Antiviral Research 88 (2010) 269–275

Contents lists available at ScienceDirect

Antiviral Research

journal homepage: www.elsevier.com/locate/antiviral

Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery

Soo-Yon Rhee[a,*], Severine Margeridon-Thermet[a], Mindie H. Nguyen[a], Tommy F. Liu[a], Ron M. Kagan[b], Bastian Beggel[c], Jens Verheyen[d], Rolf Kaiser[d], Robert W. Shafer[a]

[a] Department of Medicine, Stanford University, Stanford, CA, United States
[b] Quest Diagnostics Incorporated, San Juan Capistrano, CA, United States
[c] Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrucken, Germany
[d] Institute of Virology, University of Cologne, Cologne, Germany

ARTICLE INFO

Article history:
Received 14 May 2010
Received in revised form 7 September 2010
Accepted 17 September 2010

Keywords:
Hepatitis B virus
Reverse transcriptase (RT)
Drug resistance mutation
Database
Nucleoside and nucleotide RT inhibitors (NRTIs, NtRTIs)

ABSTRACT

Drug resistance resulting from reverse transcriptase (RT) mutations is one of the main obstacles to successful hepatitis B virus (HBV) therapy. Indeed, HBV treatment guidelines recommend HBV genotypic resistance testing for patients receiving nucleos(t)ide RT inhibitors (N(t)RTIs) who develop virological failure. N(t)RTI-resistance mutations at 10 RT positions have been well characterized in phenotypic studies, however, data are lacking on the relative frequency of these mutations in N(t)RTI-treated and untreated individuals. There are also few published data on the extent of amino acid variation at most of the 344 positions of HBV RT and the extent to which this variation is influenced by N(t)RTI treatment. We retrieved 23,871 HBV RT sequences from GenBank and reviewed the published reports of these sequences to ascertain the number of individuals from whom the sequences were obtained, the N(t)RTI treatments of these individuals, and the year and region of virus sampling. We then used these data to populate a relational database we named HBVrtDB. As of July 2010, HBVrtDB contained 6811 sequences from 3869 individuals reported in 281 references. Among these 3869 individuals, 73% were N(t)RTI-naïve and 27% received one or more N(t)RTIs. Among the 10 well-characterized N(t)RTI-resistance mutations, L80I/V, V173L, L180M, A181T, T184S, S202G and M204I/V were significantly associated with treatment with lamivudine, an L-nucleoside analog, and A181S/T/V and N236T were significantly associated with treatment with adefovir, an acyclic nucleoside phosphonate. A similar analysis of ten additional less well-characterized resistance mutations demonstrated a significant association with N(t)RTI treatment for four of the mutations: L82M, S85A, A200V, and Q215S. We also created an interactive program, HBVseq, to enable users to identify mutations in submitted sequences and retrieve the prevalence of these mutations in HBVrtDB according to genotype and N(t)RTI treatment. HBVrtDB and HBVseq are available at http://hivdb.stanford.edu/HBV/releaseNotes/.

1. Introduction

Hepatitis B virus (HBV) infects more than 400 million people worldwide and is a leading cause of mortality as a result of cirrhosis and hepatocellular carcinoma. Within the past 12 years, five nucleos(t)ide RT inhibitors (N(t)RTIs) have been licensed for HBV treatment including lamivudine (3TC), adefovir (ADV), entecavir (ETV), telbivudine (LdT), and tenofovir (TDF). Emtricitabine (FTC), which is structurally similar to 3TC, is also active against HBV and is frequently used to treat HBV because it is co-formulated with TDF (as truvada) for HIV treatment. 3TC, FTC, and LdT are L-nucleoside analogs. ETV is a deoxyguanosine analog. ADV and TDF are acyclic nucleoside phosphonates (ANPs). HBV resistance is one of the obstacles to successful anti-HBV therapy. HBV RT is functionally and structurally similar to HIV-1 RT (Bartholomeusz et al., 2004; Das et al., 2001) and has an error rate similar to that of other retroviral polymerases (Günther et al., 1999). The current HBV treatment guidelines recommend HBV genotypic resistance testing for patients who experience primary or secondary virological failure while receiving N(t)RTIs (Keeffe et al., 2008a,b; Lok and McMahon, 2009; Lok et al., 2007). Although about 15 mutations

* Corresponding author at: Division of Infectious Diseases, Room S-169, Stanford University Medical Center, 300 Pasteur Drive, Stanford, CA 94305, United States. Tel.: +1 650 736 0911; fax: +1 650 725 2088.
E-mail addresses: syrhee@stanford.edu (S.-Y. Rhee), severine@stanford.edu (S. Margeridon-Thermet), mindiehn@stanford.edu (M.H. Nguyen), tliu@stanford.edu (T.F. Liu), Ron.M.Kagan@questdiagnostics.com (R.M. Kagan), beggel@mpi-inf.mpg.de (B. Beggel), jens.verheyen@uk-koeln.de (J. Verheyen), rolf.kaiser@uk-koeln.de (R. Kaiser), rshafer@stanford.edu (R.W. Shafer).

