scientific reports

Check for updates

OPEN

# Understanding the genetics of viral drug resistance by integrating clinical data and mining of the scientific literature

An Goto[1], Raul Rodriguez-Esteban[2], Sebastian H. Scharf[2] & Garrett M. Morris[1]✉

**Drug resistance caused by mutations is a public health threat for existing and emerging viral diseases. A wealth of evidence about these mutations and their clinically associated phenotypes is scattered across the literature, but a comprehensive perspective is usually lacking. This work aimed to produce a clinically relevant view for the case of Hepatitis B virus (HBV) mutations by combining a chronic HBV clinical study with a compendium of genetic mutations systematically gathered from the scientific literature. We enriched clinical mutation data by systematically mining 2,472,725 scientific articles from PubMed Central in order to gather information about the HBV mutational landscape. By performing this analysis, we were able to identify mutational hotspots for each HBV genotype (A-E) and gene (C, X, P, S), as well as the location of disulfide bonds associated with these mutations. Through a modelling study, we also identified a mutation position common in both the clinical data and the literature that is located at the binding pocket for a known anti-HBV drug, namely entecavir. The results of this novel approach show the potential of integrated analyses to assist in the development of new drugs for viral diseases that are more robust to resistance. Such analyses should be of particular interest due to the increasing importance of viral resistance in established and emerging viruses, such as for newly developed drugs against SARS-CoV-2.**

Gene mutations that confer drug resistance to pathogens are an emerging public health and medical threat with, additionally, potential consequences for current and future pandemic response. A step towards combating this scourge is to study the way in which particular mutations lead to viral drug-resistant phenotypes. Viral species hosted by an individual patient can contain multiple mutations that interact to produce specific phenotypes. To understand mutational interactions, the scientific literature can help in interpreting existing knowledge on the biological mechanisms associated with viral mutation-phenotype relationships. Here we explore the feasibility of a novel integrated approach that combines clinical and text mining data to gather existing knowledge and produce new insights on drug resistance based on data from viral species hosted by individual patients, and in particular for the case of HBV. Such approaches have the potential to be applied to other viral diseases, singularly due to the growing acknowledgement of the critical importance of viral mutation monitoring in patients.

Hepatitis B is an infectious disease that affects approximately 292 million people worldwide. Despite being a major global health concern, it has been estimated that only 10% of individuals who are chronically infected with HBV have been diagnosed, and only 5% of those individuals who are eligible for treatment receive an antiviral therapy[1]. Chronic HBV infection is known to progress to cirrhosis of the liver in up to 40% of untreated patients[2]; and in 2015, HBV led to an estimated 887,000 deaths, which were largely caused by cirrhosis of the liver and hepatocellular carcinoma (HCC)[3]. There are two classes of treatments that are known to be effective against suppressing HBV infections: interferons and nucleotide analogues. These treatments can reduce the patient's viral load and the main impacts of the disease on the patient. However, although these treatments have been available for nearly two decades, they have not eliminated HBV[4].

HBV is classified into ten HBV genotypes, A to J, each with a distinct geographic distribution[5,6]. The infectious HBV virion, also known as the *Dane particle*, is a spherical, double-shelled structure with a diameter of 42 nm[7]. This particle consists of an outer lipoprotein envelope and an inner nucleocapsid core, which encapsulates the viral genome[6]. The genome of HBV is approximately 3.2 kilobase pairs long and has a partially double-stranded circular DNA[7]. The viral genome codes for all five viral proteins required for HBV replication: the HBV surface

[1]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK. [2]Roche Innovation Center Basel, Basel, Switzerland. ✉email: garrett.morris@stats.ox.ac.uk

antigen; the HBV core antigen; the HBV envelope antigen; the X protein; and the HBV reverse transcriptase/polymerase.

There are four known genes in the viral genome: C, X, P, and S. Gene C encodes the core protein, and it also produces the pre-core protein by promoting an upstream AUG start codon as its start codon[8]. Genes P and S encode for DNA polymerase and HBsAg, respectively. The S gene is an open long reading frame that is divided into three sections based on the start codons: pre-S1, pre-S2, and S[9]. The mechanism and function of the protein coded by gene X, HBV X antigen, remains largely unknown, but it is known to be associated with the development of liver cancer[10].

Due to its high mutation rate, with commonly accepted rates of ~ $2.0 \times 10^{-5}$ nucleotide substitutions per site per year[11], HBV can become resistant to HBV drugs relatively quickly. With the emergence of drug resistance, patients may experience a notable compromise in the efficacy of anti-viral therapy and deterioration of the disease condition[12,13]. Thus, investigation of resistance mutations is crucial for the successful development of robust HBV drugs. Lamivudine (phosphorylated lamivudine) and adefovir dipivoxil were the first nucleotide analogues that were developed but had limited success because of the development of resistant variants. Resistance mutations pertaining to nucleotide analogues are usually reported in the polymerase gene of HBV DNA (*i.e.,* gene P)[14]. The resistance phenotype is typically a consequence of nucleotide analogues failing to bind to DNA polymerase[15].

Lamivudine and adefovir diphosphate are widely known for treating HBV[16]. After conducting randomized clinical trials of lamivudine, it was found that, compared with placebo, HBV mutant variants were associated with reduced sensitivity to lamivudine in approximately 30% of patients after only a year of treatment[17,18]. In addition, mutations causing resistance to adefovir dipivoxil were detected in up to 20 to 29% of patients after five years of treatment[19,20]. However, more-recently developed nucleotide analogues, such as entecavir and tenofovir disoproxil, have been reported to dramatically decrease the rate of development of resistance.

In the case of entecavir, only 1.2% of patients developed a resistant strain after five years of treatment if they had never been treated with nucleotide analogues previously[21]. For tenofovir disoproxil, there were no clinically significant resistant variants identified during up to seven years of follow-up[22]. Cross-resistance is also known to occur and is another concern for the development of drug resistance. For example, between lamivudine- and entecavir-resistant HBV strains, the cumulative probability of developing entecavir resistant variants is more than 50% for those individuals with lack of resistance toward lamivudine[21].

In order to create new therapies that can circumvent or negate mechanisms of drug resistance, we need to understand resistance mutations better with the aid of analytics tools such as text mining. Text mining has been applied to extract evidence for generic drug resistance[23,24], impact of mutations in disease[25] and to harvest viral mutation data[26]. It has also been leveraged to create comprehensive databases of the viral mutation literature[27,28]. There is no report, however, that it has been used to produce new insights that support clinical data analysis of viral mutations with integration of literature data. This study aimed to build on a clinical study that ultra-deep sequenced HBV quasispecies from a European cohort[29] by comparing the data from this clinical study against genetic mutation information automatically extracted from the scientific literature regarding all known launched anti-HBV drugs. Such data are typically scattered across many publications; therefore, we built a pipeline to mine literature sources systematically. In order to extract mutations from the literature we applied a rule-based text mining approach[30–32]. With our approach, we were able to produce a landscape of disease-specific viral mutations associated to drug resistance, which allowed us to derive new insights into mechanisms associated to drug resistance.

