D566–D570 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 3 November 2012*
doi:10.1093/nar/gks1022

HBVdb: a knowledge database for Hepatitis B Virus

Juliette Hayer[1], Fanny Jadeau[1], Gilbert Deléage[1], Alan Kay[2], Fabien Zoulim[2,3] and Christophe Combet[1,*]

[1]Unité Bases Moléculaires et Structurales des Systèmes Infectieux; UMR 5086 CNRS - Université Claude Bernard Lyon 1; IBCP FR 3302 - 7, passage du Vercors, 69367 Lyon CEDEX 07, [2]INSERM, U1052, Viral Hepatitis Research Laboratory; Université Lyon 1, 151, cours Albert Thomas, 69003 Lyon, and [3]Hospices Civils de Lyon, Hepatology Department, 69004 Lyon, France

Received August 14, 2012; Revised September 28, 2012; Accepted October 3, 2012

ABSTRACT

We have developed a specialized database, HBVdb (http://hbvdb.ibcp.fr), allowing the researchers to investigate the genetic variability of Hepatitis B Virus (HBV) and viral resistance to treatment. HBV is a major health problem worldwide with more than 350 million individuals being chronically infected. HBV is an enveloped DNA virus that replicates by reverse transcription of an RNA intermediate. HBV genome is optimized, being circular and encoding four overlapping reading frames. Indeed, each nucleotide of the genome takes part in the coding of at least one protein. However, HBV shows some genome variability leading to at least eight different genotypes and recombinant forms. The main drugs used to treat infected patients are nucleos(t)ides analogs (reverse transcriptase inhibitors). Unfortunately, HBV mutants resistant to these drugs may be selected and be responsible for treatment failure. HBVdb contains a collection of computer-annotated sequences based on manually annotated reference genomes. The database can be accessed through a web interface that allows static and dynamic queries and offers integrated generic sequence analysis tools and specialized analysis tools (e.g. annotation, genotyping, drug resistance profiling).

INTRODUCTION

Hepatitis B virus (HBV) is a major health problem worldwide with more than 350 million people being chronic carriers. Chronic HBV infection is associated with a significantly increased risk of developing severe liver diseases, including liver cirrhosis, and hepatocellular carcinoma (HCC), one of the most common forms of human cancer. The estimated risk of HCC in chronic HBV carriers is ~100 times greater than in uninfected individuals (1). Currently available anti-HBV drugs have limitations. Indeed, interferon alpha administration is associated with adverse reactions, while nucleos(t)ide analogs are virostatic and require long-term administration (2,3).

HBV is an enveloped DNA virus that belongs to the *Hepadnaviridae*, a family of hepatotropic DNA viruses infecting certain mammalian or avian hosts (4). It contains a small (~3.2 kb), partially double-stranded relaxed-circular DNA (rcDNA) genome that replicates by reverse transcription of an RNA intermediate, the pregenomic RNA (pgRNA). The genome encodes four overlapping open reading frames (ORFs) that are translated to produce the viral core protein (5,6), the surface proteins (5,7), a polymerase/reverse transcriptase (RT) (2,4), and HBx (8,9). The HBV life cycle starts with the binding of the virus to an unknown receptor of the host cell. Then, the viral particle is internalized. The virion rcDNA is delivered to the nucleus, where it is repaired to form a covalently closed-circular cccDNA. The episomal cccDNA serves as the template for the transcription of the pgRNA and the other viral mRNAs by the host RNA polymerase II (10).

The viral genome is variable because of the spontaneous error rate of the viral polymerase and the lack of proof reading activity. There are eight genotypes of HBV designated A–H based on >8% nt variation over the entire genome. The eight HBV genotypes are distributed in distinct geographical localizations (11–13). Recombinant forms involving different genotypes have also been reported (14). However, the extensive overlap between the four encoded ORFs limits the diversity that the virus can tolerate. Indeed, every nucleotide participates in the coding of at least one viral protein attesting of an optimized small genome (15,16). Moreover, the genome variability is also constrained by environmental pressure exerted by the host immune response and the antiviral drugs for treated patients.

