Genotyping hepatitis B virus dual infections using population-based sequence data

Bastian Beggel,1 Maria Neumann-Fraune,2 Matthias Döring,1 Glenn Lawyer,1 Rolf Kaiser,2 Jens Verheyen2 and Thomas Lengauer1

1Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany
2Institute of Virology, University of Cologne, Cologne, Germany

The hepatitis B virus (HBV) is classified into distinct genotypes A–H that are characterized by different progression of hepatitis B and sensitivity to interferon treatment. Previous computational genotyping methods are not robust enough regarding HBV dual infections with different genotypes. The correct classification of HBV sequences into the present genotypes is impaired due to multiple ambiguous sequence positions. We present a computational model that is able to identify and genotype inter- and intra-genotype dual infections using population-based sequencing data. Model verification on synthetic data showed 100 % accuracy for inter-genotype dual infections and 36.4 % sensitivity in intra-genotype dual infections. Screening patient sera (n=241) revealed eight putative cases of inter-genotype dual infection (one A–D, six A–G and one D–G) and four putative cases of intra-genotype dual infection (one A–A, two D–D and one E–E). Clonal experiments from the original patient material confirmed three out of three of our predictions. The method has been integrated into geno2pheno_hbv an established web-service in clinical use for analysing HBV sequence data. It offers exact and detailed identification of HBV genotypes in patients with dual infections that helps to optimize antiviral therapy regimens. geno2pheno_hbv is available under http://www.genafor.org/g2p_hbv/index.php.

INTRODUCTION

Worldwide an approximate number of 350 million people are chronically infected with the hepatitis B virus (HBV) with mortality rates of about 600 000 per year. The 3.2 kilobase genome has been classified into eight well characterized genotypes A–H based on a nucleotide variation threshold of 8 % over the entire genome (Arauz-Ruiz et al., 2002; Naumann et al., 1993; Norder et al., 1994, 2004; Okamoto et al., 1988; Stuyver et al., 2000). Several authors suggested two additional putative genotypes, namely I (Olinger et al., 2008; Huy et al., 2008; Thuy et al., 2010) and J (Tatematsu et al., 2009). HBV genotypes have been found to influence the rate of chronicity (Ogawa et al., 2002; Suzuki et al., 2005; Zhang et al., 2008), the course of the disease (Chan et al., 2004; Kao et al., 2000a; Livingston et al., 2007; Sánchez-Tapias et al., 2002; Thakur et al., 2002; Yuen et al., 2004) and the response to interferon treatment. Genotypes A and B show a significantly better response to interferon compared with genotypes D and C, respectively (Janssen et al., 2005; Kao et al., 2000b; Wai et al., 2002). Therefore, the German guidelines for the management of HBV infection recommend genotyping before therapy (Cornberg et al., 2008). Nevertheless, treatment with nucleos(t)ide analogues does not appear to be strongly influenced by genotype (Raimondi et al., 2010; Wieand et al., 2008) and thus the major international guidelines do not recommend the determination of the HBV genotype before the start of treatment (European Association For The Study Of The Liver, 2012; Liaw et al., 2008; Lok & McMahon, 2009). Several methods have been developed for the determination of HBV genotypes. Phylogenetic analysis of the entire genome derived from PCR amplification remains the gold standard. Additionally, other computational approaches were proposed to identify the genotype or a recombination of genotypes using full- or subgenome sequence (Myers et al., 2006; Rozanov et al., 2004; Schultz et al., 2009).

Recombinants arise when an individual is infected with multiple strains that exchange segments of genetic viral material. The simultaneous infection with two different HBV strains is referred to as HBV dual infection. Dual infections can either be coinfections or superinfections. Coinfection indicates the situation where a patient was infected with two heterologous strains either simultaneously or within a brief period of time. According to the definition, the infection with the second strain occurs prior
to the immune response to the first infection. Superinfection denotes an infection with a second strain that occurs after the initial infection and after its immune response (Smith et al., 2005). Despite several studies on the natural progression of dual infections, their clinical implications are largely unknown (Hannoun et al., 2002; Kao et al., 2001; Michitaka et al., 2005).