0166-3542/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.antiviral.2010.09.012

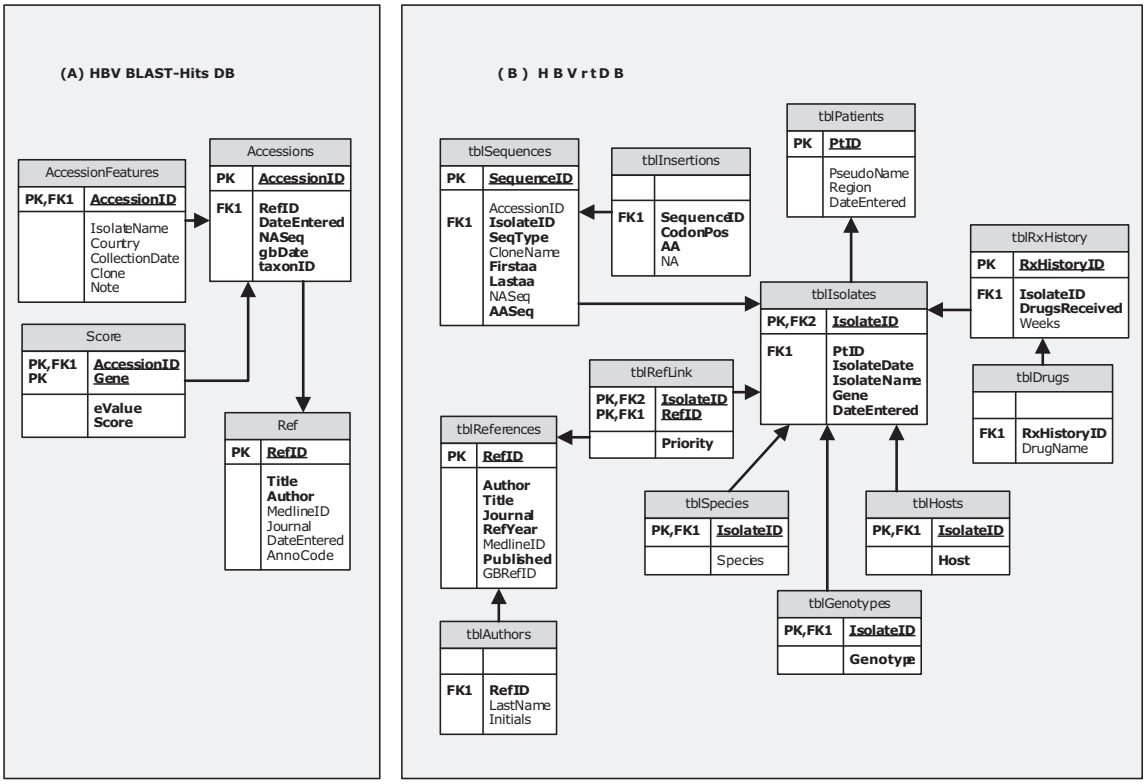

Fig. 1. HBV BLAST-Hits DB and HBVrtDB schemas. HBVrtDB was seeded with a blast search of the GenBank viral sequence files using an HBV RT amino acid sequence. Filtered BLAST hits were aggregated by the GenBank reference field and parsed to create a relational database, HBV BLAST-Hits DB (Fig. 1A). Each set of sequences from a reference was annotated and the resulting annotation was used to populate the tables in HBVrtDB (Fig. 1B).

at 10 positions are strongly associated with decreased HBV N(t)RTI susceptibility (Keeffe et al., 2008b; Lok et al., 2007), the relative frequencies of those drug resistance mutations in N(t)RTI-treated and N(t)RTI-untreated individuals are not known. Moreover, the association of some less-well characterized possible drug-resistance mutations with N(t)RTI treatment is uncertain.

There are at least eight HBV genotypes that differ from one another by about 10% of their nucleotides (Kurbanov et al., 2010). There are two HBV sequence databases that allow users to download genotype-specific alignments of different HBV genes (Gnaneshan et al., 2007; Shin et al., 2008). However, these databases do not contain information on the N(t)RTIs received by the individuals from whom the sequenced viruses were obtained. In addition, the sequence data in these databases are not linked to the additional data in the references from which the sequences were reported. A third database contains proprietary sequence data and associated clinical information that are available only for registered users (Yuen et al., 2007).

We created an HBV RT variant database, HBVrtDB to feature associations between HBV RT mutations and the N(t)RTI treatments of the individuals from whom the sequences were obtained. The database presents these associations within the context of viral genotype and geographic origin. We also created an interactive program, HBVseq, to enable users to identify mutations in submitted sequences and retrieve the prevalence of these mutations in HBVrtDB according to genotype and N(t)RTI treatment.

