The sequence files can be viewed for each genotype both in protein and nucleotide. (**B**) List of tools available on the website to analyze HDV related sequences. (**C**) List of sequence datasets available in FASTA format and updated on a regular basis. (**D**) A snapshot of the tutorial page with step-by-step instructions on how to use different tools available on this database.

The dynamic part allows the analysis of user-provided queries. The homepage presents an interactive search box that allows the visitors of our database with options to search sequences based on accession number, genotype, country and date for protein, complete genome, coding sequences for *L*-HDAg and *S*-HDAg, as well as partial sequences. The nucleotide and protein sequences queries can be genotyped using "Identify HDV sequences by genotyping" option. The webservice uses BLAST [45], which performs local alignments and scores the most relevant sequences to access the query genotype. A minimum of 75% identity score against the database is required to classify the query sequence to one

of the HDV genotypes. This threshold prevents the false positives to be classified and is based on our previous research [17].

Furthermore, we integrated computational tools for multiple sequence analysis (Clustal Omega, version 1.2.3 [47]), primer design (Primer3, version 2.3.7 [48]) and phylogenetic analyses (Phylip (PhyML), version 3.696 [49]), Figure 3. The user can also graphically visualize the phylogenetic trees on completion generated by FigTree, version 1.4.4. The request and response from these services was handled using PHP Laravel framework and bash scripting.

4. Conclusions

Hepatitis D has received a lot of attention in recent years, resulting in a flood of new findings and information, including next generation sequencing data. However, a platform capable of collecting and analyzing this growing body of data has so far been missing. Here we introduced HDVdb as a comprehensive database of human HDV sequences with a potential of expansion to the recently identified isolates from animals and insects. HDVdb allows the user to download structured data of all known HDV sequences. It also permits the user to use this data and perform comparative sequence analysis using multiple bioinformatics services available directly on HDVdb website.

Author Contributions: H.K. and M.R. conceived the idea of study. Z.U. extracted the data, created the webservices, and developed the database. Z.U. and S.V. implemented software services. Z.U., S.V., U.P., M.R., D.F. and H.K. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by internal funds of the Medical Faculty of the Technical University of Munich and by Registry Grant from European Association for the Study of the Liver (EASL). Z.U. was supported by a grant (57129429) from the German Academic Exchange Service (DAAD).

Acknowledgments: We thank all the staff working at the diagnostic sections of the Institute of Virology of the Technical University of Munich for their technical assistance. We would also like to thank the network and system administrators both at TUM Department of Bioinformatics who helped make the service available online and accessible for all.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chen, H.Y.; Shen, D.T.; Ji, D.Z.; Han, P.C.; Zhang, W.M.; Ma, J.F.; Chen, W.S.; Goyal, H.; Pan, S.; Xu, H.G. Prevalence and burden of hepatitis D virus infection in the global population: A systematic review and meta-analysis. *Gut* **2018**. [CrossRef] [PubMed]
2. Noureddin, M.; Gish, R. Hepatitis delta: Epidemiology, diagnosis and management 36 years after discovery. *Curr. Gastroenterol. Rep.* **2014**, *16*, 365. [CrossRef] [PubMed]
3. Smedile, A.; Farci, P.; Verme, G.; Caredda, F.; Cargnel, A.; Caporaso, N.; Dentico, P.; Trepo, C.; Opolon, P.; Gimson, A.; et al. Influence of delta infection on severity of hepatitis B. *Lancet* **1982**, *2*, 945–947. [CrossRef]
4. Wedemeyer, H.; Yurdaydìn, C.; Dalekos, G.N.; Erhardt, A.; Çakaloğlu, Y.; Değertekin, H.; Gürel, S.; Zeuzem, S.; Zachou, K.; Bozkaya, H.; et al. Peginterferon plus adefovir versus either drug alone for hepatitis delta. *N. Engl. J. Med.* **2011**, *364*, 322–331. [CrossRef]
5. Wedemeyer, H.; Yurdaydin, C.; Hardtke, S.; Caruntu, F.A.; Curescu, M.G.; Yalcin, K.; Akarca, U.S.; Gurel, S.; Zeuzem, S.; Erhardt, A.; et al. Peginterferon alfa-2a plus tenofovir disoproxil fumarate for hepatitis D (HIDIT-II): A randomised, placebo controlled, phase 2 trial. *Lancet Infect. Dis.* **2019**, *19*, 275–286. [CrossRef]
6. Heidrich, B.; Yurdaydin, C.; Kabacam, G.; Ratsch, B.A.; Zachou, K.; Bremer, B.; Dalekos, G.N.; Erhardt, A.; Tabak, F.; Yalcin, K.; et al. Late HDV RNA relapse after peginterferon alpha-based therapy of chronic hepatitis delta. *Hepatology* **2014**, *60*, 87–97. [CrossRef]
7. Bogomolov, P.; Alexandrov, A.; Voronkova, N.; Macievich, M.; Kokina, K.; Petrachenkova, M.; Lehr, T.; Lempp, F.A.; Wedemeyer, H.; Haag, M.; et al. Treatment of chronic hepatitis D with the entry inhibitor myrcludex B: First results of a phase Ib/IIa study. *J. Hepatol.* **2016**, *65*, 490–498. [CrossRef]

