Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR

Richard Myers,1 Caroline Clark,1 Arshad Khan,1 Paul Kellam1 and Richard Tedder2

Hepatitis B virus (HBV) genomes have been classified into eight genotypes based on phylogenetic analysis of sequence variation. Identifying and tracking the movement of HBV genotypes is important in terms of both monitoring infection rates and predicting disease and treatment. An HBV genotyping tool has been developed that compares query sequences with position-specific scoring matrices representing the eight HBV genotypes. This tool (HBV STAR) is rapid, robust and accurate and assigns genotype based on a statistically defined scoring model. HBV STAR confidently assigned 90% of 590 full-length HBV genomes to an HBV genotype (Z score > 2). Thirty-two of the residual 48 sequences were identified as non-human primate viruses and 16 sequences were identified as recombinant or putative recombinants. Receiver-Operated Characteristic (ROC) analysis was used to compare the accuracy of genotype prediction using basal core promoter sequences and surface and core genes with the accuracy achieved by using full-length sequences. A web interface to HBV STAR is available at http://www.vgb.ucl.ac.uk/starn.shtml.

INTRODUCTION

Hepatitis B virus (HBV), a 3-2 kb orthohepadnavirus, is a major human pathogen. The outcome of acute infection by HBV is variable, but is usually followed by a complete recovery. A small proportion of infections continue as a chronically infected carrier state, in which the virus persists in the liver. There are usually no initial symptoms of carriage, but, over time, cirrhosis and hepatocellular carcinoma (HCC), the major disease sequelae of carriage, may arise. HBV exists in its human host as eight clusters of viruses, each cluster displaying a similarity of sequences (genotypes A–H) and variable antigenicity (serotypes). It is estimated that some 350 million persons globally are currently infected as carriers, whilst many more will have been infected with HBV and recovered.

A number of host factors are recognized that affect the outcome of the infection, e.g. infection in early life or in the face of immunosuppression favours carriage, although less is known about virological factors that might also influence the outcome of both acute and chronic infection. There is increasing evidence that different HBV genotypes may be associated with alternative disease profiles and differing responses to antiviral therapy. Studies in Taiwan and Japan have suggested that, in patients with cirrhosis, genotype C is more common, whereas HCC is associated with genotype B in patients younger than 50 years old and with genotype C in those over 50 (Ding et al., 2001; Fujie et al., 2001; Kao et al., 2000). HBV genotype also appears to influence the viral response in carriers undergoing seroconversion from having HBV e antigen (HBeAg) in their serum to having antibody to HBeAg (anti-HBe) (Delaney et al., 2001; Lok et al., 1994).

Sequence variation within the HBV genome has been classified into the eight genotypes A–H, which have been defined by differences in their full-length genome of > 8% (Norder et al., 2004). These are distinct from serological subtypes, which are defined by the antigenicity of the HBV surface antigen (HBsAg) (Cournoué et al., 1983), determined by amino acids at specific residues in the ‘a’ determinant of HBsAg. Although improvements in molecular biology, computational power and phylogenetic algorithms have facilitated characterization and genotyping of the full-length HBV genome, many genotype predictions are often determined more practically by sequencing a smaller region of the genome, usually the surface antigen.

Antiviral therapy targeting the reverse transcription function of polymerase is increasingly used in clinical practice to
suppress virus replication in carriers. Not surprisingly, long-term use of single drugs in the face of continued replication is associated with viral mutational escape from the drug. This can be monitored by direct sequencing of polymerase (pol) (Lai et al., 2003; Tenney et al., 2004), which overlaps entirely a different open reading frame for the HBsAg (s) gene. Thus, sequencing of pol will also generate s data and provides an increasingly large sequence repository that could be used for genotyping, potentially providing both epidemiological data and further insight into host–parasite relationships.

There are currently few methods available for genotyping HBV. It is possible to analyse phylogenetic trees including novel sequences and a set of reference sequences, an approach that requires skill to interpret and is not reliable for outlying or recombinant sequences. There is also one web-based genotyping tool from the National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi), which uses the BLAST algorithm along a sliding window to compare HBV nucleotide sequences with reference subtype sequences. It does not, however, attempt to assign any confidence to the nucleotide sequences with reference subtype sequences. It therefore does not attempt to assign any confidence to the genotype sequences used to define the eight genotype PSSMs. False Positives (FP) and True Negatives (TN) were assigned on the basis of whether or not the query sequence was correctly classified as (or not) in the most likely genotype.