2. Methods

2.1. Sequence retrieval and annotation

A local BLAST search using an HBV RT amino acid sequence was performed using the GenBank viral sequence files. The BLAST search results were aggregated by the GenBank reference field and were imported to seed a relational database we call the HBV BLAST-Hits DB. Each reference in the HBV BLAST-Hits DB was annotated according to whether the set of sequences in the reference was obtained from one or more than one individual, which we refer to as the provenance of the sequence. Sequences from an individual were then annotated according to whether they were obtained at the same or different times. Sequences were next annotated according to the year and country of sampling, N(t)RTI treatments received by the individuals from whom the sequences were obtained, and the method of sequencing. Sequences that could be fully annotated, that encompassed at least RT codons 180–240, and that were obtained from neither laboratory viruses nor defective genomes were added to HBVrtDB. An initial seed BLAST search was performed on the April 15th version of the GenBank viral sequence files and then supplemented with a second incremental BLAST search performed on the June 15th version of the GenBank viral sequence files. As part of this study, we supplemented HBVrtDB with sequences from 443 individuals from one clinic in the U.S. and one in Germany (GenBank accession numbers: HM173808-HM174250).

To facilitate bimonthly incremental updates of HBVrtDB, each reference in the HBV BLAST-Hits DB is annotated according to whether the sequences from the reference have been exported to HBVrtDB. For the sequences that were not added to the HBVrtDB, one of the following annotations is assigned: (i) gene fragment: sequences that do not encompass RT positions 180–240. These sequences were typically from studies of the S gene; (ii) HBV treatment and/or other data not available: sequences in GenBank lacking the required annotation; (iii) unpublished: sequences in GenBank that were not linked to a published paper and that contained insufficient data within the GenBank entry for inclusion in HBVrtDB; (iv) sequences of laboratory viruses; and (v) chromosomally integrated HBV gene sequences.

Table 1
Genotypes, nucleos(t)ide RT inhibitor (N(t)RTI) treatments, and geographic origin of the viral sequences in HBVrtDB.

N(t)RTI-naïve

Genotype	No. persons	Region[a]
A	445	Poland (73), Haiti (57), Germany (48), USA (45), South Africa (38), Africa (26), France (16), Czech Republic (11), Laos (9), India (9)
B	530	USA (207), China (72), Malaysia (50), Japan (49), Canada (40), Vietnam (19), Taiwan (17), Australia (14), Indonesia (12), Laos (8)
C	707	China (214), Japan (101), USA (81), South Korea (62), Hong Kong (49), Canada (28), Malaysia (25), Laos (21), Australia (20), Indonesia (20)
D	759	Tunisia (111), Bulgaria (76), Germany (66), Turkey (50), Iran (41), Haiti (38), Greece (32), Belarus (31), New Zealand (28), Russia (27)
E	227	Guinea (78), Africa (65), Niger (19), Ghana (18), Haiti (12), France (6), Namibia (6), Germany (5), UK (4), Cameroon (4)
F	83	Venezuela (27), Colombia (12), Bolivia (11), Brazil (8), Spain (5), Peru (3), USA (3), Nicaragua (2), Japan (2), Costa Rica (2), France (2), Panama (2)
G	23	Germany (6), Mexico (6), Japan (5), USA (4), South Africa (1), Spain (1)
H	24	Mexico (8), Japan (7), USA (5), Nicaragua (2), Czech Republic (1), Poland (1)
I	30	Laos (23), Vietnam (4), Canada (2), France (1)

N(t)RTI-treated[b]

N(t)RTIs received[b]	No. persons	Region[a]
3TC	436	Germany (102), Italy (54), Brazil (53), Rwanda (44), Japan (39), USA (29), Bulgaria (20), Taiwan (17), Canada (16), Poland (15)
3TC, ADV	145	Germany (45), Japan (30), USA (28), France (15), Taiwan (10), Italy (7), Canada (4), Spain (3), China (1), Poland (1), Greece (1)
ADV	90	USA (43), Germany (21), France (15), China (9), Singapore (1), Spain (1)
3TC, TDF	39	Germany (29), Spain (10)
ETV	11	USA (7), Germany (3), France (1)
3TC, ETV	11	USA (6), China (2), Germany (1), Japan (1), France (1)
Other	12	USA (11), Germany (1)

3TC: lamivudine; ADV: adefovir; TDF: tenofovir; ETV: entecavir; FTC: emtricitabine.
[a] The top 10 regions from which most of HBV RT sequences were obtained are shown; number in parentheses indicates the number of patients from which HBV RT sequences were obtained in that region.
[b] N(t)RTIs received prior to viral sequencing. In some cases, individuals may not have been on the indicated N(t)RTIs at the time of virus sequencing.

2.2. Sequence variation analysis

A genotype was assigned to each sequence in HBVrtDB by calculating the uncorrected nucleotide distance to a set of reference genotype RT sequences (http://hivdb.stanford.edu/HBV/HBVseq/development/input/misc/RTsubtypes_refNA.txt). For this analysis, we used the eight well-established genotypes A–H (Kurbanov et al., 2010), and the recently described genotype I which has recently been reported from several populations in South-East Asia (Schaefer et al., 2009). Viruses that were labeled as recombinants in the GenBank record were assigned the user-defined genotype. Each nucleotide sequence was translated and the translated amino acid sequence was compared with the corresponding consensus reference genotype amino acid sequence (http://hivdb.stanford.edu/HBV/HBVseq/development/input/misc/RTgenotypes_consensusAA.fasta.txt). Differences between each sequence and the consensus reference genotype sequence were classified as mutations.