There exist manually curated databases[33–36] that aim to gather HBV resistance mutations from patients' data. However, it is hard to evaluate their coverage since they do not give a complete list of resistance mutations and instead provide a mean to analyze the genetic sequences that the user has obtained. In addition, most of these databases are not open source. Hence, the use of text mining approaches is able to mine through the literature efficiently are essential to tackle the challenge of identifying known resistance mutations.

## Methods

**Code and supplementary information.** The analysis was performed using standard Python packages (*e.g., ElementTree*[37] for parsing XML-formatted literature, *re* for identifying information about mutations, and *pandas* for data analysis). The resulting code and supplementary information (Supplementary Figures S1–S15, Tables S1–S5, and Note S1) is available at: https://github.com/angoto/HBV_Code.

**Source of data.** Mining of the scientific literature was conducted by using all 2,472,725 XML files from PubMed Central (commercial and non-commercial use), which consisted of information about each publication such as title, authors, DOI, abstract, full body content, and references[38]. Clinical data for genetic mutations of HBV was collected from a total of 186 plasma samples from a Western European cohort of chronic HBV patients during the period from 1985 to 2012[29]. HBV DNA was extracted from 200 µL of plasma using the QIAamp MinElute Virus Kit (Qiagen) or the Roche MagNA Pure LC instrument according to the manufacturer's instructions. These samples are stored at the Erasmus University Medical Center in Rotterdam, the Netherlands. The guidelines followed in the study were in accordance with the Declaration of Helsinki[39] and the principles of Good Clinical Practice[40]. The study was approved by the ethical review board of the Erasmus Medical Center, Rotterdam, the Netherlands. The clinical data used in this study was anonymized and permission was not required to access the raw data of the clinical study. The dataset used to build the anti-HBV drugs' vocabulary included 69 launched anti-HBV drugs obtained from the Cortellis Drug Discovery Intelligence database (CDDI), formerly known as Integrity[41]. We extracted each drug's code name, generic name, brand name, and drug name from the CDDI database to enrich this vocabulary, which gave 312 unique drug expressions.

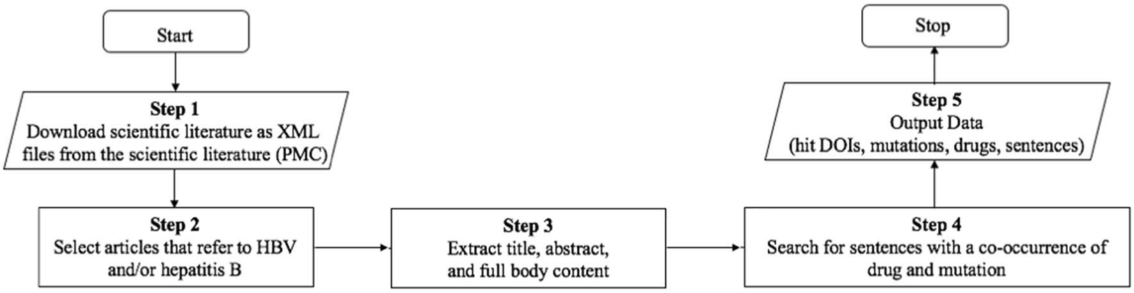
Start
Step 1
Download scientific literature as XML files from the scientific literature (PMC)
Step 2
Select articles that refer to HBV and/or hepatitis B
Step 3
Extract title, abstract, and full body content
Step 4
Search for sentences with a co-occurrence of drug and mutation
Step 5
Output Data
(hit DOIs, mutations, drugs, sentences)
Stop

**Figure 1.** Workflow for mining the scientific literature.

### Scientific literature mining.

The workflow in Fig. 1 was used to search through the sentences from journal articles in PubMed Central (PMC) that mentioned both an HBV drug and a mutation. In Step 2, we have used GNU parallel in order to select relevant literature that refer to 'hepatitis b' and/or 'hbv' (both keywords are case insensitive). In Step 4, a regular expression was used to split sentences from PMC and mutations were identified by using regular expressions for HBV mutations, as described in the next section. We avoid any problems with abbreviations that could lead to false positives, such as confusing the viral gene "C" with the element "C", by focusing our search for specific mutations found in the clinical study, and ensuring co-occurrence of HBV drugs in the same sentence. By basing our rules on co-occurrence relationships between drug and mutation, our approach is generalizable to new data, and new diseases. The output data from this workflow was later used to conduct the data analysis (*i.e.,* Step 5 in Fig. 1). Further details of Steps 1–5 are available at: https://github.com/angoto/HBV_Code.

### Regular expressions for HBV mutations.

Mutations relevant to HBV were found from articles in PubMed Central by combining regular expressions.

### Use of clinical study data.

Six different pieces of data from the Western Europe cohorts described in Mueller-Breckenridge, et al.[29] were used to conduct the analysis: sample, reference genome, gene, effect, amino acid variant, and nucleotide variant. "Sample" identifies the unique patients' number used during the clinical study. Reference genome refers to a genotype-specific reference sequence (namely, GenBank accession numbers AF090842, AB033554, AB033556, AF121240, and AB032431, for genotypes A to E, respectively) that was used to map the outputs generated from quality-trimmed demultiplexed FASTQ reads[29]. "Gene" indicates four known genes encoded by the genome: C, P, S, and X. "Effect" refers to whether a particular mutation was either an intragenic or a missense variant. For this report, we have only considered missense variants because intragenic variants from the literature were not associated with the clinical study's phenotypes. "Amino acid variant" ($n=7285$) and "nucleotide variant" ($n=10{,}658$) are two different ways to describe the same mutation, *i.e.* mutation in the amino acid (*e.g.* p.Pro156Ser) and nucleotide (*e.g.* c.314C > T) sequences, respectively. Although it is relatively straightforward to obtain HBV-related genetic mutations from the literature, it is difficult to map the position number of nucleotides to the position number of amino acids, and *vice-versa*. This is because we are unsure about the reference sequence that was used to report these mutations for each journal article in PubMed Central. In order to do this type of mapping with our current approach, we would need to manually go over each journal article that mentioned both an HBV drug and a mutation to conduct the translation of DNA sequences to amino acid sequences because an ambiguous inverse mapping would be impossible (except for Tryptophan). Hypothetical clinical study data is available on the GitHub repository for this study[42] to provide a better idea of the data structure used in the clinical study.

### Comparison of scientific literature and clinical study.

The workflow in Fig. 2 was used to compare mutations from the clinical study with the genetic mutations found in the scientific literature downloaded from PubMed Central after translating the mutations from the literature to the format used in the clinical study: amino acid (category *vi*) and nucleotide (category *ii*) variants ($n=4214$).

### Mutation hotspot for entecavir.

We independently built our own models of entecavir bound with HBV RT and DNA, using the X-ray crystal structures of HIV RT also bound to entecavir and DNA, from PDB entries 5XN1[43] and 6IKA[44]. Homology modelling was performed using SWISS-MODEL[45], and UniProt ID Q9WRJ9[46] residues 349–692 for the sequence of HBV RT genotype A. The resulting model underwent energy minimization using the open-source build of PyMOL version 2.3.0[47], relaxing the entecavir ligand and all residues and bases within 4 Å of entecavir.