8. Beilstein, F.; Blanchet, M.; Vaillant, A.; Sureau, C. Nucleic acid polymers are active against hepatitis delta virus infection in vitro. *J. Virol.* **2018**, *92*. [CrossRef]
9. Zhang, Z.; Filzmayer, C.; Ni, Y.; Sultmann, H.; Mutz, P.; Hiet, M.S.; Vondran, F.W.R.; Bartenschlager, R.; Urban, S. Hepatitis D virus replication is sensed by MDA5 and induces IFN-beta/lambda responses in hepatocytes. *J. Hepatol.* **2018**, *69*, 25–35. [CrossRef]
10. Bazinet, M.; Pantea, V.; Cebotarescu, V.; Cojuhari, L.; Jimbei, P.; Albrecht, J.; Schmid, P.; Le Gal, F.; Gordien, E.; Krawczyk, A.; et al. Safety and efficacy of REP 2139 and pegylated interferon alfa-2a for treatment-naive patients with chronic hepatitis B virus and hepatitis D virus co-infection (REP 301 and REP 301-LTF): A non-randomised, open-label, phase 2 trial. *Lancet Gastroenterol. Hepatol.* **2017**, *2*, 877–889. [CrossRef]
11. Lempp, F.A.; Urban, S. Hepatitis delta virus: Replication strategy and upcoming therapeutic options for a neglected human pathogen. *Viruses* **2017**, *9*, 172. [CrossRef] [PubMed]
12. Lai, M.M. The molecular biology of hepatitis delta virus. *Annu. Rev. Biochem.* **1995**, *64*, 259–286. [CrossRef] [PubMed]
13. Huang, C.R.; Lo, S.J. Evolution and diversity of the human hepatitis d virus genome. *Adv. Bioinform.* **2010**. [CrossRef] [PubMed]
14. Weiner, A.J.; Choo, Q.L.; Wang, K.S.; Govindarajan, S.; Redeker, A.G.; Gerin, J.L.; Houghton, M. A single antigenomic open reading frame of the hepatitis delta virus encodes the epitope(s) of both hepatitis delta antigen polypeptides p24 delta and p27 delta. *J. Virol.* **1988**, *62*, 594–599. [CrossRef]
15. Casey, J.L. RNA editing in hepatitis delta virus. *Curr. Top. Microbiol. Immunol.* **2006**, *307*, 67–89.
16. Le Gal, F.; Gault, E.; Ripault, M.P.; Serpaggi, J.; Trinchet, J.C.; Gordien, E.; Deny, P. Eighth major clade for hepatitis delta virus. *Emerg. Infect. Dis.* **2006**, *12*, 1447–1450. [CrossRef]
17. Karimzadeh, H.; Usman, Z.; Frishman, D.; Roggendorf, M. Genetic diversity of hepatitis D virus genotype-1 in Europe allows classification into subtypes. *J. Viral Hepat.* **2019**. [CrossRef]
18. Le Gal, F.; Brichler, S.; Drugan, T.; Alloui, C.; Roulot, D.; Pawlotsky, J.M.; Deny, P.; Gordien, E. Genetic diversity and worldwide distribution of the deltavirus genus: A study of 2,152 clinical strains. *Hepatology* **2017**, *66*, 1826–1841. [CrossRef]