2.3. Mutation analysis according to genotype and treatment

The prevalence of each mutation in sequences from N(t)RTI-naïve individuals was calculated for each genotype. Positions at which amino acid variation was present at more than 0.5% of viruses from N(t)RTI-naïve individuals were considered polymorphic. The prevalence of each mutation was also calculated in sequences from N(t)RTI-experienced individuals with one of the following three types of treatment histories: (i) treatment with one or more L-nucleosides and/or ETV but no ANPs; (ii) treatment with one or more ANPs but no L-nucleoside or ETV; or (iii) treatment with an L-nucleoside and/or ETV as well as with one or more ANPs.

For the purposes of calculating mutation prevalence, we performed the following analyses: (i) when sequences were available before and after an individual received an N(t)RTI, the pre-therapy sequences were used for calculating prevalence in untreated individuals and the post-therapy sequences were used for calculating prevalence in treated individuals; (ii) when an individual had more than one pre-therapy sequence or more than one post-therapy sequence with the same mutation, that mutation was counted once for each category of therapy; and (iii) when sequences of multiple clones of the same virus isolate were present, a consensus sequence containing the most common amino acid at each position was used.

Statistical associations between mutations and N(t)RTI-treatment were assessed using the Fisher's Exact test combined with the method of Benjamini and Hochberg to control the false-discovery rate (FDR) at 0.01 (Benjamini and Hochberg, 1995). We analyzed the following mutations at 10 well-characterized DRM positions L80IV, I169T, V173L, L180M, A181TV, T184SAILFG, S202GI, M204VIS, N236T, and M250V (Keeffe et al., 2008b; Lok et al., 2007). We also analyzed the mutations that have been reported to possibly contribute to resistance at 10 additional positions: L82M, V84M, S85A, A194T, A200V, V214A, Q215S, I233V, P237H and NASH238TD (Liu et al., 2010; Ogata et al., 1999; Schildgen et al., 2006; Shaw et al., 2006; Sheldon et al., 2005).

3. Results

3.1. Database

Our filtered BLAST search program produced the seed HBV BLAST-Hits DB. This database contained 23,871 HBV RT sequences from 761 references. Fig. 1A shows the schema of the HBV BLAST-Hits DB that contains four tables with data from the GenBank BLAST search. Of the 761 references in the HBV BLAST-Hits DB, sufficient annotation for HBVrtDB inclusion was available for 281 (37%) references containing 6811 sequences from 3896 individuals. Fig. 1B shows the schema of HBVrtDB which contains the annotations for each of the GenBank sequences including the provenance of the sequence, the year the sequenced sample was obtained, the country from which the sequenced sample was obtained, and the N(t)RTIs received by the individual from whom the sequence was obtained. The following web page (http://hivdb.stanford.edu/HBV/DB/cgi-

Pos	N(t)RTI Naive Persons (by Genotype)										N(t)RTI Treated Persons (by Treatment)		
	A 438	B 531	C 705	D 753	E 227	F 82	G 23	H 24	I 32	pooled 2805	L-Nuc. and/or ETV 458	ANPs 90	L-Nuc. and/or ETV + ANPs 191
80	L	L	L	L	L	L	L	L	L	L	L **I**$^{10.4}$ **V**$^{2.9}$	L	L I$^{6.4}$ V$^{1.3}$
169	I	I	I	I	I	I	I	I	I	I	I	I	I
173	V	V	V	V	V	V	V	V	V	V	V **L**$^{4.8}$	V	V L$^{4.8}$ M$^{0.5}$
180	L	L	L	L	L	L	L	L	L	L	L **M**$^{31.7}$	L M$^{1.1}$	L M$^{29.1}$
181	A	A	A	A	A	A S$^{1.2}$	A	A	A	A	A T$^{0.9}$	A **V**$^{11.1}$ **T**$^{3.3}$ **S**$^{2.2}$	A V$^{3.2}$ T$^{1.6}$ S$^{1.1}$
184	T	T	T	T	T	T I$^{1.2}$	T	T	T	T	T **S**$^{1.1}$	T	T S$^{1.6}$ A$^{0.5}$
202	S	S	S	S	S	S N$^{1.2}$	S	S	S	S	S G$^{0.7}$	S	S G$^{1.0}$
204	M I$^{0.9}$	M	M	M	M	M	M	M	M	M	M **V**$^{21.7}$ **I**$^{20.6}$	M I$^{1.1}$	M V$^{24.0}$ I$^{13.7}$
236	N	N	N	N	N	N	N	N	N	N	N	N **T**$^{14.4}$	N T$^{2.2}$ R$^{0.6}$
250	M	M	M	M	M	M	M	M	M	M	M	M	M L$^{0.6}$