## Results

### Comparison between the clinical study and the literature.

The number of publications found to have a sentence co-occurrence of an approved HBV drug and a mutation in the same sentence was 30,686. There were 4,214 unique mutations mentioned in those sentences and, 7.5% of those, a total of 316 mutations (*i.e.,* 254 amino acid and 62 nucleotide variants) were also found in the clinical data from Western Europe cohorts ($n=182$ patients, 31,977 mutations). After analyzing the clinical data using solely those unique mutations that

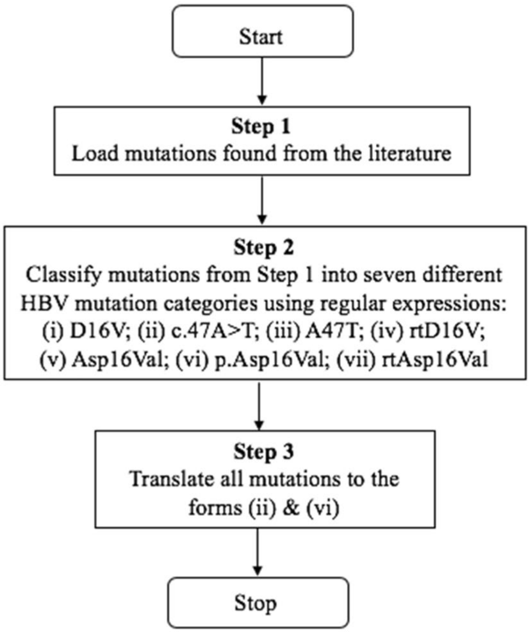
Start
Step 1
Load mutations found from the literature
Step 2
Classify mutations from Step 1 into seven different HBV mutation categories using regular expressions: (i) D16V; (ii) c.47A>T; (iii) A47T; (iv) rtD16V; (v) Asp16Val; (vi) p.Asp16Val; (vii) rtAsp16Val
Step 3
Translate all mutations to the forms (ii) & (vi)
Stop

**Figure 2.** Normalization of DNA and amino acid mutations found from text mining the literature, to enable the direct comparison with clinical study mutations.

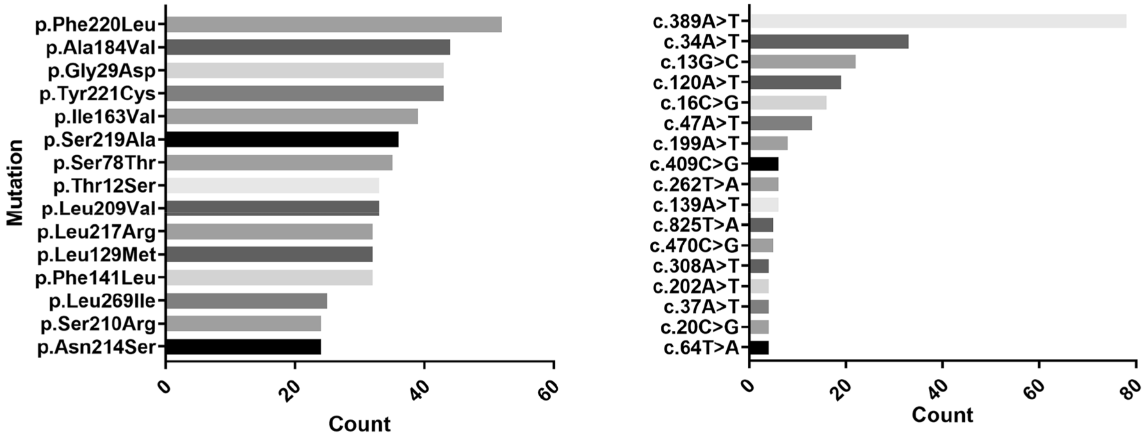

**Figure 3.** Top 10 most frequent amino acid mutations (left) and nucleotide mutations (right) in the clinical data that were also found in the literature.

were common between the clinical study and the literature, we found that 180 out of 182 patients (corresponding to 1750 amino acid and 302 nucleotide variants) had mutations that also appeared in the literature.

**Prevalence of genetic mutations in the clinical study.** Figure 3 shows the ten most frequent amino acid (left) and nucleotide variants (right) in the clinical study that were also found in the literature. Their prevalence ranged between 1 and 52 patients for amino acid variants and between 1 and 78 patients for nucleotide variants. Supplementary Figures S1–S2 show the frequency of mutations in the clinical data (*i.e.,* number of patients reported to have each mutation type).

**Mutation likelihood of patients in the clinical study.** The 180 patients presented between 16 and 753 HBV mutations. Out of these mutations, between 1 and 35 appeared in the literature (Supplementary Figure S3-1), with an average of 11.2 mutations and a median of 10 mutations. Supplementary Figure S3-2 shows the patients in the clinical study who had the most mutations matching those found in the literature.

**Mutation hotspots for HBV genotypes in the clinical study.** We identified amino acid mutations in the four genes X, P, C, and S, across five HBV genotypes, A, B, C, D, and E, that emerged during the clinical study. There were 182 patients in the clinical study, of which there were 56 patients with genotype A, 19 with genotype