To allow researchers to investigate the genetic variability of HBV sequences and viral resistance to treatment, several databases and repositories (17–19) have been published to date. Moreover, tools specific to HBV genotyping (20–22) are available for virologists, as well

*To whom correspondence should be addressed. Tel: +33 4 37 65 29 47; Fax: +33 4 72 72 26 04; Email: christophe.combet@ibcp.fr

as tools aimed at drug resistance mutations analysis, among which some are freely accessible (23) and others need a registration (24). We have developed HBVdb, a database that contains a collection of computer-annotated HBV sequences thanks to manually annotated reference genomes. The sequences taken as input are the ones publicly available in the INSDC (25), including partial and complete HBV genomes. The database can be queried via a web interface and the query results can be further analysed with the numerous integrated generic and specialized (e.g. annotation, genotyping, drug resistance profiling) analysis tools.

DATABASE BUILDING

We developed a fully automated procedure to annotate all HBV sequences from the European Nucleotide Archive (ENA) (26), using a reference set of 16 manually annotated and non-recombinant complete genomes representing the 8 genotypes. The HBVdb building process starts with the retrieval of all the HBV entries in ENA. The second step is the automatic annotation of these entries in their text format (flat file ENA format). The annotated HBVdb entries are then loaded into a PostgreSQL relational database system. Finally, sequence datasets are extracted and multiple sequence alignments together with associated data are computed.

The HBVdb is updated on a monthly basis. The program for the automatic annotation, as well as for the querying and the management of the database, are implemented in Java and SQL programming languages.

ANNOTATION PROCEDURE

A standard numbering system of HBV genomes exists, defined by the (often hypothetical) EcoR1 restriction site as the origin of the genome (27). However, the circularity of the HBV genome leads to the deposit of sequences in generic public databases that do not follow this system. Such sequences will result in one or several partial pairwise alignments with the reference genomes. To circumvent this problem, we modeled the HBV by duplicating the sequence of each reference genome (from ~3.2 kb to ~6.4 kb). The first step of the automated annotation procedure is a similarity search, using the FASTA program (28) (Figure 1A) in order to identify the most similar reference genome to the sequence to annotate. The second step of the annotation procedure checks if the query sequence follows the EcoR1 numbering by looking if the query sequence is aligned to only one replicate or if it overlaps both. In the latter case, the part of the query sequence that is aligned on the second replicate is shifted to the corresponding region of the first replicate. If there are gaps between the shifted part and the fixed part, they are replaced by 'n'. This checking ensures that all the sequences follow the EcoR1 numbering system. The third step consists in optimizing the global query-reference pairwise alignment in order to avoid erroneous translation or to detect non-functional sequences. For each coding sequence (CDS) found in the query, a pairwise alignment is produced from the global one and divided into non-overlapping 3-nt windows (i.e. codons). If a window contains one or two gaps, the process tries to optimize gaps in order to have only 3 nt in the window or three gaps (codon deletion). If the optimization fails, the entry is discarded from the database. In the fourth step, the reference genome features (e.g. *CDS*, *mat_peptide*) are mapped onto the sequence to annotate when they are present. The sequence is genotyped in a fifth step. The last step corresponds to the drug resistance profiling. Finally, the annotated HBV entry is formatted as an ENA text entry for its later inclusion in the relational database.