Fig. 2. Prevalence of well-characterized drug resistance mutations in HBVrtDB according to genotype and treatment. A schematic representation of the type of data produced by HBVseq or available in the "HBV RT Mutations According to Genotype and Treatment" page. This figure shows ten rows of data corresponding to the ten most widely recognized HBV N(t)RTI-resistance positions (Keeffe et al., 2008b; Lok et al., 2007). The first column indicates the RT position and columns 2 through 10 indicate the percent prevalence of mutation at these 10 positions in N(t)RTI-naïve individuals according to virus genotype. Column 11 contains the pooled (across all genotypes) percent prevalence of mutations in N(t)RTI-naïve individuals. Column 12 contains the percent prevalence of each mutation in individuals who received L-nucleosides and/or ETV but no acyclic nucleoside phosphonates (ANPs). Column 13 contains the percent prevalence of each mutation in individuals who received ANPs but no L-nucleosides or ETV. Column 14 contains the percent prevalence of each mutation in individuals who received L-nucleoside and/or ETV as well as one or more ANPs. In each column, the consensus amino acid is shown at the top of each cell and reported variants along with their percent prevalence (shown as superscripts) are indicated below the consensus. Each mutation represents a hyperlink to the references in which the mutation was reported. Mutations significantly associated with L-nucleosides and/or ETV (column 12) or ANPs (column 13) are shown in bold.

bin/gbReferenceHBVRT.cgi) contains a summary of all references in the HBV BLAST-Hits DB according to whether they have been added to HBVrtDB.

Table 1 is a summary of the virus sequences in HBVrtDB by their genotypes and the N(t)RTI treatment and geographic origins of the individuals from whom the sequenced viruses were obtained. Of the 3896 individuals in the database, 73% (2848) individuals were NRTI-naïve and 27% (1048) received one or more N(t)RTIs: 62% received L-nucleosides or ETV, 12% received ANPs, and 26% received N(t)RTIs of both classes.

3.2. Sequence variation in viruses from N(t)RTI-naïve individuals

The mean uncorrected pair-wise amino acid differences between HBV RT genotypes from N(t)RTI-naïve individuals ranged from 8% to 11% for all genotype pairs with three exceptions: A vs G (6%), E vs G (6%), F vs H (5%) and G vs I (5%). The mean intra-genotype nucleotide distance was 1% for genotypes E and H, 2% for genotype G and I, and 3% for genotypes A, B, C, D and F.

Among the 344 RT positions in sequences from the N(t)RTI-naïve individuals, 59 (17.2%) positions had different genotype-specific consensus amino acids. At 53 (15.4%) positions each of the genotypes had the same consensus but differences from the consensus were present in ≥0.5% of the pooled viruses from untreated persons. Among the remaining 232 (67.4%) positions without polymorphisms, 87 positions were nearly completely conserved containing no more than one virus with a non-consensus mutation, whereas 145 positions had viruses with more than one non-consensus mutation but none of these mutations were present at a level of >0.5%.

3.3. Mutations associated with therapy

The web page "HBV RT Mutation Prevalence According to Genotype and Treatment" (http://hivdb.stanford.edu/HBV/DB/cgi-bin/MutPrevByGenotypeRxHBV.cgi) dynamically generates a tabular summary of mutational data for each of the 344 HBV RT positions. Column 1 contains the RT position. Columns 2–10 contain the percent prevalence of each mutation in N(t)RTI-naïve individuals infected according to genotype (A–I). Column 11 contains the percent prevalence of each mutation in the pooled sequences of different genotypes from N(t)RTI-naïve individuals. Column 12 contains the percent prevalence of each mutation in individuals who received L-nucleosides and/or ETV but no ANPs. Column 13 contains the percent prevalence of each mutation in individuals who received ANPs but no L-nucleosides or ETV. Column 14 contains the percent prevalence of each mutation in individuals who received L-nucleoside and/or ETV as well as one or more ANPs (either in sequence or combination). The total numbers of individuals from whom sequences were available in HBVrtDB are listed in the table header.

The top of each cell in columns 2–10 contains the consensus amino acid for the position by genotype. However, because columns