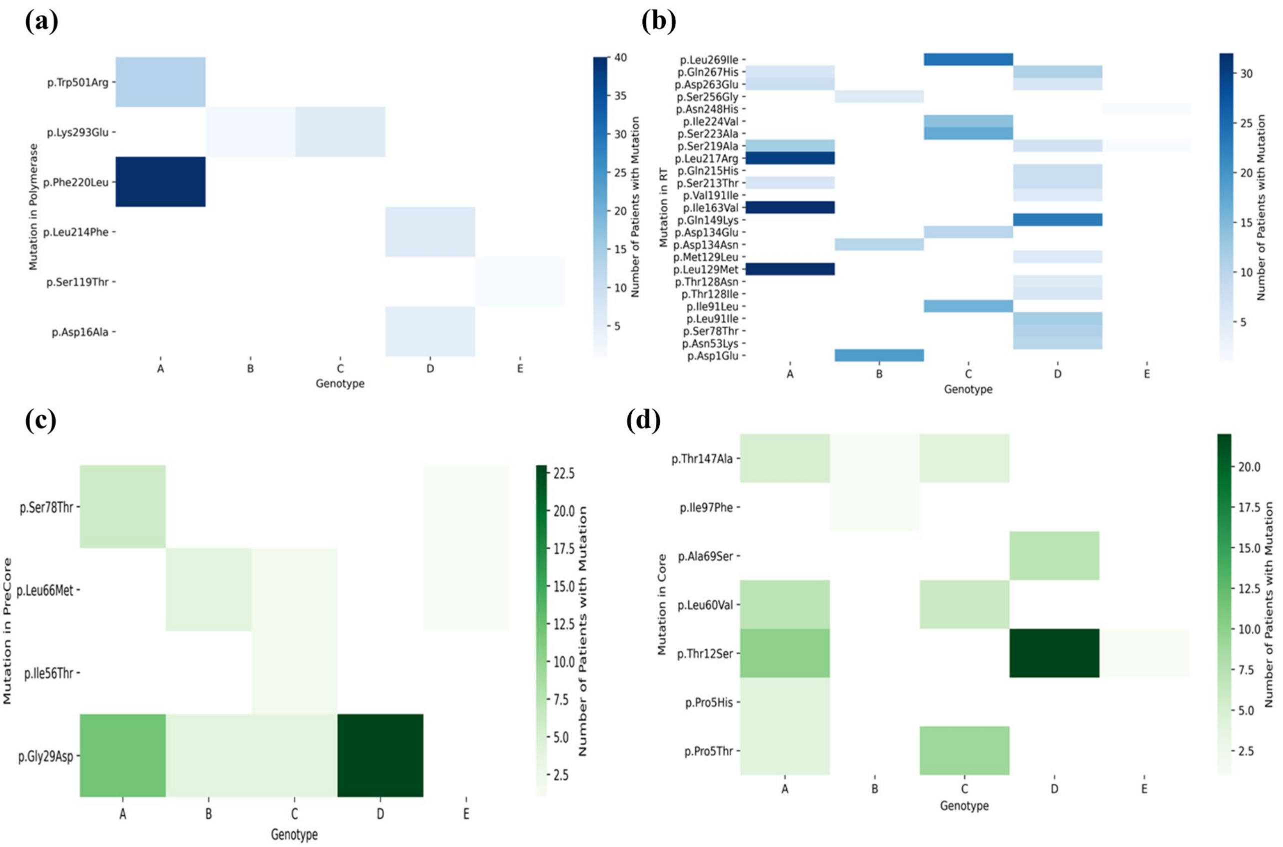

**Figure 4.** Heatmap of mutation hotspots for gene P in genotypes A-E: (**a**) HBV polymerase, and (**b**) HBV reverse transcriptase (RT) and heatmap of mutation hotspots for gene C in genotypes A-E: (**c**) HBV precore, and (**d**) HBV core. The white regions represent counts of mutations that are null, while dark blue (in **a** and **b**) or dark green (in **c** or **d**) indicates the largest number of mutations.

B, 43 with genotype C, 62 with genotype D, and 2 with genotype E. In order to create a mutation hotspot map for each of these genotypes (Supplementary Figures S4–S8) where there was at least one report in the literature of that mutation, overlaps between the mutations from the literature and the clinical study mutations were identified. It should be noted that there are mutations in the patients that were present in the clinical study but were not found in the analysis of the literature, and hence are not shown in these heatmaps.

We defined a "mutation hotspot" as a mutation with a count above the average for that particular gene and genotypes. For example, for HBV RT genotype A, mutation V214A had a count of 1 patient, while the average of the counts was 5.18 patients, so this was excluded from the map; while mutation L217R had a count of 30 patients, which was above the average for HBV RT genotype A, and thus was included as a mutation hotspot. For nucleotide variants, similar plots were made based on the genotypes (A-E), with mutations sorted in order of position number (Supplementary Figures S9–S13). Figures 4, 5, 6 represent heatmaps of the mutational hotspots that we identified for eight different gene products (*i.e.,* polymerase, reverse transcriptase, X, precore, core, PreS1, PreS2, and HBsAg) and nucleotide variants, with darker, more saturated colors indicating more mutations at that position in that genotype. The summary tables of hotspots for each HBV genotype for amino acid and nucleotide variants are shown in Supplementary Tables S1–S5.

Our analysis revealed there were "genotype-common" mutation hotspots, *i.e.,* mutations with the same position number, in two or more genotypes (Figs. 4, 5, 6), above the average count of patient. These were: polymerase at position K293; reverse transcriptase at positions S213, S219, D263, and Q267; PreS1 at positions S5, F141, and N214; PreS2 at position R16; HBsAg at positions E2, L77, K122, P127, M133, Y161, W182, A184, S204, S210, and F220; X at position P33; precore at positions G29, L66, and S78; core at positions P5, T12, L60, and T147; and nucleotide variants at positions 34A, 120A, and 389A.

Thus, although we were unable to discriminate HBV drug-resistant mutations unambiguously from other mutations, we were able to identify common mutations in the clinical study and the literature, across genotypes A-E. This can further our understanding of HBV mutations by highlighting mutations that may be potentially relevant to resistance.

**Mutations occurring in disulfide bonds.** We investigated the number of common mutations between the clinical study and the literature that arose from mutating wild type cysteine residues. Cysteine residues can be responsible for the formation of disulfide bonds, which play an important role in folding and stability of the proteins[48]. The HBV envelope proteins, which corresponds to gene S, are known to form an intermolecu-

