JOURNAL OF GENERAL VIROLOGY

RESEARCH ARTICLE

Forbi *et al.*, *Journal of General Virology* 2017;98:1048–1057

DOI 10.1099/jgv.0.000776

MICROBIOLOGY SOCIETY

Molecular epidemiology of hepatitis B virus infection in Tanzania

Joseph C. Forbi,[1,*] Michael Dillon,[2] Michael A. Purdy,[1] Bakary S. Drammeh,[3] Alexandra Tejada-Strop,[1] Daniel McGovern,[1] Guo-liang Xia,[1] Yulin Lin,[1] Lilia M. Ganova-Raeva,[1] David S. Campo,[1] Hong Thai,[1] Gilberto Vaughan,[1] Dunstan Haule,[4] Regina P. Kutaga,[5] Sridhar V. Basavaraju,[3] Saleem Kamili[1] and Yury E. Khudyakov[1]

Abstract

Despite the significant public health problems associated with hepatitis B virus (HBV) in sub-Saharan Africa, many countries in this region do not have systematic HBV surveillance or genetic information on HBV circulating locally. Here, we report on the genetic characterization of 772 HBV strains from Tanzania. Phylogenetic analysis of the S-gene sequences showed prevalence of HBV genotype A (HBV/A, *n*=671, 86.9 %), followed by genotypes D (HBV/D, *n*=95, 12.3 %) and E (HBV/E, *n*=6, 0.8 %). All HBV/A sequences were further classified into subtype A1, while the HBV/D sequences were assigned to a new cluster. Among the Tanzanian sequences, 84 % of HBV/A1 and 94 % of HBV/D were unique. The Tanzanian and global HBV/A1 sequences were compared and were completely intermixed in the phylogenetic tree, with the Tanzanian sequences frequently generating long terminal branches, indicating a long history of HBV/A1 infections in the country. The time to the most recent common ancestor was estimated to be 188 years ago [95 % highest posterior density (HPD): 132 to 265 years] for HBV/A1 and 127 years ago (95 % HPD: 79 to 192 years) for HBV/D. The Bayesian skyline plot showed that the number of transmissions 'exploded' exponentially between 1960–1970 for HBV/A1 and 1970–1990 for HBV/D, with the effective population of HBV/A1 having expanded twice as much as that of HBV/D. The data suggest that Tanzania is at least a part of the geographic origin of the HBV/A1 subtype. A recent increase in the transmission rate and significant HBV genetic diversity should be taken into consideration when devising public health interventions to control HBV infections in Tanzania.

INTRODUCTION

Hepatitis B virus (HBV) infection remains a major public health concern, with ~400 million people chronically infected worldwide. It is associated with a broad spectrum of clinical presentations, ranging from asymptomatic infection to serious liver diseases such as liver cirrhosis and hepatocellular carcinoma [1, 2]. The World Health Organization estimates that about 600 000 people die every year due to acute or chronic hepatitis B [2, 3]. Sub-Saharan Africa accounts for a large majority of these infections and has been described as a hyper-endemic region for the disease [4]. In Tanzania (located in East Africa), the prevalence of HBV infection has reached 17.3 % among HIV-infected adults [5]. With the presence of social networks among men who have sex with men and persons who inject drugs in Tanzania [6], the rate of infection is expected to remain high, as these groups have been associated with increased transmission and acquisition of HBV infection [7]. Only limited information is available on the genetic composition of the HBV population circulating in Tanzania. However, this information is relevant for estimating the potential severity and progression of chronic hepatitis B, the response to antiviral therapy and the risk of liver cancer [8]. Genetic characterization of currently circulating HBV strains should provide crucial insights into the global expansion and evolution of HBV and will help in creating more efficient disease interventions and patient management strategies.

The HBV genome is a small partially double-stranded DNA that is ~3.2 kb in length and contains four partially overlapping ORFs encoding the P (polymerase), C (core), S (surface) and X proteins [9]. The S protein is a critical target of

Received 9 September 2016; Accepted 15 March 2017

Author affiliations: [1]Division of Viral Hepatitis, National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Rd, NE, Atlanta, Georgia, USA; [2]CDC Tanzania, Division of Global HIV/AIDS, Center for Global Health, Centers for Disease Control and Prevention, Dar es Salaam, Tanzania; [3]HIV Prevention Branch, Division of Global HIV/AIDS, Center for Global Health, CDC, Dar es Salaam, Tanzania; [4]Tanzania National Blood Transfusion Services, Ministry of Health and Social Welfare, Dar es Salaam, Tanzania; [5]US Centers for Disease Control and Prevention, Dar es Salaam, Tanzania.

***Correspondence:** Joseph C. Forbi, gzf7@cdc.gov *or* JForbi@cdc.gov

Keywords: hepatitis B virus; molecular epidemiology; evolution; Tanzania.