GENOTYPING

Starting from all the pairwise query-reference alignments (Figure 1B), the algorithm computes a matrix containing the identity percentage at each position of the query for each genotype. The identity percentage is computed by summing the identity percentage over overlapping sliding windows (window length 301 nt, window step 1 nt) divided by the number of windows used at each query position. The maximum identity percentage calculated for each query position is taken from the matrix to fill an array with the corresponding genotype letter, as long as the maximum mean identity percentage is above or equal to 90%. The genotype is then computed from this array. The procedure is able to process only sequences with lengths equal or above the window size. Overall, the genotyping algorithm allows the identification of 'pure' genotype sequence as well as recombinant ones (14,29). In the result file, users can find the number of informative positions used in the genotype computation. This value can be used as a confidence value of the genotype prediction. The accuracy of our genotyping tool is similar to Oxford (20,30), jpHMM (21) and NCBI (22) HBV genotyping tools, including recombinant genomes detection. The HepSEQ genotyping tool (17) uses only polymerase/surface genes in genotype computation and is less accurate in recombinant detection (Supplementary Tables S1 and S2).

DRUG RESISTANCE PROFILING

An alignment between the query protein sequence and a reference reverse transcriptase (RT) sequence is computed. The algorithm searches, in the query sequence, for mutations defining known resistance profiles to lamivudine, telbivudine, entecavir, adefovir and tenofovir drugs (31). If all the mutations of one profile are found, the profile is reported with the associated drug and resistance status. There are three possible resistance statuses designated by 'Sensitive' (S), 'Intermediate' (I, reduced susceptibility) and 'Resistant' (R). The algorithm output provides the detected profile, the status and mutation positions in the query sequence and according to the RT numbering system (32). In its current implementation, our resistance tool does not look for antiviral drug-associated potential vaccine-escape mutants [ADAPVEM (33)] contrarily to the HepSeq polymerase annotator (17).

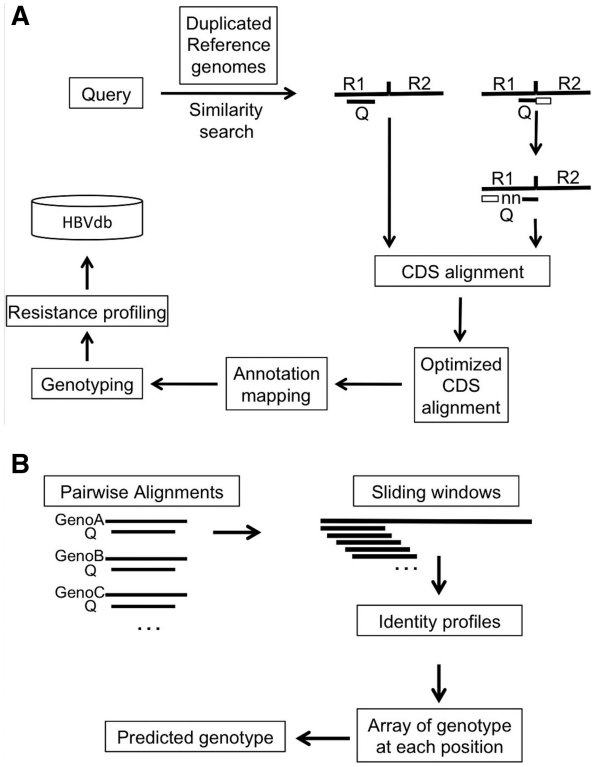

Figure 1. Annotation and genotyping processes. (**A**) The annotation procedure starts with the computation of a pairwise alignment between the query sequence (Q) and the most similar duplicated reference genome sequence (R1, R2: replicate 1, 2; nn indicates missing nucleotides between shifted and fixed parts of the query sequence). This alignment is split up into CDS alignments that are optimized before the mapping of features and the transfer of annotations. (**B**) Genotyping process. These pairwise alignments between query (Q) and reference sequences (e.g. GenoA, genotype A reference genome) are iterated using sliding windows to compute a matrix of mean identities. The matrix is used to produce the array of genotype at each query sequence position. Predicted genotype is deduced from this array.