To genotype HBV accurately and robustly, our HBV STAR program was populated with sequence alignments representing all HBV genotypes. Neighbour-joining phylogenetic analysis of 590 full-length HBV genomes, including 23 genotype reference sequences (Fig. 1), identified GenBank sequences (of unknown genotype) that clustered with these reference sequences. Bootstrapped neighbour-joining trees of individual genotypes and recombination analysis enabled the unambiguous genotyping of 441 of 567 query sequences, giving a population of 464 phylogenetically defined full-length sequences. Genotypes A, B, C and D were well-populated (Table 1), reflecting the extent to which these genotypes have been studied in both Europe and Asia. Tight clusters of sequences were formed around reference sequences from genotypes A, B and D; however, the genotype C cluster was greatly extended. Only sequences that clustered in close proximity to the NCBI reference sequences were used to populate the genotype C dataset. Sequences clustering with reference sequences from genotypes E, F, G and H formed distinct and discrete clusters; however, the number of sequence representatives in each genotype was lower.

RESULTS

Genotyping analysis was performed by using HBV STAR. This method converts genotype-specific alignments into position-specific scoring matrices (PSSMs) for each genotype and then compares a query sequence of unknown genotype to each PSSM, as described previously for HIV (Myers et al., 2005). The scores that are generated by this comparison (eight scores in the case of HBV) are transformed into Z scores, giving the eight-point distribution a mean of zero and a standard deviation of 1. The genotype PSSM generating the highest Z score and the magnitude of this Z score are used to predict the genotype of the query sequence and the confidence of that genotype prediction, respectively. Z scores >2.0 are considered indicative of a significant genotype prediction. The PSSM generating the highest raw score also generates the highest Z score; however, Z-score transformation has been developed to normalize HBV STAR scoring such that longer sequences do not generate arbitrarily higher scores. In cases where the query sequence is recombinant, two genotype PSSMs will produce high raw scores; however, the resulting maximal Z score will be lowered. Query sequences that generate low Z-score predictions (<2.0) are therefore recomputed with the genotype PSSMs to detect the presence of putative HBV recombination in a two-stage process. The query sequence is genotyped as defined previously and then a separate process analysing the sequences for recombination is performed. Recombination detection was conducted by using the difference in sequence identity relative to the ascribed genotype along a sliding window of 150 bp with a step interval of one base. Sequences containing a segment in excess of 150 bp where the mean sequence identity was more similar to a different genotype and that diverged from the ascribed genotype by >1% were considered as potential recombinants. This procedure can be initiated automatically on the basis of a user-definable minimum Z-score threshold. Regions of the query sequence that are more similar to a PSSM that is not the predicted genotype are identified by accumulating the difference between the normalized positional nucleotide frequencies.

Receiver-Operated Characteristic (ROC) analysis was used to assess and compare the performance of HBV STAR by determining the sensitivity and specificity of genotype predictions over a range of Z-score thresholds. At any Z-score value within the range examined, True Positives (TP) were those with correctly assigned genotype with Z scores above the given threshold, and False Negatives (FN) were those falling below a Z-score threshold. This was performed for 464 HBV genotype sequences used to define the eight genotype PSSMs. False Positives (FP) and True Negatives (TN) were assigned on the basis of genotype prediction for a set of non-human HBV sequences. FP scoring above and TN below a Z-score threshold, respectively. Sensitivity was calculated as TP/(TP+FN) and specificity was calculated as TN/(TN+FP).

METHODS

Five hundred and ninety available full-length HBV genomes were downloaded from GenBank (available from the authors on request). The accession numbers of 23 genotype reference sequences (see Supplementary Table S1, available in JGV Online) were obtained from the NCBI and identified within the dataset of 590 sequences. Alignment of HBV sequences was performed by using CLUSTAL_X (Thompson et al., 1997), followed by manual curation of the resulting alignment. Neighbour-joining phylogenies and bootstrapping were used to genotype GenBank sequences relative to genotype reference sequences and were performed by using the PHYLIP package (Felsenstein, 1996). Recombination analysis was performed by using HBV STAR and SimPlot (Lole et al., 1999).