Several studies analysed the prevalence of HBV dual infections. In a study conducted among 183 Chinese HBV carriers dual infections with genotypes B and C were identified in eight (4.4 %) patients (Ding et al., 2003). In an international study that included 256 sera from HBV carriers in Bangladesh, Cameroon, Japan, South Africa, USA and Uzbekistan dual infections with two distinct genotypes were found in 28 (10.9 %) patients (Kato et al., 2003). In a Swedish study of patients with chronic hepatitis B infection of genotype A, signs of dual infection were found in 20 of 30 (67 %) patients treated with interferon (Hannoun et al., 2002). Recently, a study described the distribution of genotypes prevalent in the USA and reported that 15 of 946 (1.6 %) samples exhibited infections with multiple genotypes (Mallory et al., 2011).

Several in vitro methods were developed to identify and genotype dual infections with divergent genotypes (Kato et al., 2003; Naito et al., 2001; Osiowy & Giles, 2003). This study aims to develop and validate the first in silico genotyping method (to the best of our knowledge) that determines the genotypes involved in a multiple HBV infection of a single host using population-based sequence data. In contrast to previous in silico methods which focused on genotyping recombinants it was designed to genotype dual infections. The accurate detection of all genotypes present in a patient sample optimizes the choice of antiretroviral treatment. It also facilitates the determination of the frequencies of dual infections using population-based sequencing in different settings or regions and possibly determines clinical correlations in different patient cohorts.

RESULTS

In silico validation of the dual-infection model on public and synthetic data

A maximum-likelihood classifier, referred to as the dual-infection model, was derived from position- and genotype-specific nucleotide distribution. Profiles were generated using full genome sequence data from GenBank annotated with genotype (Benson et al., 2012). The dual-infection model was applied to synthetic data and to clinical patient sequences to identify and genotype HBV dual infections.

Synthetic inter- and intragenotype dual infections were created by randomly combining pairs of hepatitis B surface antigen (HBsAg) sequences with known genotypes that were obtained from GenBank. All 2800 (100 %) of the intergenotype dual infections were correctly identified including the correct combination of genotypes. Intragene

type infections were correctly identified as such by the dual-infection model in 291 of 800 cases (36.4 %). For the remaining 509 (63.6 %) intragenotype samples, the correct genotype was inferred but the dual infection was overlooked. Additionally, the dual-infection model genotyped all 1791 sequences of the GenBank dataset correctly.

Model predictions were very robust with respect to randomly introduced ambiguities. Ambiguity rates in the range of 1–20 % did not lead to any false-positive predicted intergenotype dual infection. Rare false-positive intragenotype dual infections were observed depending on the level of noise. Ambiguities of 10 % on the HBsAg resulted in 4 of 320 (1.25 %) false-positive intragenotype predictions. The remaining predictions were not impaired. All four of these false positives were of genotype E.

Previously, authors identified intergenotype dual infections by a set of genotype reference sequences. The sequence dissimilarity to the closest reference sequence was computed and a dissimilarity cut-off of 6.0 % was used to identify intergenotype dual infections (Mallory et al., 2011). We applied this methodology to our dataset of synthetic intra- and intergenotype dual-infection sequences. The resulting sequence dissimilarities for intragenotype versus intergenotype dual infections showed distinct distributions (P<0.0001, Wilcoxon rank-sum test) that overlap between 2.9 and 5.2 % sequence dissimilarity (Fig. 1). Thus, the utilization of the dissimilarity score did not facilitate complete separation, and inferred neither intragenotype nor intergenotype dual infection. Additionally, we found that the 6.0 % dissimilarity cut-off proposed by Mallory et al. was too high for optimizing accuracy in our setting where the HBsAg subsegment was used, as only 57.5 % of the intergenotype dual-infection sequences exceeded this cut-off. In contrast, a cut-off of 3.0 % sequence dissimilarity yielded sensitivity in the identification of intergenotype dual infections of 99.6 %, specificity of 94.1 % and the accuracy amounted to 98.4 %, while a cut-off of 5.0 % yielded sensitivity of 84.3 %, specificity of 99.8 % and accuracy of 87.7 %. We also applied the dissimilarity measure to our dataset with randomly introduced ambiguities and obviously found that, as soon as the percentage of generated ambiguities exceeded the dissimilarity cut-off, all sequences were classified as intergenotype dual infections. These evaluations show that methods based on the sequence dissimilarity score could not identify intergenotype dual infections as accurately as the dual-infection model on our synthetic dataset and easily get hampered in the presence of sequencing noise.