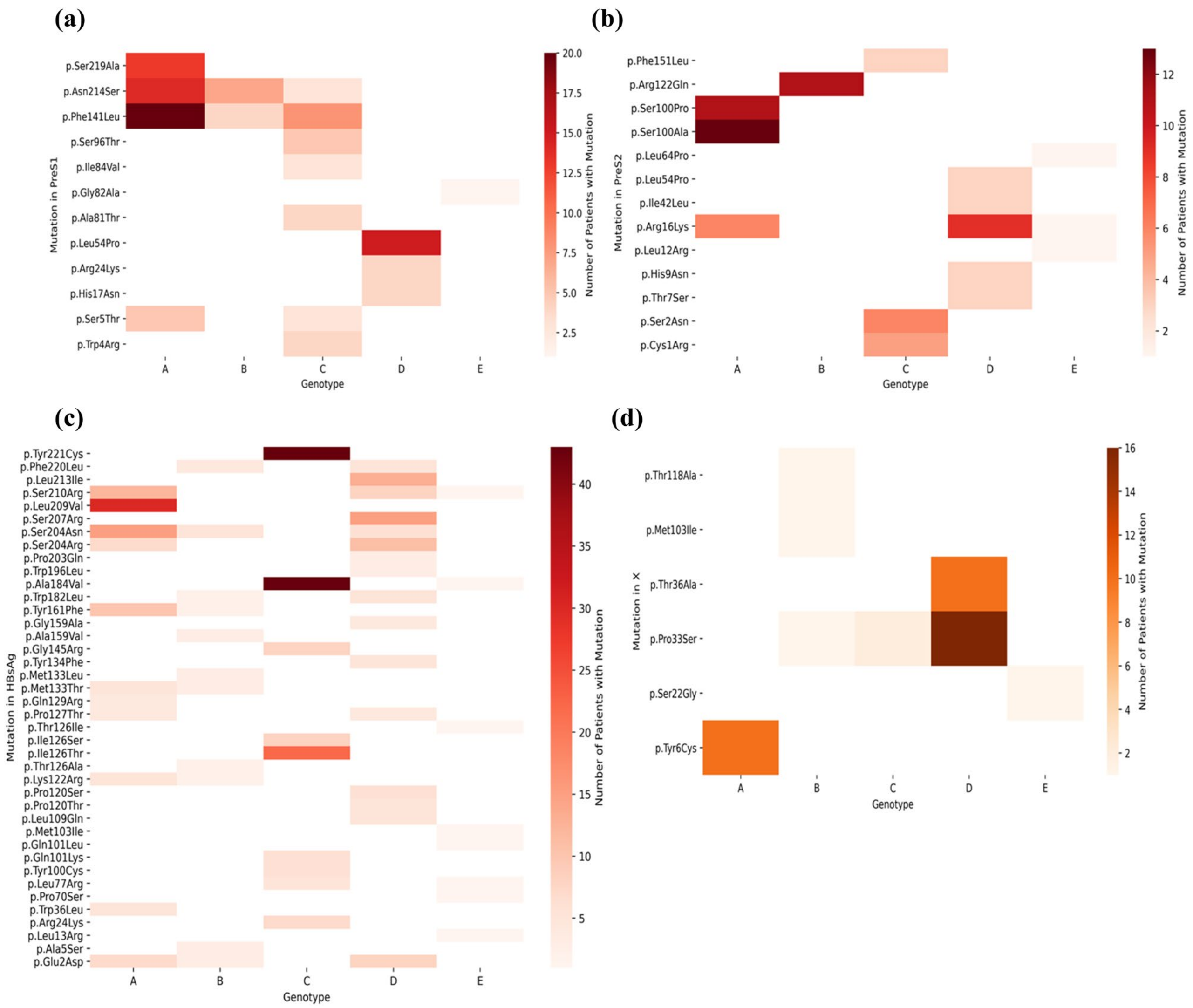

**Figure 5.** Heatmap of mutation hotspots for gene S in genotypes A-E: (**a**) HBV PreS1, and (**b**) HBV PreS2; heatmap of mutation hotspots for: (**c**) HBV HBsAg (gene S), and (d) HBV gene X in genotypes A-E. The white regions represent counts of mutations that are null, while dark red indicates the largest number of mutations.

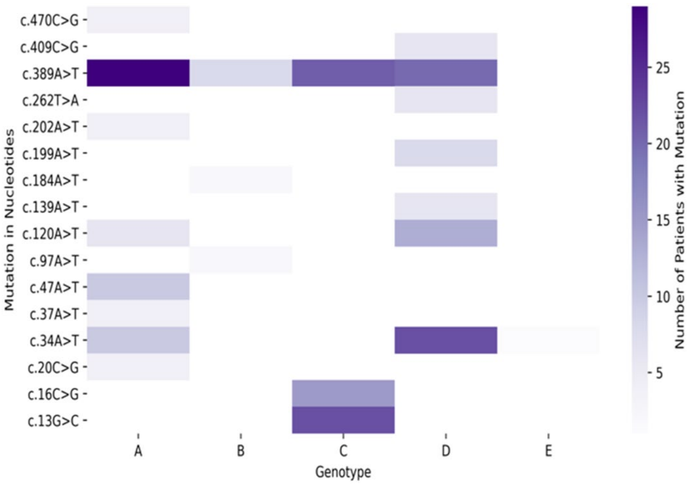

**Figure 6.** Heatmap of mutation hotspots for HBV nucleotide variants in genotypes A-E. The white regions represent counts of mutations that are null, while dark blue indicates the largest number of mutations.

| Genotype | Cysteine Residues |
|---|---|
| A | p.Cys149Arg (S), p.Cys48Gly (C) |
| B | p.Cys1Arg (S), p.Cys1Ser (S), p.Cys48Gly (C) |
| C | p.Cys149Arg (S), p.Cys1Arg (S) |
| D | p.Cys124Arg (S), p.Cys149Arg (S) |
| E | N/A |

**Table 1.** Cysteine residues in genotypes A-E for common mutations between the clinical study and the literature with corresponding genes labelled in parentheses (*i.e.*, gene S, and gene C).

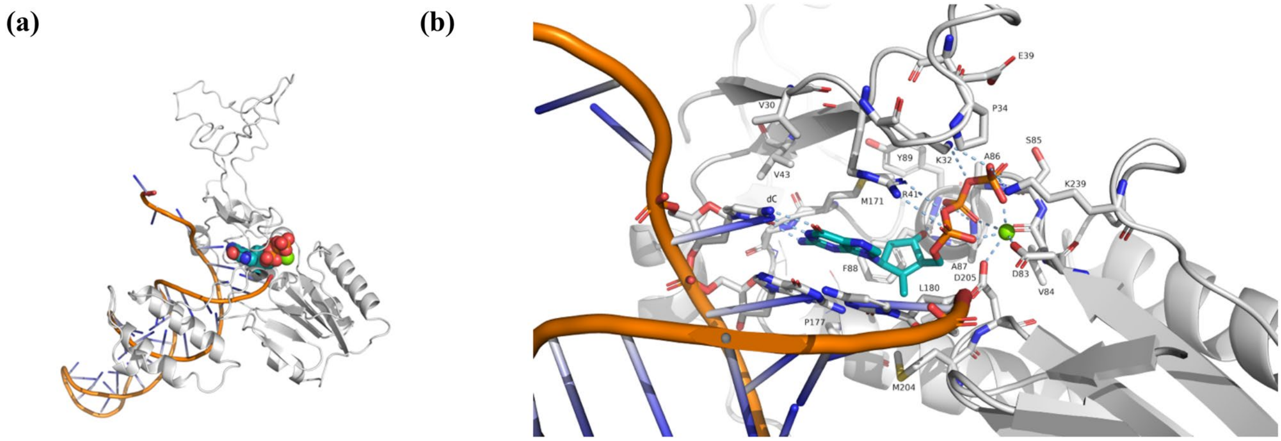

**Figure 7.** Model of HBV RT (genotype A) based on the HIV RT-DNA-entecavir complex X-ray crystal structure, PDB ID 5XN1[43]. (**a**) Entecavir-triphosphate is shown as spheres with teal-colored carbon atoms, red oxygen, blue nitrogen, and orange phosphorus atoms. Entecavir is at the 3′ end of the DNA strand (orange cylinders). The secondary structure of the HBV RT model is shown as white coils ($\alpha$-helices), white arrows ($\beta$-strands), and white loops. (**b**) Close-up view of the binding pocket, showing entecavir-triphosphate with teal-colored carbons, and a $Mg^{2+}$ cation as a green sphere. Note that hydrogen bonds and metal bonds are shown as light blue dashed lines. The deoxyguanine (dG)-moiety of entecavir can be seen "base-pairing" with deoxycytosine (dC) in a second strand of DNA.

lar disulfide network through cysteine residues in the cysteine-rich antigenic loop that are in positions 102 to 161[49–51]. Hence, mutations for cysteine residues reported in Table 1 for positions C125 and C149 are in fact in regions of disulfide bonds. For gene C, the cysteine residue in position C48 is expected to form a disulfide bond with C149[52]. We were unable to find any evidence of a disulfide bond involving residue C1 in gene S.

Therefore, by comparing the literature and the clinical study, we were able to identify which gene, and in what particular position, mutations occurred at disulfide bond. We found mutations located at C124 and C149 in gene S to occur in regions of disulfide bonds, which are likely to have an effect towards destabilizing the HBV proteins. Table 1 shows the list of cysteine residues in each genotype, A-E, for mutations that were common between the clinical data and the literature.