Downloaded from www.microbiologyresearch.org by IP: 54.70.40.11 On: Mon, 07 Jan 2019 03:35:06
Eight PSSMs, each representing a single HBV genotype, were formed (Table 1) and the 464 sequences used to populate these PSSMs were genotyped by using HBV STAR (Fig. 2). A resampling method was used to assess genotype-prediction accuracy where the 464 sequences were each removed individually from their respective genotype PSSM prior to use as a query sequence in order to avoid biasing this genotype prediction. By using full-length HBV genome PSSMs, 458 (98.7%) sequences were reclassified correctly with a Z score > 2.0. Six (1.3%) of the 464 sequences were assigned the correct genotype, but with a Z score < 2.0. These six sequences (genotypes B = 4, C = 1, D = 1) were predicted to contain putative genotype recombinations, although the size of the recombinant fragment ranged from the limit of detection of 150 bp to approximately 1500 bp. The remaining 126 sequences from the initial dataset of 590 sequences were also genotyped by using HBV STAR (Fig. 2). Seventy-eight (62%) of these sequences were predicted confidently (Z score > 2.0) as belonging to one of the eight HBV genotypes (A = 13, B = 6, C = 54, D = 6), suggesting that our phylogenetic assignment of genotype was more rigorous than necessary. Forty-eight (38%) sequences that scored below the Z-score threshold of 2.0 were not assigned a confident genotype prediction. The most prominent grouping amongst these 48 sequences that were not assigned a genotype prediction were 32 HBV genomes that were similar to genotype E. Analysis of the GenBank annotation associated with these sequences showed them to be HBV genomes derived from non-human primate sources. These non-human HBV-like genomes were independent from the HBV genotypes by using phylogenetic analysis and clustered together (Fig. 1). They were also predicted to be closest to genotype E by using HBV STAR, but with Z scores below 2.0 (Fig. 2). The remaining 16 sequences were all identified by HBV STAR as containing putative HBV genotype recombination (AB = 1, AC = 3, AD = 3, BC = 3, CD = 2, CE = 4). An example of recombination detection within HBV STAR is illustrated by the analysis (Fig. 3) of a low-scoring query sequence (gi16751309), initially predicted to be genotype C with a low Z score, which was reclassified correctly as a genotype C/D recombinant.

Table 1. HBV STAR genotyping of 464 full-length, surface, core and BCP sequences

| Genotype | n  | Full-length |  |  |  |  |  |  |  |  |  |  |  |  |
|----------|----|-------------|  |  |  |  |  |  |  |  |  |  |  |  |
|          | FP | FN†         |  |  |  |  |  |  |  |  |  |  |  |  |
| A        | 105 | 0 | 0 | 5 | 0 | 7 | 1 | 0 | 4 |
| B        | 120 | 0 | 4 | 0 | 12 | 2 | 24 | 0 | 0 |
| C        | 143 | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 0 |
| D        | 72  | 0 | 1 | 3 | 14 | 0 | 1 | 0 | 2 |
| E        | 6   | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| F        | 3   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G        | 10  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| H        | 5   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total    | 464 | 0 | 6 | 8 | 31 | 9 | 27 | 0 | 6 |

*FP, False Positive (incorrect genotype prediction, Z score > 2.0).†FN, False Negative (genotype prediction, Z score < 2.0).
ROC analysis of full-length HBV genomes (Fig. 4) was performed by assessing sensitivity \[\frac{TP}{TP + FN}\] and specificity \[\frac{TN}{TN + FP}\] across a range of Z-score thresholds (−2 to +3). HBV STAR using full-length genomes performed very impressively, achieving sensitivity and specificity scores close to 1 between Z scores of −2 and +2.5. This illustrated that, within this range, there was no reduction in TN- and minimal reduction (1-0-0.97) in TP-detection rates. The definitive Z-score threshold of 2.0 was established on the basis of the ROC curve analysis.

HBV STAR was very effective at assigning accurate and robust genotypes to full-length HBV genomes; however, the majority of HBV sequences that are generated in clinical practice do not represent the entirety of the HBV genome. To ensure that HBV STAR could be used to analyse the subgenomic sequence databases, the sequence-alignment algorithm that matched the query sequence to a template of the genotype PSSMs was modified such that a local, rather than a global, alignment could be performed. This allowed short regions of query sequence to be aligned with an HBV template representing full-length HBV PSSMs without the need for independent PSSMs representing those subgenomic regions of sequence. To examine whether using shorter query sequences had any potential for increasing error in genotype prediction, ROC analysis was performed on HBV STAR by using subgenomic portions of the HBV genome comprising HBV surface, core and BCP regions (Fig. 5). ROC analysis showed that HBV STAR predicted the genotype of HBV surface sequences with the same high degree of sensitivity and specificity as for full-length HBV genomes. At a Z-score

![Graph](image-url)
threshold of 2·0 for both surface and full-length sequences, there were no instances where a surface query sequence was predicted to be of an incorrect genotype (no FP). The only error in the prediction of surface and full-length sequences occurred in six instances where sequences scored below the Z-score threshold of 2·0 (FN; Table 1). The identities of six full-length and the six surface gene sequences scoring below a Z score of 2·0 differed between the two genotyping methods, although, in both cases, these sequences failed because of a prediction of genotype recombination. This showed that only 1·3% (six out of 464) of sequences were not classified correctly by using either full-length or surface regions and that, in both cases, the misclassification was not prediction of the wrong genotype, but rather prediction of no genotype.