Abbreviations: Befi-BaTS, Bayesian tip-associationsignificance testing; HBV, hepatitis B virus; HBsAg, hepatitis B surface antigen; HKY, Hasegawa–Kishino–Yano; HPD, highest posterior density; tMRCA, time to the most recent common ancestor.

Accession numbers of new sequences: KU594654–KU595425.

000776

the host's neutralizing antibodies [10, 11] and the S gene is frequently used for genotyping, phylogenetic and evolutionary analyses [12, 13]. Here, we characterize HBV strains from Tanzania and analyse their genetic history and geographic distribution. The data suggest that Tanzania is, at least in part, the geographic origin of the HBV/A1 subtype and HBV has a long history in the country.

RESULTS

Phylogenetic analysis

Phylogenetic analysis of the 772 HBV S-gene sequences (Northern zone, *n*=22; Eastern, *n*=211; Southern, *n*=94; Western, *n*=7; Zanzibar, *n*=29; Lake, *n*=373; Southern Highlands, *n*=36) identified three HBV genotypes circulating in Tanzania, HBV/A (*n*=671; 86.9 %), HBV/D (*n*=95; 12.3 %) and HBV/E (*n*=6; 0.8 %) (Fig. 1). The genotype distribution by zones is shown in Fig. 2. To identify subtypes of the sampled HBV strains, phylogenetic analysis was performed using a set of reference sequences from HBV strains with known HBV/A and HBV/D subtypes. It was found that while HBV/A reference sequences clustered according to their subtypes, HBV/A1, HBV/A2 and HBV/A3, the HBV/D sequences of different subtypes were intermixed, indicating that accurate subtyping using the S-gene sequence cannot be accomplished for HBV/D strains (Fig. 3).

All HBV/A sequences from Tanzania clustered together with the HBV/A1 reference sequences, suggesting the predominance of the subtype in the country. It is important to note that 84 % of HBV/A1 sequences sampled in this study were unique, while only 44 and 60 % of the S-gene sequences from the recently identified HBV/A1 Asian-American and African clusters, respectively, were unique [14]. Additionally, many sequences formed long terminal branches in the phylogenetic tree, reflecting significant genetic differences among the Tanzanian HBV/A1 strains (Figs 1 and 3). The nucleotide diversity among the Tanzanian HBV/A1 sequences was 2.9 %, while it was 1.6 and 1.4 % among HBV/A1 sequences from the African and Asian-American clusters, respectively. The HBV/D sequences from Tanzania, however, did not intermix with any of the reference sequences used here, suggesting a significant genetic distinction of the Tanzanian HBV/D strains from those collected from GenBank. Similar to HBV/A1, the HBV/D sequences from Tanzania were genetically heterogeneous and formed long branches in the phylogenetic tree. The genetic diversity among the Tanzanian HBV/D sequences was 3.1 %, while it was 1.9 % among the all other reference HBV/D sequences. No significant difference in diversity was found between the Tanzanian HBV/A1 and HBV/D (P>0.05).

Genetic history

Bayesian coalescent analysis estimated the average time to the most recent common ancestor (tMRCA) for HBV/A1 to be 188 years ago (95% HPD: 132 to 265) and 127 years ago (95% HPD: 79 to 192) for HBV/D. Skyline plots indicate that HBV/A1 went through a continuous expansion that started in ~1890 and experienced an extraordinarily rapid inflation starting in ~1960. After 1970, the effective population of HBV/A1 became constant and it has remained constant to the present day (Fig. 4). HBV/D also experienced a population expansion that went in two phases: slow from ~1920–1965 and rapid from ~1970–1990 (Fig. 5). The total increase in the effective population seen in HBV/D is only about 2 logs, as compared to that for HBV/A which expanded by about 4 logs (Figs 4 and 5).