DATABASE CONTENT

The text format of an HBVdb entry is an extension of the ENA-Annotation format as used for the euHCVdb (34). Some elements of the ENA-Annotation entry are conserved such as the accession number, the organism name, the creation date, and the references. After the annotation procedure, some elements are corrected and/or completed in the entry, mainly in the features. Indeed, a set of new qualifiers that store specific data is added to some features. The qualifier *PRABI_genotype* is added to the *source* feature to indicate a provisional genotype predicted by the genotyping tool. The qualifier *PRABI_name* is added to each feature to ensure standard names across all the database entries. Concerning the protein annotations, some qualifiers are added to the *mat_peptide* features. These qualifiers, noted *PRABI_prodft*, follow the feature table format of the UniProtKB database (35). The *PRABI_prodft* qualifiers describe the protein chains, the protein domains, some sites like active sites (*act_site*) that designate the catalytic residues of enzymes, and the resistance mutations (*res_mut*) with the drug and the resistance status.

WEB INTERFACE

The HBVdb is accessible through a website (Figure 2A) divided into two parts. In the static part, the user can find general information about HBV and the nomenclature used through genome organization, protein descriptions (Figure 2B), the reference genomes and the genotypes. The user can also access pre-computed 'Nucleotide' and 'Protein' datasets sorted according to the genotype (rows) and the protein or CDS names (columns). The datasets provide full-length sequences in Pearson/Fasta format, their multiple sequence alignments computed with Muscle (36) and displayed in Clustal W format (Figure 2C), and the corresponding residue repertoires with Shannon entropies that are useful for analysing conserved/variable alignment positions. The user can download the corresponding files for further analysis. Furthermore, the alignments can be interactively edited with the 'EditAlignment' applet developed by our team. In the dynamic part, users can extract their own dataset by combining multiple criteria (e.g. genotype and sequence length and protein name). The datasets can be exported as Pearson/Fasta sequences, accession number lists and entry flat files for further analysis with the integrated analysis tools.

The available analysis tools are either generic or specialized. The generic analysis tools (e.g. BLAST (37) or Clustal W (38)) are available through the NPS@ server (39), that is an integrated sequence analysis web server. 'Annotate' (Figure 2D), 'Genotype' and 'Resistance' (Figure 2E) specialized tools allow the analysis of one or several HBV sequences uploaded by the user. The annotation of a nucleotide sequence produces a result page listing the CDS found in the query sequence, presenting the predicted genotype, and giving the drug resistance status (with a link to resistance output file) if the nucleotide sequence contains the CDS of the RT domain. The results page also gives access to the global pairwise alignment between the query sequence and the reference genome, as well as the pairwise alignments for each CDS. The user can also access the text entry format of the annotated sequence. The annotation of a protein sequence ends up with a result page indicating the most similar sequence used for its annotation, with links to the entry and the pairwise alignment, and the resistance status if the sequence contains the RT domain. The 'Genotype' tool allows the user to genotype nucleotide sequences. It produces a result page giving the predicted genotype, and the genotype computed at each position of the query sequence. The 'Resistance' tool enables the detection of known drug resistance mutations from nucleotide or protein sequences. The output lists the drug, the resistance status (R, I, S) or 'n.a.' if the query sequence does not contain the HBV RT domain, and the identified mutations.

STATISTICS

HBVdb is available since June 2012. The release 4 (September 2012) comprises 39 289 sequences, including 3606 complete genomes.