Pos	N(t)RTI Naive Persons (by Genotype)										N(t)RTI Treated Persons (by Treatment)		
	A 438	B 531	C 705	D 753	E 227	F 82	G 23	H 24	I 32	pooled 2805	L-Nuc. and/or ETV 458	ANPs 90	L-Nuc. and/or ETV + ANPs 191
82	L R$^{0.5}$	L	L	L	L	L	L	L S$^{8.3}$	L	L	L **M$^{1.0}$**	L	L M$^{1.9}$
84	V	V	V I$^{0.5}$	V	V	V	V	V	V	V	V	V	V
85	S	S	S T$^{0.6}$	S	S	S	S	S	S	S	S	S **A$^{1.7}$**	S A$^{1.9}$
194	A	A	A	A	A	A	A	A	A	A	A	A	A S$^{0.5}$ T$^{0.5}$
200	A	A	A	A	A	A	A	A	A	A	A **V$^{1.1}$**	A	A V$^{0.5}$
214	V A$^{0.5}$ I$^{0.5}$	V	V I$^{0.6}$	V A$^{0.8}$ E$^{0.7}$	V E$^{0.9}$	V P$^{1.2}$	V	V	V	V A$^{0.5}$	V A$^{0.7}$	V	V
215	Q	Q H$^{0.9}$	Q	Q S$^{4.3}$ H$^{3.0}$ P$^{1.9}$	Q	Q	Q	Q	Q	Q S$^{1.1}$ H$^{1.0}$ P$^{0.5}$	Q **S$^{3.9}$** H$^{1.1}$	Q S$^{4.6}$ H$^{1.1}$	Q S$^{1.6}$
233	I	I V$^{0.6}$	I V$^{1.3}$	I V$^{1.1}$	I	I	I	I	I	I V$^{0.7}$	I V$^{0.5}$	I V$^{1.1}$	I V$^{0.5}$
237	P	P	P	P T$^{6.4}$	P S$^{0.9}$	T	P	T	P	P T	P T	P T	P T H$^{0.5}$ C$^{0.5}$
238	N H$^{1.0}$ S$^{1.0}$ T$^{0.7}$	H Q$^{3.9}$	N T$^{8.7}$ A$^{2.4}$ H$^{2.4}$ S$^{0.6}$	N H$^{2.3}$ D$^{0.9}$ S$^{0.8}$ T$^{0.8}$	N H$^{1.8}$ T$^{1.8}$ A$^{0.5}$ S$^{0.5}$	S A$^{17.7}$ F$^{1.3}$	N T$^{10.5}$ D$^{5.3}$	A	N	N A S H T$^{2.8}$ Q$^{0.8}$	N A S H T$^{1.7}$	N A S H T$^{6.8}$ D$^{2.3}$ E$^{1.1}$	N A S H T$^{4.4}$ D$^{0.5}$

Fig. 3. Prevalence of less well-characterized drug resistance mutations in HBVrtDB according to genotype and treatment. A schematic representation of the type of data produced by HBVseq or available in the "HBV RT Mutations According to Genotype and Treatment" page. This figure shows ten rows of data corresponding to the ten less well-characterized drug-resistance mutations (see text). The first column indicates the RT position and columns 2 through 10 indicate the percent prevalence of mutation at these 10 positions in N(t)RTI-naïve individuals according to virus genotype. Column 11 contains the pooled (across all genotypes) percent prevalence of mutations in N(t)RTI-naïve individuals. Column 12 contains the percent prevalence of each mutation in individuals who received L-nucleosides and/or ETV but no acyclic nucleoside phosphonates (ANPs). Column 13 contains the percent prevalence of each mutation in individuals who received ANPs but no L-nucleosides or ETV. Column 14 contains the percent prevalence of each mutation in individuals who received L-nucleoside and/or ETV as well as one or more ANPs. In each column, the consensus amino acid is shown at the top of each cell and reported variants along with their percent prevalence (shown as superscripts) are indicated below the consensus. Each mutation represents a hyperlink to the references in which the mutation was reported. Mutations significantly associated with L-nucleosides and/or ETV (column 12) or ANPs (column 13) are shown in bold.

11–14 contain data belonging to pooled genotypes, these cells often contain more than one consensus amino acid at the top of the cell. For all columns with mutational prevalence data, differences from the consensus amino acid are listed and the prevalence of each mutation is indicated as a superscript. Each of these mutations is also a hyperlink to a page that contains the references, accession numbers, countries of origin and genotypes of sequences containing the mutation and complete N(t)RTI treatment of individuals from whom the sequences were obtained.

Fig. 2 shows the prevalence of amino acid variants in HBVrtDB at 10 well-characterized HBV RT drug-resistance positions (Keeffe et al., 2008b; Lok et al., 2007) by genotype and N(t)RTI treatment history. Twelve mutations at the eight of these 10 positions were significantly associated with N(t)RTI therapy (Fisher's Exact test; Benjamini–Hochberg adjusted p value <0.01): L80I/V, V173L, L180M, A181T, T184S, S202G and M204I/V were associated with the L-nucleoside analog lamivudine; A181S/T/V and N236T were associated with the ANP adefovir. Too few sequences from entecavir-treated individuals were available to identify treatment associations for two (I169T and M250V) of the four (I169T, T184SAILFG, S202GI, M250V) entecavir-associated mutation positions. Each of these ten well-characterized drug-resistance mutation positions was non-polymorphic in untreated individuals.

Fig. 3 shows the prevalence of amino acid variants at ten less well-characterized possible drug-resistance mutations (Ogata et al., 1999; Schildgen et al., 2006; Shaw et al., 2006; Sheldon et al., 2005). For these mutations, a significant association with L-nucleosides was present for L82M, A200V and Q215S and an association with ANPs was present for S85A (Fisher's Exact test; Benjamini–Hochberg adjusted p value <0.01). Supporting these associations were the findings that each of the viruses with L82M or A200V had one or more established L-nucleoside mutation (L180M and/or M204IV) and that each of the viruses with S85A had the ANP mutation N236T. Additionally, nine of the 17 mutations with Q215S had one or more L-nucleoside associated mutations. Among these ten less well-characterized mutations, three were polymorphic (prevalence >0.5%) in the 2804 pooled untreated individuals: Q215S (1.1%), I233V (0.7%), and NASH238T (2.8%).