In vitro validation of the genotyping method using patient sera

Sequence data obtained from chronically infected patients (n=241) who were diagnosed within the routine diagnostics and therapy procedure covered the reverse-transcriptase
In addition to the in silico analysis, cloning experiments were performed. In three of 12 suspected cases of HBV dual infections a total number of 54 clones were picked, sequenced and genotyped by the National Center for Biotechnology Information (NCBI) genotyping web-service (Rozanov et al., 2004). In nine cases a shortage of material or insufficient PCR amplification obviated potential experiments. For the first patient sample (A–G dual infection, GenBank accession no. JQ776529) 21 clones were successfully sequenced of which 19 were of genotype G, one was of genotype A and one clone was A/G recombinant. The recombinant sequence was analysed with the jumping profile hidden Markov model (Schultz et al., 2009) that inferred two breakpoints. The sequence was estimated to be of genotype G between nucleotide positions (with respect to AM282986) 131 and 375 and between 726 and 1053 and of genotype A between nt positions 376 and 725. For the second patient sample (A–G dual infection, GenBank accession no. JQ776530) 17 clones were sequenced of which 14 were of genotype G, two were of genotype A and one was a recombinant of genotypes A and G. The jumping profile hidden Markov model estimated that nt positions 226–518 were genotype A and nt positions 519–998 were of genotype G. For the third patient sample (A–G dual infection, GenBank accession no. JQ776532) 16 clones were successfully sequenced. Thirteen were of genotype G and three were found to be G/A recombinant. The three recombinant sequences covered nt positions 388–1132, 402–1171 and 389–1093 with estimated breakpoints at nt positions 725, 765 and 782. Thus, the predictions of the dual-infection model were confirmed in vitro for these three patient samples as clonal HBV variants, including the predicted genotypes were found in the respective patient sera. The three original patient sequences did not exceed the 5 % dissimilarity cut-off and were falsely classified as genotype G by the use of the dissimilarity score method. Three of the four putative intragenotype dual infections showed evidence of multiple subgenotypes. Predictions were further investigated by a follow-up analysis where subgenotype-specific nucleotide distributions were employed by the dual-infection model. The A–A dual-infection sequence (GenBank accession no. JQ776527) was identified as an A1–A2 intersubgenotype dual infection. The sequence contained 19 of 454 (4.2 %) ambiguities at the nucleotide level, which implies the presence of two subgenotypes. Further, the HBsAg encoded the ambiguities s207SN and s209L which are characteristic for the presence of subtypes A1 and A2. A1 usually encodes asparagine at position 207 and leucine at position 209 and A2 usually encodes serine at position 207 and valine at position 209 (Norder et al., 2004). Both D–D dual infections were classified as D2–D3 intersubgenotype dual infections. The first of these two sequences (GenBank accession no. HM174233) contained two different serotypes ayw2 and ayw3 as arginine was encoded at position 122, lysine at position 160, glycine at position 159 and threonine at position 140 in addition to the HBsAg ambiguity s127PT (Purdy et al., 2007). In this sequence alanine was present at position 128, while the other sequence (GenBank accession no. HM174239) expressed alanine and valine at position 128 and threonine at position 127 of the HBsAg. In both patient samples, alanine and threonine were present at position 118, and methionine and threonine were expressed at position 125. Thus, the HBsAg amino acid sequences were in 100 % accordance with the conserved amino acid residues of a combination of the subgenotypes D2 and D3 (Tallo et al., 2008). The sequence HM174233 contained 16 of 394 (4.1 %) and the sequence HM174239 contained 13 of 469
The dual-infection model is also capable of identifying intragenotype dual infections. In our synthetic dataset these dual infections were identified in 36.4 % of the cases. The model’s ability to distinguish between intragenotype dual infections and sequencing noise was demonstrated by the false-positive rate of 1.25 % for intragenotype dual infection in the presence of a very high level of sequencing noise (10 % sequence ambiguities).