19. Paul Dény, C.D.; Gatherer, D.; Kay, A.C.; Le Gal, F.; Drugan, T. Hepatitis delta virus clades and genotypes: A practical approach. In *Hepatitis D. Virology, Management and Methodology*; Mario Rizzetto, A.S., Ed.; Il Pensiero Scientifico: Rome, Italy, 2019.
20. Ivaniushina, V.; Radjef, N.; Alexeeva, M.; Gault, E.; Semenov, S.; Salhi, M.; Kiselev, O.; Deny, P. Hepatitis delta virus genotypes I and II cocirculate in an endemic area of Yakutia, Russia. *J. Gen. Virol.* **2001**, *82*, 2709–2718. [CrossRef]
21. Lee, C.M.; Changchien, C.S.; Chung, J.C.; Liaw, Y.F. Characterization of a new genotype II hepatitis delta virus from Taiwan. *J. Med. Virol.* **1996**, *49*, 145–154. [CrossRef]
22. Sakugawa, H.; Nakasone, H.; Nakayoshi, T.; Kawakami, Y.; Miyazato, S.; Kinjo, F.; Saito, A.; Ma, S.P.; Hotta, H.; Kinoshita, M. Hepatitis delta virus genotype IIb predominates in an endemic area, Okinawa, Japan. *J. Med. Virol.* **1999**, *58*, 366–372. [CrossRef]
23. Marino, Q.; Cisse, M.; Gerber, A.; Dolo, O.; Sayon, S.; Ba, A.; Brichler, S.; Tata Traore, F.; Gordien, E.; Togo, J.; et al. Low hepatitis D seroprevalence in blood donors of Bamako, Mali. *Infect. Dis. (Lond.)* **2019**, *51*, 622–624. [CrossRef] [PubMed]
24. Casey, J.L.; Brown, T.L.; Colan, E.J.; Wignall, F.S.; Gerin, J.L. A genotype of hepatitis D virus that occurs in northern South America. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 9016–9020. [CrossRef]
25. Gomes-Gouvea, M.S.; Soares, M.C.; Bensabath, G.; de Carvalho-Mello, I.M.; Brito, E.M.; Souza, O.S.; Queiroz, A.T.; Carrilho, F.J.; Pinho, J.R. Hepatitis B virus and hepatitis delta virus genotypes in outbreaks of fulminant hepatitis (Labrea black fever) in the western Brazilian Amazon region. *J. Gen. Virol.* **2009**, *90*, 2638–2643. [CrossRef]
26. Quintero, A.; Uzcategui, N.; Loureiro, C.L.; Villegas, L.; Illarramendi, X.; Guevara, M.E.; Ludert, J.E.; Blitz, L.; Liprandi, F.; Pujol, F.H. Hepatitis delta virus genotypes I and III circulate associated with hepatitis B virus genotype F in Venezuela. *J. Med. Virol.* **2001**, *64*, 356–359. [CrossRef]
27. Scarponi, C.F.O.; Silva, R.D.N.D.; Souza Filho, J.A.; Guerra, M.R.L.; Pedrosa, M.A.F.; Mol, M.P.G. Hepatitis delta prevalence in South America: A systematic review and meta-analysis. *Rev. Soc. Bras. Med. Trop.* **2019**, *52*, e20180289. [CrossRef]