3.4. HBVseq

We created an on-line program HBVseq (http://hivdb.stanford.edu/HBV/HBVseq/development/HBVseq.html) that enables clinical

Fig. 4. The HBVseq user interface. The HBVseq user interface including the sequence submission form (Fig. 4A); program output showing the genotype of the submitted sequence, the established drug resistance RT mutations identified in the sequence and a table showing the percent prevalence of each identified mutation in HBVrtDB according to genotype and nucleos(t)ide RT inhibitor (N(t)RTI) treatment (Fig. 4B); and the detailed sequence information page accessible through each mutation's hyperlink (Fig. 4C). The example shown here is a genotype G sequence containing a well-characterized drug resistance mutation V173L. The circled text in (Fig. 4B) shows the percent prevalence of V173L in L-nucleoside and/or ETV-treated individuals. By clicking the V173L hyperlink (Fig. 4B), the list of published reports of V173L is shown on a new page (Fig. 4C).

laboratories and researchers to analyze new HBV RT sequences in the context of previously published sequence data. Users can enter up to 100 HBV RT sequences in fasta format. For each of the submitted sequences, the HBVseq output includes: (a) summary data comprising the genotype of the sequence and a list of well-characterized drug-resistance mutations in the sequence (Keeffe et al., 2008b; Lok et al., 2007); (b) an assessment of sequence quality; and (c) a table containing the mutation percent prevalence in HBVrtDB by genotype and treatment for each identified mutation. HBVseq also creates a phylogenetic tree with all submitted sequence(s) and a set of representative sequences belonging to each of the nine genotypes.

The genotype is determined by comparisons to reference sequences (http://hivdb.stanford.edu/HBV/HBVseq/development/input/misc/RTsubtypes_refNA.txt). For sequences with less than 96% nucleoside identity to each of the reference sequences, a note is provided together with links to two other online genotyping tools that are a capable of detecting recombinant sequences. The "Sequence Quality Assessment" section lists stop codons, frame shifts, highly ambiguous nucleotides, and mutations at nearly completely conserved positions defined as positions at which no more than one mutation has been reported in viruses from N(t)RTI-naïve and N(t)RTI-experienced individuals (http://hivdb.stanford.edu/HBV/HBVseq/development/input/profiles/conserved.positions.txt).

The columns in table containing the mutation prevalence in HBVrtDB according to genotype and treatment are the same as those in table "Mutation Prevalence According to Genotype and Treatment". Each mutation in this table is also a hyperlink to the references that report the mutation. Fig. 4 summarizes the flow of information to the user showing the sequence upload form (Fig. 4A), the genotype and mutation percent prevalence output (Fig. 4B), and the detailed isolate information page accessible through each mutation's hyperlink (Fig. 4C). Sequences submitted to HBVSeq for analysis are not stored on the server.

4. Discussion

HBVrtDB was constructed by curating and annotating more than 250 studies in GenBank and was supplemented by the contribution of well-characterized HBV sequences from two large clinical populations. HBVrtDB provides novel data on the extent of polymorphism at each RT position according to HBV genotype and on the relative prevalence of each of the well-characterized drug-resistance mutations in individuals who received L-nucleoside and/or ETV but no ANPs and in individuals who received ANPs but no L-nucleosides or ETV. The statistical associations with therapy for the well-characterized mutations provide strong evidence of the validity of the data in HBVrtDB.

Although we cannot be sure that the presence of a mutation in a virus from an individual who received a specific N(t)RTI resulted from selective drug pressure, the large compilation of sequences also makes it possible to test putative associations between specific mutations and drug therapy and to identify new associations that should be further tested in vitro. This was demonstrated in the evaluation of the ten less well-characterized possible drug-resistance mutations in which a treatment association was detected for four of the ten mutations. As the number of sequences from N(t)RTI-treated and untreated individuals belonging to different genotypes increases, it may become possible to evaluate treatment associations independently within viruses belonging to different genotypes and to distinguish mutation profiles to individual drugs within HBV N(t)RTI drug classes.

HBVseq interrogates HBVrtDB to assist researchers and clinical laboratories that perform HBV RT sequencing and that are confronted with large numbers of sequence variants that can be difficult to interpret. HBVseq identifies mutations in submitted sequences, returns their prevalence in HBVrtDB according to genotype and treatment, and links to the references that report the mutations. HBVseq was modeled after a similar program we previously created for HIV drug resistance (Shafer et al., 2000). By examining the prevalence of mutations in HBVrtDB, users can quickly check if the mutations identified in their sequences are common polymorphisms vs rare variants that may reflect sequencing error or potentially new drug-resistance mutations. HBVseq demonstrates that the body of published sequence data on HBV RT can be made available in real time to clinical laboratories and researchers sequencing new HBV RT isolates.

5. Conclusion

HBVrtDB demonstrates that the large numbers of sequences in GenBank – when properly annotated – provides a powerful tool for mutation discovery and sequence analysis. Analysis of the aggregate data in HBVrtDB makes it possible to test statistical associations between HBV RT mutations and treatment exposure and to generate hypotheses about mutations that can be further tested in vitro. HBVseq makes it possible for those performing HBV RT sequencing to identify previous reports of the RT mutations in their sequence.