The dual-infection model identified eight intergenotype and four intragenotype dual infections in our patient cohort (n=241). The majority of the intergenotype cases were A–G and D–G dual infections. This confirms previous observations that genotype G frequently co-occurs with different genotypes (Kato et al., 2002; Osiowy et al., 2008). Additionally, one A–D dual infection was identified (GenBank accession no. IQ776528). This sequence was classified as genotype A by standard genotyping web-services (Myers et al., 2006; Rozanov et al., 2004) in which the risk of interferon treatment failure would have been underestimated due to the overlooked presence of a genotype D strain. Datta et al. (2009) identified and validated five dual-infection sequences (four A–D and one A–C dual infection) with GenBank accession numbers EU275341, EU275342, EU275344, EU275345 and EU275338. These sequences display the same inaccurate classification by standard genotyping methods but are identified correctly by the dual-infection model. The NCBI web-service classifies all five as genotype A and HBV STAR classifies the four A–D dual infections as genotype A and the A–C dual infection remains undetermined.

Validation of our predictions with clonal experiments from the original patient material confirmed three of three predictions. Clonal HBV variants of the predicted genotypes were found in the respective patient samples. No prediction was falsified. But only few samples could be investigated in vitro due to the shortage of original patient material. The analysis of the clonal data using the NCBI genotyping web-service and the jumping profile hidden Markov model for HBV (Schultz et al., 2009) revealed complex quasispecies compositions as two of the three samples contained genotype A variants, genotype G variants and A/G recombinants.
The detailed picture of the viral populations we obtained from the clonal experiments could not be elucidated by the use of the dual-infection model alone. Firstly, the dual-infection model does not account for combinations of more than two viral strains. Secondly, the dual-infection model does not account for recombinants as it is addressed to analyse relatively short sequence data acquired for routine diagnostics. These sequences usually cover the part of the reverse-transcriptase domain, which contain the major drug resistance mutations (approx. aa positions 80–250). Full genome sequence data that arise from dual infections with at least one recombinant strain involved could for instance be analysed by combining the dual-infection model with a sliding window approach (Myers et al., 2006; Rozanov et al., 2004) or by combination with the so called jumping profile hidden Markov model (Schultz et al., 2009).

We predicted four cases of intragenotype dual infection within our patient cohort. The intragenotype predictions are difficult to validate and might only reflect the viral quasispecies within the patient. For three of the four predicted cases we have strong evidence of intersubgenotype dual infections. Firstly, the total number of ambiguous sequence positions was high (13, 16 and 19). Secondly, the expressed amino acids were in concordance with the conserved residues of the combination of subgenotypes predicted by the dual-infection model. It is more likely that the observed combination of amino acids results from a dual infection with existing subgenotypes than that it is the result of intra-patient evolution. Intra-patient evolution is unlikely to develop the very same combination of exact amino acids which are conserved in patient evolution is unlikely to develop the very same combination of exact amino acids which are conserved in

The dual-infection model provides an in silico alternative to labour- and money-intensive clonal experiments. It requires only HBV sequences obtained by routine diagnostics. Alternatives such as massively parallel sequencing approaches, also called next generation sequencing, are able to analyse this problem with higher sensitivity with respect to minor subpopulations, but none of the systems is expected to replace the current population-based sequencing in the clinical routine setting within the near future due to cost reasons.

The dual-infection model outperformed the identification of dual infection based on sequence similarity. It could genotype the sample sequences, achieved higher accuracy on synthetic data (100 vs 88.5 % with 5 % dissimilarity cut-off) and was more robust with respect to sequencing errors (with 10 % sequencing noise: 0 % false-positive intergenotype dual infections vs 100 % false-positive intergenotype dual infections). In our patient cohort using the 5 % dissimilarity cut-off identified only two of the eight intergenotype dual infections reported by the dual-infection model. Three validated dual infections did not exceed the 5 % dissimilarity cut-off. Lower cut-offs were not suitable for our patient cohort either, due to false-positive intergenotype predictions. Other standard genotyping methods do not account for dual infections and thus bear the risk of leading to suboptimal-treatment decisions when present genotypes are overlooked.