The ROC curve analysis of HBV STAR genotyping of BCP and core sequences highlighted a reduction in performance relative to full-length and surface sequence genotyping (Fig. 5). Between 6 and 7% of sequences (31 BCP and 27 core) gave an FN genotype result at a Z-score threshold of 2·0. Genotyping based on BCP and core regions also resulted in FP predictions (Table 1). Eight BCP sequences were predicted falsely (five as genotype A, three as genotype D), as were nine core sequences (two as genotype B, seven as genotype A).

### DISCUSSION

Phylogenetic analysis and partitioning of HBV sequences into sequence variation-derived genotypes is widely established (Norder et al., 2004). We used 464 genotyped sequences to establish HBV genotype-specific PSSMs. HBV STAR classification of sequences ascribed genotype correctly to 458 of the 464 sequences at a Z = 2·0 threshold. It also predicted genotypes in 78 of the group of 126 sequences that were not genotyped by phylogenetic analysis and, in the remaining 48 sequences, highlighted the presence of 32 non-human primate HBV sequences and 16 putative recombinant sequences. Overall, HBV STAR assigned accurate genotypes to 90% of all available full-length HBV sequences and identified reasons for not ascribing genotype.

Analysis of the performance of HBV STAR compared genotyping of HBV sequences of differing length. Clearly, reductions in the query sequence length could result in a loss of genotype-specific signal, thereby causing a greater likelihood of errors in the genotype prediction. Comparisons between genotyping sensitivity and specificity of subgenomic regions of HBV relative to full-length genomes showed that genotyping using the surface gene remained of high specificity. In some ways this was not surprising, as the surface/polymerase-encoding region of HBV has long been used for phylogenetic assignment of genotype. This indicates a strong relationship between sequence variation and genotype, even though variation is constrained by the overlap of surface and polymerase genes. The finding of a 1·3% FN-prediction rate when reclassifying the surface gene and full-length sequences within HBV STAR represented the only classification errors. Non-assignment of genotype to a query sequence is safer than incorrect assignment of genotype and was detected here on the basis of low Z scores. A proportion of the low-Z-scoring sequences (n = 16) showed evidence of recombination. HBV STAR recombination analysis of full-length sequences identified true recombinant sequences (gi16751309, gi15419825, gi10443814, gi10443806 and gi10443822) and other putative recombinant sequences. The fact that HBV STAR generated no FP results during analysis of full-length and surface HBV sequences shows that this tool performs well using these regions of HBV sequence.

In the absence of surface gene sequence, core and BCP sequences may give an indication of genotype. However, whilst BCP and core sequences were genotyped accurately 92·5% and 93·2% of the time, respectively, there was a slightly increased risk of FP prediction (1·7 and 1·9%). This risk increased when predictions of genotypes A and D were considered by using BCP sequences (4·5%) and when predictions of genotype A were considered by using core sequences (4%). Even though they contain more variation than the surface gene, the core and BCP regions of the HBV genome are difficult to use as predictors of HBV genotype. This arises because the sequence variation encoded within these regions does not appear to be solely genotype-specific. We surmise that the sequence variation within core and BCP of HBV may be a function of both genotype and interaction with the host. Seroconversion from HBeAg to anti-HBe in carriers is associated with changes in the core/precore-encoding region (Delaney et al., 2001; Lok et al., 1994). It seems entirely plausible that changes in the BCP corresponding to different liver transcription-factor binding may provide selection-driven sequence changes that confuse attempts to genotype based on this region. These findings suggest that prediction of genotype using HBV core and BCP sequences would be more error-prone than that using full-length sequences and surface-encoding regions.

Here, we have developed a genotyping tool, HBV STAR, that can perform rapid, accurate and statistically robust analysis of HBV sequences without the need to generate phylogenetic trees. We have shown when comparing HBV reference genomes that it is accurate using full-length sequences, as well as when using clinically derived subgenomic sequences. It is also possible to detect recombinant genotypes. The ease of use, sequence-region flexibility and rapidity of the tool allow large databases of HBV clinical sequences to be analysed. As the number of HBV antiviral drugs inevitably increases, so will the requirement for direct sequencing of pol. Given that the surface gene of HBV overlies with pol, this will result in an expanded dataset of surface sequences. We therefore believe that the utility of HBV STAR will increase accordingly. Whilst HBV genotypes are constrained traditionally by the global region from which the host originates, ever-increasing population movements and migrations mean that HBV genotypes will move into new geographical...
regions and new human populations. Identifying and tracking this HBV genotype movement will be critical in terms of both national monitoring of infection rates and predicting the disease and its treatment.

ACKNOWLEDGEMENTS

We thank Dr Yasu Takeuchi and colleagues in the Division of Infection for interesting discussions and constructive comments during the writing of the paper.

REFERENCES