Acknowledgements

SYR, TFL and RWS were supported by a NIH grant, AI06858. SMT was supported by Hoffmann La Roche.

Contribution. SYR and RWS designed the study and wrote the manuscript. SYR and TFL performed the computer programming required for the database and sequence analysis program. MHN, RMK, BB, JV, and RK collected, annotated, and contributed to the study. SMT and RWS reviewed and annotated the studies in the HBVrtDB. All authors reviewed the manuscript and approved its content.

References

Bartholomeusz, A., Tehan, B.G., Chalmers, D.K., 2004. Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. Antivir. Ther. 9, 149–160.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Das, K., Xiong, X., Yang, H., Westland, C.E., Gibbs, C.S., Sarafianos, S.G., Arnold, E., 2001. Molecular modeling and biochemical characterization reveal the mechanism of hepatitis B virus polymerase resistance to lamivudine (3TC) and emtricitabine (FTC). J. Virol. 75, 4771–4779.

Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M., Green, J., 2007. HepSEQ: international public health repository for hepatitis B. Nucleic Acids Res. 35, D367–D370.

Günther, S., Fischer, L., Pult, I., Sterneck, M., Will, H., 1999. Naturally occurring variants of hepatitis B virus. Adv. Virus Res. 52, 25–137.

Keeffe, E.B., Dieterich, D.T., Han, S.H., Jacobson, I.M., Martin, P., Schiff, E.R., Tobias, H., 2008a. A treatment algorithm for the management of chronic hepatitis B virus infection in the United States: 2008 update. Clin. Gastroenterol. Hepatol. 6, 1315–1341.

Keeffe, E.B., Dieterich, D.T., Pawlotsky, J.M., Benhamou, Y., 2008b. Chronic hepatitis B: preventing, detecting, and managing viral resistance. Clin. Gastroenterol. Hepatol. 6, 268–274.

Kurbanov, F., Tanaka, Y., Mizokami, M., 2010. Geographical and genetic diversity of the human hepatitis B virus. Hepatol. Res. 40, 14–30.

Liu, B.M., Li, T., Xu, J., Li, X.G., Dong, J.P., Yan, P., Yang, J.X., Yan, L., Gao, Z.Y., Li, W.P., Sun, X.W., Wang, Y.H., Jiao, X.J., Hou, C.S., Zhuang, H., 2010. Characterization of potential antiviral resistance mutations in hepatitis B virus reverse transcriptase sequences in treatment-naive Chinese patients. Antiviral Res. 85, 512–519.

Lok, A.S., McMahon, B.J., 2009. Chronic hepatitis B: update 2009. Hepatology 50, 661–662.

Lok, A.S., Zoulim, F., Locarnini, S., Bartholomeusz, A., Ghany, M.G., Pawlotsky, J.M., Liaw, Y.F., Mizokami, M., Kuiken, C., 2007. Antiviral drug-resistant HBV: standardization of nomenclature and assays and recommendations for management. Hepatology 46, 254–265.

Ogata, N., Fujii, K., Takigawa, S., Nomoto, M., Ichida, T., Asakura, H., 1999. Novel patterns of amino acid mutations in the hepatitis B virus polymerase in association with resistance to lamivudine therapy in japanese patients with chronic hepatitis B. J. Med. Virol. 59, 270–276.

Schaefer, S., Magnius, L., Norder, H., 2009. Under construction: classification of hepatitis B virus genotypes and subgenotypes. Intervirology 52, 323–325.

Schildgen, O., Sirma, H., Funk, A., Olotu, C., Wend, U.C., Hartmann, H., Helm, M., Rockstroh, J.K., Willems, W.R., Will, H., Gerlich, W.H., 2006. Variant of hepatitis B virus with primary resistance to adefovir. N. Engl. J. Med. 354, 1807–1812.

Shafer, R.W., Jung, D.R., Betts, B.J., 2000. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. Nat. Med. 6, 1290–1292.

Shaw, T., Bartholomeusz, A., Locarnini, S., 2006. HBV drug resistance: mechanisms, detection and interpretation. J. Hepatol. 44, 593–606.

Sheldon, J., Camino, N., Rodes, B., Bartholomeusz, A., Kuiper, M., Tacke, F., Nunez, M., Mauss, S., Lutz, T., Klausen, G., Locarnini, S., Soriano, V., 2005. Selection of hepatitis B virus polymerase mutations in HIV-coinfected patients treated with tenofovir. Antivir. Ther. 10, 727–734.

Shin, I.T., Tanaka, Y., Tateno, Y., Mizokami, M., 2008. Development and public release of a comprehensive hepatitis virus database. Hepatol. Res. 38, 234–243.

Yuen, L.K., Ayres, A., Littlejohn, M., Colledge, D., Edgely, A., Maskill, W.J., Locarnini, S.A., Bartholomeusz, A., 2007. SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. Antiviral Res. 75, 64–74.