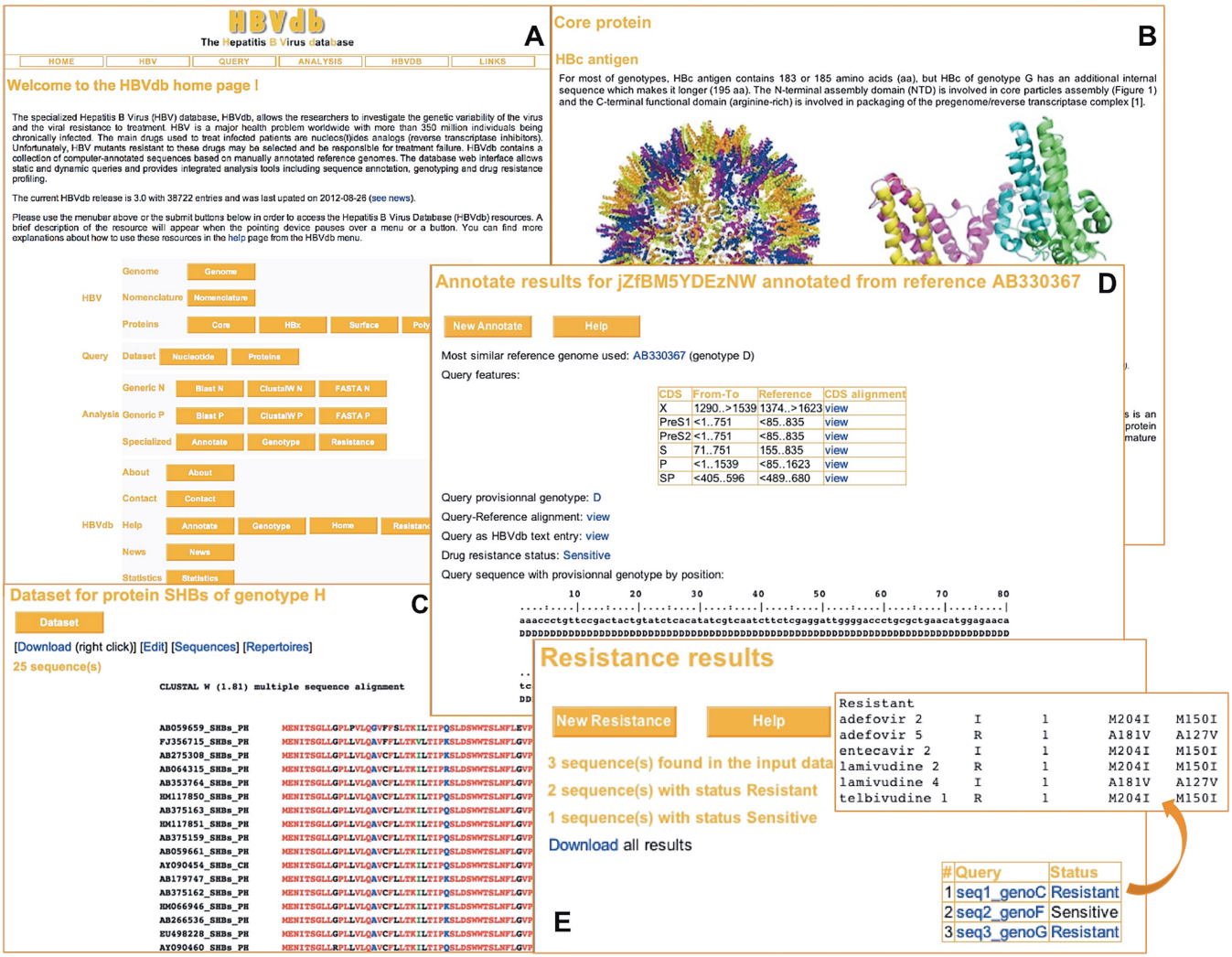

Figure 2. Snapshot of the HBVdb web interface. (**A**) The HBVdb homepage, with the menu bar and all the menus repeated as buttons for ease-of-use. (**B**) The HBV core protein information page. (**C**) The multiple sequence alignment of the genotype H SHBs proteins. (**D**) The *Annotate* tool detailed result page for an input nucleotide sequence showing the CDS positions with links to their alignment, the genotype and the resistance status. (**E**) The *Resistance* tool result page showing the resistance status and the resistance profile.