In summary, we presented the first in silico genotyping method that can identify and genotype HBV intergenotype dual infections reliability. The method can elucidate the frequency of dual infections in the routine diagnostics and can be helpful in optimizing antiviral therapy. It is freely available as part of geno2pheno_hbv at http://www.genafor.org/g2p_hbv/index.php and can be used with HBV sequences generated within the clinical routine.

**METHODS**

**GenBank sequence data.** HBV full genome sequences (3796) were downloaded from GenBank in August 2011 (Benson et al., 2012). The sequence meta-information was parsed to annotate the genotype. Sequences without genotype annotation were excluded along with sequences whose annotations included the keywords 'defective' or 'non-functional' or 'recomb'. Further, we excluded sequences that contain at least one stop codon within the HBsAg or within the reverse-transcriptase domain. Additionally, the sequences with GenBank accession numbers AB486012, AF461362, AY293309, EF103284, EU305546, EU939634, GQ377573, GU357844 and HQ231877–HQ231885 were removed from the dataset due to evidence from the NCBI genotyping web-service implicating recombinants or false-genotype annotations (Rozanov et al., 2004). Duplicates with identical HBsAg sub-sequence were also removed. The goal of the filtering step was to obtain a dataset of high quality, non-redundant and non-recombinant sequences with correct genotype annotation that could be used to train and to validate our genotyping methodology. This procedure resulted in 1791 genotype annotated sequences (270 for genotype A, 339 for genotype B, 636 for genotype C, 304 for genotype D, 159 for genotype E, 54 for genotype F, 12 for genotype G and 17 for genotype H). Subgenotype predictions were based on sequence data and subgenotype annotations supplied in the literature (Norder et al., 2004; Tallo et al., 2008).

**Prediction model.** The type of infection was indicated by the single infection or dual-infection models based on the maximum-likelihood principle. The likelihood of the test sequences are evaluated under each model. The model which provides the highest data likelihood is determined according to the maximum-likelihood principle. Data likelihoods were derived from position- and (sub)genotype-specific nucleotide distributions (PGSN) separately for each (sub)genotype or combination of (sub)genotypes.

The PGSNs used here were based on the GenBank sequence data. After aligning the sequences to the reference sequence GenBank accession no. AM282986 the observed frequency of the four DNA bases per reference sequence position and HBV (sub)genotype were computed. \( p_{gtj} \) denotes the PGSND for (sub)genotype \( g \) and position \( j = 1,...,3221 \). These distributions encode information about (sub)genotype-specific polymorphisms at the nucleotide level and are available as supplementary table (Table S1, available in JGV Online). Conditioning on a single (sub)genotype or on a combination of (sub)genotypes – in case of a dual infection – the data likelihood of an HBV sequence represented by a set of sequence positions \( r_{i,...,ri,...} \) was computed based on the PGSND. Here, each \( r \)
Table 2. Dual-infection sequences identified by the dual-infection model

Patient sera (n=241) were sequenced, genotyped and screened for dual infections. The dual-infection model identified eight intergenotype and four intragenotype dual infections. The sequence dissimilarities to a set of genotype reference sequences were computed. All polymorphisms which increased the odds of the dual infection by a factor of at least 2.0 are reported.