**Mutation hotspot for entecavir.** Based on a search conducted using DrugBank[16], one out of 69 approved drugs listed in the CDDI database had a modelling study published with a known target and binding site[53]. By using our own models of entecavir bound with HBV RT and DNA, we identified which of these binding pocket residues were common between the clinical data and the scientific literature. For entecavir, we found common mutations located at I169, M204, and N238 for genotype A (within 12 Å of entecavir in our model).

Figure 7 shows the binding pocket of entecavir, together with the binding pocket residues within 5 Å of entecavir. The original modelling study by Langley, et al.[53] used a sequence alignment between HIV and HBV reverse transcriptase to determine the most conserved domains, and they used the HIV reverse transcriptase DNA X-ray structure (PDB ID 1RTD)[54] to build the HBV RT model.

We also investigated the location of the mutational hotspots we identified for HBV RT in genotype A and mapped these to our model. It can be seen from Fig. 8 that the mutations occur throughout the structure.

Monitoring of drug resistance driven by viral mutations is relevant for antiviral drugs beyond those used for HBV, such as human cytomegalovirus (HMCV)[55], HIV[56], hepatitis C virus[57], influenza virus[58], SARS-CoV-2[59] and more. Hence, the method proposed could be useful for a range of antiviral drugs and associated diseases.

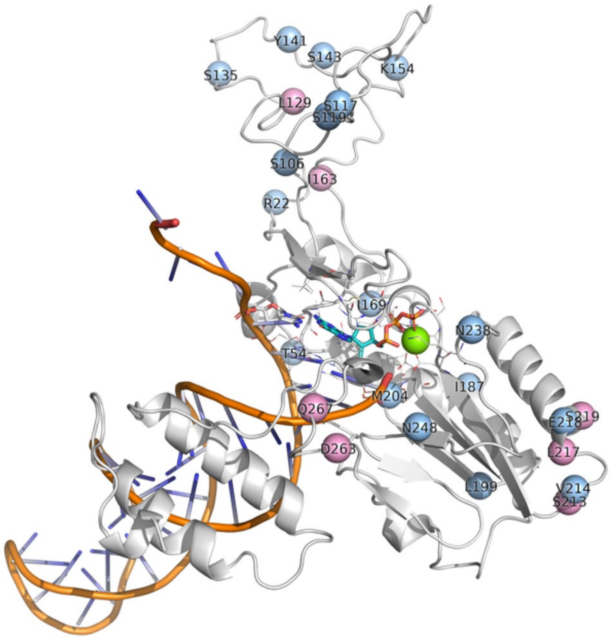

**Figure 8.** Locations of the mutations in HBV RT found both in patients in the clinical study, and in our analysis of the literature, shown as spheres placed at the α-carbon atom of the amino acid. The pink spheres are mutations with above-average counts of patients, and are at positions 129, 163, 213, 217, 219, 263, and 267; while the light blue spheres indicate positions of mutations below the average threshold. The secondary structure of our model of the HBV RT is shown as a white cartoon, while the DNA backbone is in orange. The location of the DNA, entecavir-triphosphate (shown with teal carbons), and $Mg^{2+}$ (green sphere) was modelled on PDB ID 5XN1[43].

## Discussion and conclusions

This study showed the feasibility of integrating clinical and text mined viral mutation data and its potential to produce new insights on disease-relevant viral mutations. The focus of the study was on HBV, but the methods are applicable to other viruses. While mutational hotspots can be identified by considering only the clinical study's mutations, the difficulty associated with this process is that, even when new mutational hotspots are identified through the study, they are often disregarded because there are no literature reports available or there are not enough resources to comb the literature to confirm its role in resistance mutations. Thus, it is typical for clinical studies to report only those mutational hotspots that have been reported in the past. Hence, the methodology presented in this study serves as a bridge to fill in that gap by leveraging the full text of 2.47 million articles from PubMed Central to produce a landscape of disease-specific viral mutations. Although it is not possible for this method to confirm a direct correlation between the mutational hotspots and resistance mutations, it is able to provide new hypotheses that are potentially relevant to the resistance phenotype.

Our study identified known hotspots that were found in four genes, P, C, S, and X, in genotypes A to E, as shown in Figs. 4, 5, 6. There are other HBV mutational hotspots that are known such as L180, which is a compensatory mutation for resistance to entecavir, lamivudine, and telbivudine[60]; rt202 which causes resistance to entecavir in Asian population[61]; and rt236 which is responsible for resistance to adefovir dipivoxil in Caucasian[62]. Hence, the list of mutational hotspots that were identified in this study was not necessarily conclusive. In addition, two amino acid positions (*i.e.,* C124 and C149) exhibited mutations in disulfide bonds, which are likely to impact the structure of HBV proteins. This finding is confirmatory since these regions of the disulfide bonds have already been identified in the past[49–51]. Moreover, we identified a mutation in a position that was relevant to the binding site for the anti-HBV drug, entecavir. However, the I169 mutation, which is known to be the primary mutation responsible for entecavir's resistance was found to be reported below the average threshold when comparing mutational hotspots between the clinical study and the literature[63]. Therefore, further refinement of the definition of mutational hotspots used in this study may be necessary in the future.

One limitation of our approach lies in terms of the amount of data we collected from the literature. This resulted in an overall coverage of the number of genetic mutations reported in the literature against the common mutations of the literature and the clinical study of 7.5%. In order to further increase the coverage of genetic mutations data, literature sources could be increased to include repositories from publishers such as ScienceDirect[64] and Springer Nature[65]. We could also obtain additional data from the PubMed Central literature by mining the information included in tables and figures. It is also important to be aware of errors that may arise during DNA sequencing and patient data collection during the clinical study.

In addition, we considered the rate of false positive identifications of mutations. An example of a false positive would be the following sentence: *"Some non-resistance-associated mutations of rtD134N (ranging from 20.33 to*

*74.63%), rtL145M (ranging from 2.83 to 78.82%), rtF151Y (ranging from 2.92 to 75.51%) and rtS223A (ranging from 5.77 to 18.44%) increased significantly with ADV monotherapy, then declined with the addition of LdT"*[66]. In this sentence, it satisfies the criterion for co-occurrence of an HBV drug and a mutation, but it is in fact not referring to resistance mutations. False positives could also be considered hypothetical mutations, such as the one described in the sentence: *"The lamivudine resistance-linked G529A (rtD134N) site in HBV was found to be associated with HCC outcomes, which implied potential correlation between resistance to the anti-HBV nucleoside analog lamivudine and HCC prognosis"*[67]. Furthermore, it is possible that some drug-resistant mutants involve multiple locations, so in the future we must consider such combinatorial possibilities as well. To estimate the false positive rate, we selected 50 random sentences with co-occurrence of drug and mutation and inspected them manually. It was found that two of them were false positives. Hence, we estimate the rate of false positives to be approximately 4% of the overall data.