CONCLUSIONS AND PERSPECTIVES

HBVdb is a collection of human HBV sequences, annotated thanks to a computer-automated system. The automatic annotation process used to generate HBVdb entries guarantees standard annotations and allows efficient keyword searches. The HBVdb website allows researchers to access pre-computed datasets and to extract data thanks to a dynamic query system. The extracted data can be further analysed with a set of bioinformatics tools available in the NPS@ server. HBVdb website also enables users to annotate their own sequences through the HBVdb automatic annotation process. In the future, the annotation will be enriched with new data (e.g. mapping of promoters) and analysis tools (e.g. ADAPVEM detection) will be added. The HBVdb and related programs will be useful to face the analysis of the flood of sequences that are generated thanks to the Next-Generation Sequencing (NGS) technologies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2.

ACKNOWLEDGEMENTS

We acknowledge members of the French ANRS hepatitis network for stimulating discussions through the last years in order to make HBVdb a useful tool for the researchers.

FUNDING

J.H. is recipient of a doctoral fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche. HBVdb is hosted on the Pôle Rhône-Alpes de BioInformatique platform funded by the Groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie (GIS IBiSA). Funding for open access

charge: Groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie (GIS IBiSA).

Conflict of interest statement. None declared.

REFERENCES

1. Nguyen,D.H., Ludgate,L. and Hu,J. (2008) Hepatitis B virus-cell interactions and pathogenesis. *J. Cell. Physiol.*, **216**, 289–294.
2. Zoulim,F. and Locarnini,S. (2009) Hepatitis B virus resistance to nucleos(t)ide analogues. *Gastroenterology*, **137**, 1593–608.e1-2.
3. European Association For The Study Of The Liver. (2009) EASL Clinical Practice Guidelines: management of chronic hepatitis B. *J. Hepatol.*, **50**, 227–242.
4. Nassal,M. (2008) Hepatitis B viruses: reverse transcription a different way. *Virus Res.*, **134**, 235–249.
5. Bruss,V. (2004) Envelopment of the Hepatitis B Virus nucleocapsid. *Virus Res.*, **106**, 199–209.
6. Wynne,S.A., Crowther,R.A. and Leslie,A.G.W. (1999) The crystal structure of the human Hepatitis B Virus capsid. *Mol. Cell*, **3**, 771–780.
7. Glebe,D. and Urban,S. (2007) Viral and cellular determinants involved in hepadnaviral entry. *World J. Gastroenterol.*, **13**, 22–38.
8. Benhenda,S., Cougot,D., Buendia,M.-A. and Neuveut,C. (2009) Chapter 4 Hepatitis B Virus X Protein: molecular functions and its role in virus life cycle and pathogenesis. In: Woude,G.F.V. and Klein,G. (eds), *Advances in Cancer Research*. Academic Press, Vol. 103, pp. 75–109.
9. Bouchard,M.J. and Schneider,R.J. (2004) The enigmatic X gene of hepatitis B virus. *J. Virol.*, **78**, 12725–12734.
10. Beck,J. and Nassal,M. (2007) Hepatitis B virus replication. *World J Gastroenterol*, **13**, 48–64.
11. Bartholomeusz,A., Tehan,B.G. and Chalmers,D.K. (2004) Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. *Antivir. Ther.*, **9**, 149–160.
12. Schaefer,S. (2007) Hepatitis B Virus taxonomy and Hepatitis B Virus genotypes. *World J. Gastroenterol.*, **13**, 14–21.
13. Norder,H., Couroucé,A.M., Coursaget,P., Echevarria,J.M., Lee,S.D., Mushahwar,I.K., Robertson,B.H., Locarnini,S. and Magnius,L.O. (2004) Genetic diversity of Hepatitis B Virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology*, **47**, 289–309.
14. Simmonds,P. and Midgley,S. (2005) Recombination in the genesis and evolution of Hepatitis B Virus genotypes. *J. Virol.*, **79**, 15467–15476.
15. Kay,A. and Zoulim,F. (2007) Hepatitis B Virus genetic variability and evolution. *Virus Res.*, **127**, 164–176.