<table>
<thead>
<tr>
<th>GenBank accession no.</th>
<th>Prediction of the dual-infection model</th>
<th>Most similar genotype reference</th>
<th>Dissimilarity to most similar genotype reference</th>
<th>Polymorphisms indicative of dual infection†</th>
</tr>
</thead>
<tbody>
<tr>
<td>JQ776527</td>
<td>A–A</td>
<td>A</td>
<td>4.18 %</td>
<td>C409AC (F85FL), G436GA (L94L), T451TC (D99D), C498AC (T115T), A519GA (K122KR), C528TC (T125MT), T555AT (F134FY), T581AT (T143ST), G616GA (S154S), T619TC (S155S), G774GA (S207NS), G779GT (V209LV), C820TC (L222L)</td>
</tr>
<tr>
<td>JQ776528</td>
<td>A–D</td>
<td>A</td>
<td>6.26 %</td>
<td>G436GA (L94L), T451TC (D99D), A494AT (T114ST), A499AT (T115T), T505TC (S117S), A519GA (K122KR), C544AC (G130G), T555AT (F134FY), A562AC (S136S), A574AC (T140T), T581AT (T143ST), T586TC (D144DV), C630GC (A159AG), T657TC (V168AV), A791AT (I213IL)</td>
</tr>
<tr>
<td>JQ776529</td>
<td>A–G</td>
<td>G</td>
<td>4.58 %</td>
<td>G436GA (L94L), T451TC (D99D), A481GA (L109L), A493AC (S113S), A496GA (T114S), T555AT (F134FY), T581AT (T143ST), T586TC (D144D), T619TC (S155S), T724TC (V190V), G765GA (S204NS)</td>
</tr>
<tr>
<td>JQ776530</td>
<td>A–G</td>
<td>G</td>
<td>2.52 %</td>
<td>A481GA (L109L), A493AC (S113S), A494AT (T114ST), A496GA (T114S), A499AC (T115T), T555AT (F134FY), T581AT (T143ST), T586TC (D144D), T619TC (S155S)</td>
</tr>
<tr>
<td>HM174195</td>
<td>A–G</td>
<td>G</td>
<td>5.35 %</td>
<td>T555AT (F134FY), T581AT (T143ST), T586TC (D144DG), T619TC (S155S), T724TC (V190V), G765GA (S204SN), T832TC (I226I)</td>
</tr>
<tr>
<td>JQ776531</td>
<td>A–G</td>
<td>G</td>
<td>4.49 %</td>
<td>G436GA (L94L), T451TC (D99D), A481GA (L109L), A493AC (S113S), A494AT (T114ST), A496GA (T114S), A499AC (T115T), T555AT (F134FY), T581AT (T143ST), T586TC (D144D), T619TC (S155S), T724TC (V190V), G765GA (S204NS)</td>
</tr>
<tr>
<td>JQ776532†</td>
<td>A–G</td>
<td>A/G</td>
<td>4.99 %</td>
<td>T451TC (D99D), A481GA (L109L), A493AC (S113S), A496GA (T114S), A499AC (T115T), T555AT (F134FY), T581AT (T143ST), T586TC (D144D), T619TC (S155S), T724TC (V190V), G765GA (S204NS)</td>
</tr>
<tr>
<td>JQ776533</td>
<td>A–G</td>
<td>G</td>
<td>4.78 %</td>
<td>T300TC (L49LP), C342TC (T63IT), T357TC (I68IT), G436GA (L94L), T451TC (D99D), A493AC (S113S), A494AT (T114ST), A496GA (T114S), A499AC (T115T), T555AT (F134FY), T581AT (T143ST), T586TC (D144D), T619TC (S155S), T724TC (V190V)</td>
</tr>
<tr>
<td>HM174233</td>
<td>D–D</td>
<td>D</td>
<td>4.57 %</td>
<td>A453AC (Y100SY), T472TC (V106V), A481GA (L109L), A493TC (S113S), A506GA (T118AT), C512AC (P120PT), A518AC (K122R), C523AC (T123T), C528TC (T125MT), C531AC (T126NT), C533AC (P127PT), T784GT (S210RS), A791AT (I213IL)</td>
</tr>
<tr>
<td>HM174239</td>
<td>D–D</td>
<td>D</td>
<td>3.41 %</td>
<td>T472TC (V106V), T493TC (S113S), A506GA (T118AT), C527TC (T118TM), A518AC (K122R), C523AC (T123T), C528TC (T125MT), C533A (P127T), T537TC (A128AV), G567GA (C138CY), G570GA (C139CY), A706AC (V184V), T784GT (S210RS), T814AT (F220FL)</td>
</tr>
<tr>
<td>HM174203</td>
<td>D–G</td>
<td>G</td>
<td>4.98 %</td>
<td>A481GA (L109L), A496GA (T114S), T505TC (S117S), A514AC (P120P), A519GA (K122KR), C544AC (G130G), A564AC (N131NT), A562AC (S136S), A574AC (T140T), T619TC (S155S), C630GC (A159AG), T657TC (V168AV), T724TC (V190V), G765GA (S204NS), G774GA (S207NS)</td>
</tr>
</tbody>
</table>
Reference sequences X02763, D00329, X01587, V01460, X75657, X75658, AF160501, AY090460 were used for HBV genotypes A–H.