To improve the methods used here, additional text mining strategies could be used to improve the extraction, such as using additional synonyms from UMLS and SNOMED for both the drugs and the disease, which would decrease the number of false negatives. By using machine learning algorithms, thus we could also increase our confidence that sentences describe HBV-related information. For example, the following sentence refers to two mutations, M204I and M204V, which are not captured by simple regular expressions: *"LVDr is well characterized and arises through replacement of M204 within the YMDD motif of the HBV RT with isoleucine or valine, with or without the adaptive change L180M"*[46]. By performing the literature mining in this manner, we could be able to develop a more powerful text mining tool that would allow us to identify a more comprehensive set of resistance mutations. However, it is important to note that, while more-recently published mutation-extraction approaches are machine-learning based[68,69], these require costly gold standard corpora, which hinders their re-application to different viral species, each with its own notation particularities. Thus, the development of machine-learning algorithms may lead to improved performance but could result in more restricted application to a particular disease.

## Data availability

The code and supplementary information (Supplementary Figures S1–S15, Tables S1–S5, and Note S1) are available at: https://github.com/angoto/HBV_Code. Access to the data in reference 29 can be requested by anyone by contacting Alan Mueller-Breckenridge at Roche, using the email address: alan_james.mueller-breckenridge@roche.com.

Received: 15 December 2021; Accepted: 30 July 2022
Published online: 25 August 2022

## References

1. Polaris Observatory Collaborators. Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: A modelling study. *Lancet Gastroenterol. Hepatol.* **3**, 383–403 (2018).
2. Sarin, S. K. *et al.* Asian-pacific clinical practice guidelines on the management of hepatitis B: A 2015 update. *Hepatol Int.* **10**, 1–98 (2016).
3. World Health Organization. Hepatitis B. *World Health Organization* https://www.who.int/en/news-room/fact-sheets/detail/hepatitis-b (2019).
4. Lok, A. S. F. *et al.* Antiviral therapy for chronic hepatitis B viral infection in adults: A systematic review and meta-analysis. *Hepatology* **63**, 284–306 (2016).
5. Spradling, P., Hu, D. & McMahon, B. J. Epidemiology and prevention. In *Viral Hepatitis* (eds Thomas, H. *et al.*) (Wiley, 2013).
6. Tang, L. S. Y., Covert, E., Wilson, E. & Kottilil, S. Chronic hepatitis B infect a review. *JAMA* **319**, 1802–1813 (2018).
7. Liang, T. J. Hepatitis B: The virus and disease. *Hepatology* **9**, S13–S21 (2009).
8. Kay, A. & Zoulim, F. Hepatitis B virus genetic variability and evolution. *Virus Res.* **127**, 164–167 (2007).
9. Buti, M., Rodriguez-Frias, F., Jardi, R. & Esteban, R. Hepatitis B virus genome variability and disease progression: The impact of pre-core mutations and HBV genotypes. *J. Clin. Virol.* **34**, S79–S82 (2005).
10. Beck, J. & Nassal, M. Hepatitis B virus replication. *World J. Gastroenterol.* **13**, 48–64 (2007).
11. Orito, E. *et al.* Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **86**, 7059–7062 (1989).
12. Cross, J. C., Wen, P. & Rutter, W. J. Transactivation by hepatitis B virus X protein is promiscuous and dependent on mitogen-activated cellular serine/threonine kinases. *Proc. Natl. Acad. Sci. USA* **90**, 8078–8082 (1993).
13. Hu, Z., Zhang, Z., Kim, J. W., Huang, Y. & Liang, T. J. Altered proteolysis and global gene expression in hepatitis B virus X transgenic mouse liver. *J. Virol.* **80**, 1405–1413 (2006).
14. Milich, D. & Liang, T. J. Exploring the biological basis of hepatitis B e antigen in hepatitis B virus infection. *Hepatology* **38**, 1075–1086 (2003).
15. Zhang, Z., Torii, N., Hu, Z., Jacob, J. & Liang, T. J. X-deficient woodchuck hepatitis virus mutants behave like attenuated viruses and induce protective immunity in vivo. *J. Clin. Invest.* **108**, 1523–1531 (2001).
16. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
17. Dienstag, J. L. *et al.* Lamivudine as initial treatment for chronic hepatitis B in the United States. *N. Engl. J. Med.* **341**, 1256–1263 (1999).
18. Chan, H. L. *et al.* Two-year lamivudine treatment for hepatitis B e antigen-negative chronic hepatitis B. a double-blind, placebo-controlled trial. *Antivir. Ther.* **12**, 345–353 (2007).
19. Marcellin, P. *et al.* Long-term efficacy and safety of adefovir dipivoxil for the treatment of hepatitis B e antigen-positive chronic hepatitis B. *Hepatology* **48**, 750–758 (2008).
20. Hadziyannis, S. J. *et al.* Long-term therapy with adefovir dipivoxil for HBeAg-negative chronic hepatitis B for up to 5 years. *Gastroenterology* **131**, 1743–1751 (2006).
21. Tenney, D. J. *et al.* Long-term monitoring shows hepatitis B virus resistance to entecavir in nucleoside-naïve patients is rare through 5 years of therapy. *Hepatology* **49**, 1503–1514 (2009).
22. Buti, M. *et al.* Seven-year efficacy and safety of treatment with tenofovir disoproxil fumarate for chronic hepatitis B virus infection. *Dig Dis Sci.* **60**, 1457–1464 (2015).
23. Bui, Q. C., Nualláin, B. O., Boucher, C. A. & Sloot, P. M. Extracting causal relations on HIV drug resistance from literature. *BMC Bioinf.* **11**, 101 (2010).