28. di Filippo Villa, D.; Cortes-Mancera, F.; Payares, E.; Montes, N.; de la Hoz, F.; Arbelaez, M.P.; Correa, G.; Navas, M.C. Hepatitis D virus and hepatitis B virus infection in Amerindian communities of the Amazonas state, Colombia. *Virol. J.* **2015**, *12*, 172. [CrossRef]
29. Makuwa, M.; Caron, M.; Souquiere, S.; Malonga-Mouelet, G.; Mahe, A.; Kazanji, M. Prevalence and genetic diversity of hepatitis B and delta viruses in pregnant women in Gabon: Molecular evidence that hepatitis delta virus clade 8 originates from and is endemic in central Africa. *J. Clin. Microbiol.* **2008**, *46*, 754–756. [CrossRef]
30. Radjef, N.; Gordien, E.; Ivaniushina, V.; Gault, E.; Anais, P.; Drugan, T.; Trinchet, J.C.; Roulot, D.; Tamby, M.; Milinkovitch, M.C.; et al. Molecular phylogenetic analyses indicate a wide and ancient radiation of African hepatitis delta virus, suggesting a deltavirus genus of at least seven major clades. *J. Virol.* **2004**, *78*, 2537–2544. [CrossRef]
31. Santos, M.D.; Gomes-Gouvêa, M.S.; Nunes, J.D.; Barros, L.M.; Carrilho, F.J.; Ferreira, A.e.S.; Pinho, J.R. The hepatitis delta genotype 8 in Northeast Brazil: The North Atlantic slave trade as the potential route for infection. *Virus Res.* **2016**, *224*, 6–11. [CrossRef]
32. Niro, G.A.; Smedile, A.; Andriulli, A.; Rizzetto, M.; Gerin, J.L.; Casey, J.L. The predominance of hepatitis delta virus genotype I among chronically infected Italian patients. *Hepatology* **1997**, *25*, 728–734. [CrossRef] [PubMed]
33. Su, C.W.; Huang, Y.H.; Huo, T.I.; Shih, H.H.; Sheen, I.J.; Chen, S.W.; Lee, P.C.; Lee, S.D.; Wu, J.C. Genotypes and viremia of hepatitis B and D viruses are associated with outcomes of chronic hepatitis D patients. *Gastroenterology* **2006**, *130*, 1625–1635. [CrossRef] [PubMed]
34. Barros, L.M.; Gomes-Gouvea, M.S.; Pinho, J.R.; Alvarado-Mora, M.V.; Dos Santos, A.; Mendes-Correa, M.C.; Caldas, A.J.; Sousa, M.T.; Santos, M.D.; Ferreira, A.S. Hepatitis Delta virus genotype 8 infection in Northeast Brazil: Inheritance from African slaves? *Virus Res.* **2011**, *160*, 333–339. [CrossRef]
35. Chang, W.S.; Pettersson, J.H.; Le Lay, C.; Shi, M.; Lo, N.; Wille, M.; Eden, J.S.; Holmes, E.C. Novel hepatitis D-like agents in vertebrates and invertebrates. *Virus Evol.* **2019**, *5*, vez021. [CrossRef]
36. Wille, M.; Netter, H.J.; Littlejohn, M.; Yuen, L.; Shi, M.; Eden, J.S.; Klaassen, M.; Holmes, E.C.; Hurt, A.C. A divergent hepatitis D-like agent in birds. *Viruses* **2018**, *10*, 720. [CrossRef] [PubMed]
37. Hetzel, U.; Szirovicza, L.; Smura, T.; Prahauser, B.; Vapalahti, O.; Kipar, A.; Hepojoki, J. Identification of a novel deltavirus in boa constrictors. *mBio* **2019**, *10*. [CrossRef]
38. Perez-Vargas, J.; Amirache, F.; Boson, B.; Mialon, C.; Freitas, N.; Sureau, C.; Fusil, F.; Cosset, F.L. Enveloped viruses distinct from HBV induce dissemination of hepatitis D virus in vivo. *Nat. Commun.* **2019**, *10*, 2098. [CrossRef]
39. Kuiken, C.; Korber, B.; Shafer, R.W. HIV sequence databases. *Aids Rev.* **2003**, *5*, 52–61.
40. 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*, D566–D570. [CrossRef]
41. Kuiken, C.; Hraber, P.; Thurmond, J.; Yusim, K. The hepatitis C sequence database in Los Alamos. *Nucleic Acids Res.* **2008**, *36*, D512–D516. [CrossRef] [PubMed]
42. Kefalakes, H.; Koh, C.; Sidney, J.; Amanakis, G.; Sette, A.; Heller, T.; Rehermann, B. Hepatitis D Virus-specific CD8(+) T cells have a memory-like phenotype associated with viral immune escape in patients with chronic hepatitis D virus infection. *Gastroenterology* **2019**, *156*, 1805–1819. [CrossRef] [PubMed]
43. Karimzadeh, H.; Kiraithe, M.M.; Oberhardt, V.; Salimi Alizei, E.; Bockmann, J.; Schulze Zur Wiesch, J.; Budeus, B.; Hoffmann, D.; Wedemeyer, H.; Cornberg, M.; et al. Mutations in hepatitis D virus allow it to escape detection by CD8(+) T cells and evolve at the population level. *Gastroenterology* **2019**, *156*, 1820–1833. [CrossRef] [PubMed]
44. Benson, D.A.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2015**, *43*, D30–D35. [CrossRef]
45. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
46. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef] [PubMed]
47. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef] [PubMed]

48. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—New capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, e115. [CrossRef] [PubMed]
49. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321. [CrossRef]