16. Araujo,N.M., Waizbort,R. and Kay,A. (2011) Hepatitis B Virus infection from an evolutionary point of view: how viral, host, and environmental factors shape genotypes and subgenotypes. *Infect. Genet. Evol.*, **11**, 1199–1207.
17. Gnaneshan,S., Ijaz,S., Moran,J., Ramsay,M. and Green,J. (2007) HepSEQ: International Public Health Repository for hepatitis B. *Nucleic Acids Res.*, **35**, D367–D370.
18. Shin-I,T., Tanaka,Y., Tateno,Y. and Mizokami,M. (2008) Development and public release of a comprehensive hepatitis virus database. *Hepatol. Res.*, **38**, 234–243.
19. Panjaworayan,N., Roessner,S.K., Firth,A.E. and Brown,C.M. (2007) HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in Hepatitis B Virus sequences. *Virol. J.*, **4**, 136.
20. Alcantara,L.C., Cassol,S., Libin,P., Deforche,K., Pybus,O.G., Van Ranst,M., Galvão-Castro,B., Vandamme,A.M. and de Oliveira,T. (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res.*, **37**, W634–W642.
21. Schultz,A.K., Bulla,I., Abdou-Chekaraou,M., Gordien,E., Morgenstern,B., Zoaulim,F., Dény,P. and Stanke,M. (2012) jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic Acids Res.*, **40**, W193–W198.
22. Rozanov,M., Plikat,U., Chappey,C., Kochergin,A. and Tatusova,T. (2004) A web-based genotyping resource for viral sequences. *Nucleic Acids Res.*, **32**, W654–W659.
23. Rhee,S.Y., Margeridon-Thermet,S., Nguyen,M.H., Liu,T.F., Kagan,R.M., Beggel,B., Verheyen,J., Kaiser,R. and Shafer,R.W. (2010) Hepatitis B Virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. *Antiviral Res.*, **88**, 269–275.
24. Yuen,L.K., Ayres,A., Littlejohn,M., Colledge,D., Edgely,A., Maskill,W.J., Locarnini,S.A. and Bartholomeusz,A. (2007) SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. *Antiviral Res.*, **75**, 64–74.
25. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. (2012). International Nucleotide Sequence Database Collaboration. (2012) The International Nucleotide Sequence Database collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
26. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
27. Ono,Y., Onda,H., Sasada,R., Igarashi,K., Sugino,Y. and Nishioka,K. (1983) The complete nucleotide sequences of the cloned Hepatitis B Virus DNA; subtype adr and adw. *Nucleic Acids Res.*, **11**, 1747–1757.
28. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
29. Kramvis,A., Arakawa,K., Yu,M.C., Nogueira,R., Stram,D.O. and Kew,M.C. (2008) Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. *J. Med. Virol.*, **80**, 27–46.
30. de Oliveira,T., Deforche,K., Cassol,S., Salminen,M., Paraskevis,D., Seebregts,C., Snoeck,J., van Rensburg,E.J., Wensing,A.M., Vijver,D.A. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.
31. European Association for the Study of the Liver. (2012) EASL Clinical Practice Guidelines: management of chronic Hepatitis B Virus infection. *J. Hepatol.*, **57**, 167–185.
32. Stuyver,L.J., Locarnini,S.A., Lok,A., Richman,D.D., Carman,W.F., Dienstag,J.L. and Schinazi,R.F. (2001) Nomenclature for antiviral-resistant human Hepatitis B Virus mutations in the polymerase region. *Hepatology*, **33**, 751–757.
33. Locarnini,S.A. and Yuen,L. (2010) Molecular genesis of drug-resistant and vaccine-escape HBV mutants. *Antivir. Ther.*, **15**, 451–461.
34. Combet,C., Garnier,N., Charavay,C., Grando,D., Crisan,D., Lopez,J., Dehne-Garcia,A., Geourjon,C., Bettler,E., Hulo,C. *et al.* (2007) euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res.*, **35**, D363–D366.
35. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
36. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
37. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
38. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
39. Combet,C., Blanchet,C., Geourjon,C. and Deleage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147.