The sequence position is computed as divergence of genotypes need only be 8 % or greater. The likelihood of both strains. This case is not unusual since the overall intergroup population of viruses that exhibit mutations at certain positions, but all originate from the same initial strand.

**Single-infection likelihood.** For the single genotype case, the likelihood of an unambiguous read \( r_i = b \) (with \( b \) being one of the four bases found in DNA: adenine, cytosine, guanine or thymine) is computed with \( \text{Pr}(r_i=b \mid gt)=p_{gt}[b] \). Here, \( p_{gt}[b] \) denotes the fraction of base \( b \) being observed at position \( j \) for genotype \( gt \) and therefore equals the likelihood that base \( b \) is present given genotype \( gt \). For an ambiguous read \( r_i=(b_1,b_2) \) where \( b_1 \) and \( b_2 \) are two divergent bases the likelihood of the read is computed as \( \text{Pr}(r_i=(b_1,b_2) \mid gt)=p_{gt}[b_1] \cdot p_{gt}[b_2] \). Ambiguous positions for a single genotype model can be explained by the concept of viral quasispecies, which describe a population of viruses that exhibit mutations at certain positions, but all originate from the same initial strand.

**Dual-infection likelihood.** A dual infection is described by a combination of two not necessarily different genotypes \( gt_1 \) and \( gt_2 \). The likelihood of the sequence positions \( r_{f_1...f_{n+1}} \) are computed using the assumption that population-based sequencing provides the union of the bases of the two underlying viral strains. Thus, if the sequence position \( r_i \) is unambiguous, there is only a single base \( b \) at position \( j \) in both strains. This case is not unusual since the overall intergroup divergence of genotypes need only be 8 % or greater. The likelihood of the sequence position is computed as \( \text{Pr}(r_i=b \mid gt_1, gt_2)=p_{gt_1}[b] \cdot p_{gt_2}[b] \).

The case of an ambiguous sequence position of cardinality two \( r_i=(b_1, b_2) \) is difficult to interpret because the phase information of the two bases is lost in the sequencing process, i.e., we do not know to which genotype, \( gt_1 \) or \( gt_2 \), each of the two bases, \( b_1 \) and \( b_2 \), belong. This corresponds to two biological events: in the first, \( b_1 \) is present in a strain of genotype \( gt_1 \) and \( b_2 \) is present in a strain of \( gt_2 \); in the second, \( b_1 \) is present in a strain of \( gt_1 \) and \( b_2 \) is present in a strain of \( gt_2 \). The likelihood of the first event is \( p_{gt_1}[b_1] \cdot p_{gt_2}[b_2] \), while the likelihood of the second event is \( p_{gt_2}[b_1] \cdot p_{gt_1}[b_2] \). As these events are mutually exclusive, we can compute the likelihood of the sequence position by summation: \( \text{Pr}(r_i=(b_1,b_2) \mid gt_1, gt_2)=p_{gt_1}[b_1] \cdot p_{gt_2}[b_2] + p_{gt_2}[b_1] \cdot p_{gt_1}[b_2] \). The likelihood computation is extended by a regularization term that accounts for sequencing errors in the input sequence and inaccuracies in the PGSN. The adjusted likelihood of a sequence position is computed as \( \text{Pr}'(r_i)=\text{Pr}(r_i) \cdot k_j \cdot (1-\text{Pr}(r_i)) \cdot \text{Pr}(r_i) \), where \( k_j \) equals 1/4 if \( r_j \) (A,C,G,T) is unambiguous and 1/6 if \( r_j \) (K,M,R,Y,W) is ambiguous with cardinality two. Ambiguous sequence positions with cardinality greater than two are not included in the likelihood computation. The error constant \( p_{error} \) is set to the reported error rate of Sanger sequencing which is 1 % (Keith et al., 1993). Predictions are based on the adjusted total likelihood that summarizes all adjusted positional likelihoods \( \text{Pr}^*(r_i...r_i+n)=\text{Pr}^*(r_i) \cdot \text{Pr}^*(r_{i+1}) \cdot \text{...} \cdot \text{Pr}^*(r_{i+n}) \). Sequence positions known to be associated with drug resistance are excluded from the analysis as they might bias the model (Shaw et al., 2006). All other sequence positions are included in the computation with the position- and (sub)genotype-specific nucleotide probabilities observed in the GenBank dataset.