24. Khalid, Z. & Sezerman, O. U. ZK DrugResist 20: A TextMiner to extract semantic relations of drug resistance from PubMed. *J Biomed Inform.* **69**, 93–98 (2017).
25. Naderi, N. & Witte, R. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics.* **13**(Suppl 4), S10 (2012).
26. Roberts, K. *et al.* TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *J. Am. Med. Inform. Assoc.* **27**, 1431–1436 (2020).
27. Davey, N. E. *et al.* The HIV mutation browser: A resource for human immunodeficiency virus mutagenesis and polymorphism data. *PLoS Comput. Biol.* **10**, e1003951 (2014).
28. Wang, Y. *et al.* ViMIC: A database of human disease-related virus mutations, integration sites and cis-effects. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkab779 (2021).
29. Mueller-Breckenridge, A. J. *et al.* Machine-learning based patient classification using Hepatitis B virus full-length genome quasispecies from Asian and European cohorts. *Sci Rep.* **9**, 18892 (2019).
30. Caporaso, J. G., Baumgartner, W. A. Jr., Randolph, D. A., Cohen, K. B. & Hunter, L. MutationFinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics* **23**, 1862–1865 (2007).
31. Doughty, E. *et al.* Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics* **27**, 408–415 (2011).
32. Thomas, P., Rocktäschel, T., Hakenberg, J., Lichtblau, Y. & Leser, U. SETH detects and normalizes genetic variants in text. *Bioinformatics* **32**, 2883–2885 (2016).
33. National Center for Biotechnology Information: U.S. National Library of Medicine. Hepatitis B Virus. *NCBI* https://www.ncbi.nlm.nih.gov/genome/?term=hbv (2019).
34. Hayer, J. *et al.* HBVdb: a knowledge database for Hepatitis B Virus. *Nucleic Acids Res.* **41**, D566–D570 (2013).
35. Yuen, L. K. *et al.* SeqHepB: A sequence analysis program and relational database system for chronic hepatitis B. *Antiviral Res.* **75**, 64–74 (2007).
36. EuResist Network GEIE. EuResist Network. *euresist network* https://www.euresist.org/ (2021).
37. Python. xml.etree.ElementTree - The Element Tree XML API. *Python.* https://docs.python.org/3/library/xml.etree.elementtree.html (2019).
38. National Center for Biotechnology Information: U.S. National Library of Medicine. PMC. *NCBI* https://www.ncbi.nlm.nih.gov/pmc/ (2019).
39. World Medical Association. World medical association declaration of Helsinki. *JAMA* **310**, 2191–2194 (2013).
40. Vijayananthan, A. & Nawawi, O. The importance of Good Clinical Practice guidelines and its role in clinical trials. *Biomed. Imaging Interv. J.* **4**, e5 (2008).
41. Clarivate Analytics. Integrity. *Clarivate Analytics* https://integrity.clarivate.com//integrity/xmlxsl/ (accessed on August 14, 2019).
42. GitHub. HBV_Code. *GitHub* https://github.com/angoto/HBV_Code (2021).
43. Yasutake, Y. *et al.* HIV-1 with HBV-associated Q151M substitution in RT becomes highly susceptible to entecavir: Structural insights into HBV-RT inhibition by entecavir. *Sci Rep.* **8**, 1624–1624 (2018).
44. Yasutake, Y. *et al.* Active-site deformation in the structure of HIV-1 RT with HBV-associated septuple amino acid substitutions rationalizes the differential susceptibility of HIV-1 and HBV against 4′-modified nucleoside RT inhibitors. *Biochem. Biophys. Res. Commun.* **509**, 943–948 (2019).
45. Bienert, S. *et al.* The SWISS-MODEL Repository – new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017).
46. UniProt. UniProtKB - Q9WRJ9 (Q9WRJ9_HBV). *UniProt* https://www.uniprot.org/uniprot/Q9WRJ9 (2020).
47. The PyMOL Molecular Graphics System, Version 2.3.0 Schrödinger, LLC.
48. Sevier, C. S. & Kaiser, C. A. Formation and transfer of disulfide bonds in living cells. *Nat. Rev. Mol. Cell Biol.* **3**, 836–847 (2002).
49. Nassal, M., Rieger, A. & Steinau, O. Topological analysis of the hepatitis B virus core particle by cysteine-cysteine cross-linking. *J. Mol. Biol.* **225**, 1013–1025 (1992).
50. Mangold, C. M. & Streeck, R. E. Mutational analysis of the cysteine residues in the hepatitis B virus small envelope protein. *J. Virol.* **67**, 4588–4597 (1993).
51. Mangold, C. M., Unckell, F., Werr, M. & Streeck, R. E. Analysis of intermolecular disulfide bonds and free sulfhydryl groups in hepatitis B surface antigen particles. *Arch. Virol.* **142**, 2257–2267 (1997).
52. Mangold, C. M., Unckell, F., Werr, M. & Streeck, R. E. Secretion and antigenicity of hepatitis B virus small envelope proteins lacking cysteines in the major antigenic region. *Virology* **211**, 535–543 (1995).
53. Langley, D. R. *et al.* Inhibition of hepatitis B virus polymerase by entecavir. *J. Virol.* **81**, 3992–4001 (2007).
54. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282**, 1669–1675 (1998).
55. Guermouche, H. *et al.* Characterization of the dynamics of human cytomegalovirus resistance to antiviral drugs by ultra-deep sequencing. *Antiviral Res.* **173**, 104647 (2020).
56. da Silva, J. *et al.* Monitoring emerging human immunodeficiency virus drug resistance in Sub-Saharan Africa in the era of Dolutegravir. *J. Infect. Dis.* **225**(3), 364–366 (2022).
57. Popping, S. *et al.* The need for a European hepatitis C programme monitoring resistance to direct-acting antiviral agents in real life to eliminate hepatitis C. *J. Virus Erad.* **4**(3), 179–181 (2018).
58. Lina, B. *et al.* Five years of monitoring for the emergence of oseltamivir resistance in patients with influenza A infections in the Influenza Resistance Information Study. *Influenza Other Respir Viruses.* **12**(2), 267–278 (2018).
59. Mari A. *et al.* Global surveillance of potential antiviral drug resistance in SARS-CoV-2: proof of concept focussing on the RNA-dependent RNA polymerase. medRxiv 2020.12.28.20248663.
60. Lai, C. L. *et al.* A 1-year trial of telbivudine, lamivudine, and the combination in patient with hepatitis B e antigen-positive chronic hepatitis B. *Gastroenterology* **129**, 528–536 (2005).
61. Kim, H. S. *et al.* Management of entecavir-resistant chronic hepatitis B with adefovir-based combination therapies. *World J. Gastroenterol.* **21**, 10874–10882 (2015).
62. Angus, P. *et al.* Resistance to adefovir dipivoxil therapy associated with the selection of a novel mutation in the HBV polymerase. *Gastroenterology* **125**, 292–297 (2003).
63. Lok, A. S. & McMahon, B. J. Chronic hepatitis B: update of recommendations. *Hepatology* **39**, 857–861 (2004).
64. Elsevier. ScienceDirect https://www.sciencedirect.com/ (2019).
65. Springer Nature. Springer Nature https://www.springernature.com/gp (2019).
66. Zhang, X. *et al.* Pre-existing mutations related to tenofovir in chronic hepatitis B patients with long-term nucleos(t)ide analogue drugs treatment by ultra-deep pyrosequencing. *Oncotarget* **7**, 70264–70275 (2016).
67. Lin, C., Chien, R., Hu, C., Lai, M. & Yeh, C. Identification of hepatitis B virus rtS117F substitution as a compensatory mutation for rtM204I during lamivudine therapy. *J Antimicrob Chemother.* **67**, 39–48 (2012).
68. Cejuela, J. M. *et al.* nala: Text mining natural language mutation mentions. *Bioinformatics* **33**, 1852–1858 (2017).
69. Wei, C. H. *et al.* tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics* **34**, 80–87 (2018).

## Acknowledgements
We thank the Roche Innovation Center Basel for their computational support.

## Author contributions
A.G., R.R., G.M.M., and S.S. contributed to the conceptual idea of the study. A.G., R.R., and G.M.M. conceived and designed the methods. A.G. performed the computational work and analyzed the results. A.G., R.R., and G.M.M. wrote the paper. All authors read and approved the final manuscript.

## Funding
A.G. and G.M.M. thank the EPSRC and MRC Centre for Doctoral Training (CDT) in Systems Approaches to Biomedical Science (EP/L016044/1). A.G. thanks the Clarendon and Oxford Kobe Scholarships. G.M.M. also thanks the EPSRC CDT in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research—SABS:R$^3$ (EP/S024093/1).

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-17746-3.

**Correspondence** and requests for materials should be addressed to G.M.M.