**Synthetic test data.** Testing data consisting of synthetic intra- and intergenotype dual infections were created by combining pairs of sequences of known genotype from the GenBank dataset. This procedure generated sequence data with high numbers of ambiguities representing dual infections that simulate population-based sequence data of \( \text{in vivo} \) dual infections. For each combination of genotypes A–H, 100 test sequences were created by randomly selecting two sequences of the respective genotypes and then constructing the position-wise union of the sequence bases after alignment to the reference strain AM282986. This position-wise union of two sequences emulates the perception of how population-based sequence data of patients with HBV dual infections arise, when both strains are above the sequencing detection sensitivity. The procedure generated 3600 (36 combinations times 100 samples) test sequences, of which 800 were intragenotype (i.e. A and A) and 2800 were intergenotype (i.e. A and B).

The impact of sequencing errors on the identification and genotyping of dual infections was analysed. For each HBV genotype, 40 sequences were randomly selected (with replacement) from the GenBank dataset. Ambiguities of 1–20 % were randomly introduced to simulate sequencing errors.

All computational experiments based on the GenBank sequence data were performed with fivefold cross-validation and were carried out on the HBsAg sub-sequence. Accuracy, sensitivity and specificity were used to assess prediction quality.

**Genotype references.** Reference sequences X02763, D00329, X01587, V01460, X75657, X75658, AF160501, AY090460 were used for HBV genotypes A–H (Arauz-Ruiz et al., 2002; Stuyver et al., 2001). Sequences obtained from clonal experiments were analysed using the NCBI genotyping web-service (Rozanov et al., 2004).

**HBV sequence data.** HBV sequences from routine diagnostics performed between 2002 and 2010 were employed in this study. The HBV genome region encoding partly the polymerase and the HBsAg...
were amplified and sequenced as described previously (Schildgen et al., 2004).

Clonal analyses: the oligonucleotides HBrt-AccIII-fw (5'-GTCACTGAGACATGTTTACAGAAGCC-3') and HBrt-Rsr-II-rev (5'-GCGACATGCTCAGACGAATAAGGCC-3') were used to amplify the HBV polymerase of HBV isolates from the original serum samples. The amplification products and the vector (pCH9-3091) were cut with AccIII and RsrII preceding a ligation step. After transformation using JM110 (Promega) plasmids were picked and analysed by sequencing. All sequencing was performed using oligonucleotides HBsAg-Kpol-fw (5'-GTCACTGTTACCATGGA-GAGCAACAATCAGGATT-3') and HBsAg-MH-fw (5'-GTCACTCATAATGCTTCTATCCTGGTATG-3') and HBV S6 antisense (5'-CTTTGACADCTTTCCAATCATAG-3'). The products were sequenced with the ABI Prism 3130xl and edited with DNASTAR Lasergene.

ACKNOWLEDGEMENTS

The study was supported by the German Bundesministerium für Bildung und Forschung (grant no. BMBF01ES0822). The initial idea to develop a computational genotyping model for HBV dual infections was raised by Martin Däumer.

REFERENCES