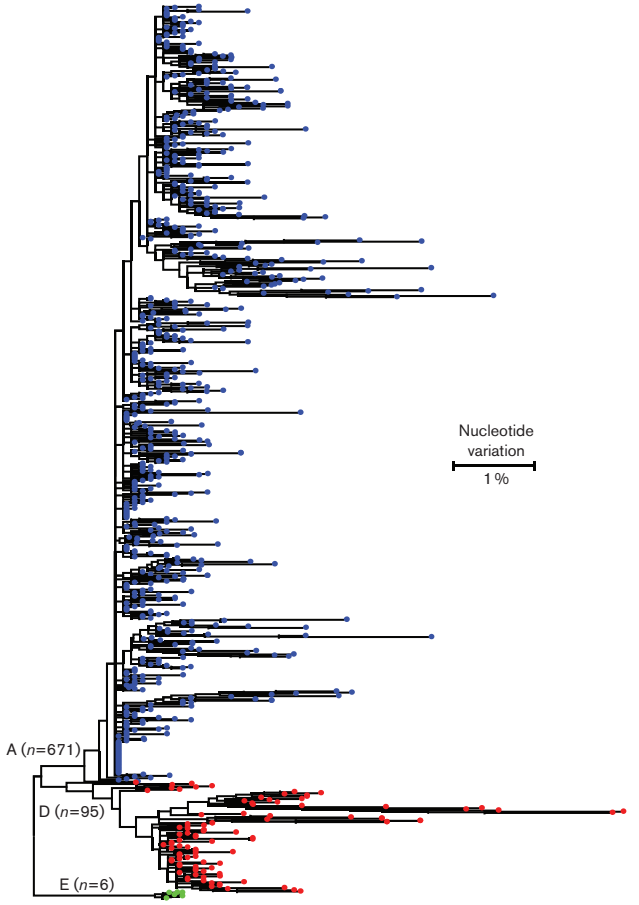

Fig. 1. Maximum likelihood tree for the HBV S-gene sequences determined in this study, showing genotype classification: HBV/A (blue), HBV/D (red) and HBV/E (green).

Phylogeographic association

The distribution of the HBV genotypes by zone of collection was as follows: East (HBV/A=186, HBV/D=22, HBV/E=3); Lake (HBV/A=304, HBV/D=68, HBV/E=1); North (HBV/A=20, HBV/D=2, HBV/E=0); Southern Highland (HBV/A=32, HBV/D=2, HBV/E=2); South (HBV/A=94, HBV/D=0, HBV/E=0); West (HBV/A=6, HBV/D=1, HBV/E=0); Zanzibar (HBV/A=29, HBV/D=0, HBV/E=0). Bayesian tip-association significance testing (Befi-BaTS) was used to test the potential association between the geographic locations and the phylogenetic relationships among HBV strains. All the

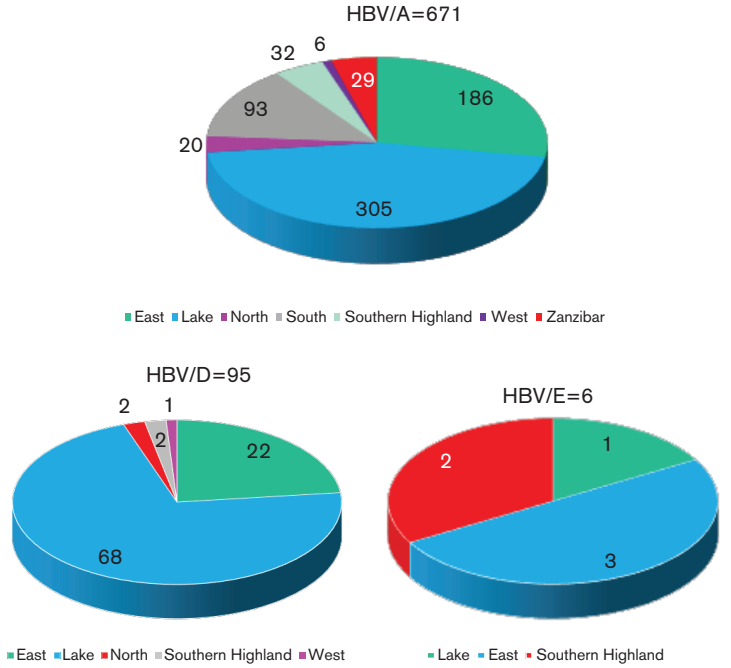

Fig. 2. Pie chart showing the distribution of HBV genotypes by geographic zones (Tanzania). The numbers represent the exact number of isolates and the colors represent the different zones.

methods in Befi-BaTS indicate that geographic distribution reflects phylogenetic relationships among HBV/A1 strains (P=0.01, Table 1). The monophyletic clade size allowed us to examine each geographic region individually. Sequences from the North, East, South, Lake, Zanzibar and Southern Highland zones were not distributed randomly on the tips of the corresponding phylogenetic trees (P=0.01), while sequences from the West zone appeared to be randomly distributed on the tree (P=1.0). A similar observation was made for HBV/D, indicating that the geographic location is non-randomly distributed across the tree (P=0.05). This finding was supported by all measures but the net relatedness index (P=0.6, Table 2). While the sequences from the Lake (P=0.01) and East zones were not found to be randomly associated with tips of the HBV/D phylogenetic tree (P=0.01), the sequences from the North, East and Southern Highland zones were randomly distributed (P=1.0), which is likely due to the small sequence numbers from the latter regions. HBV/D strains were only sampled from five (North, East, West, Southern Highland and Lake) of the seven collection zones in Tanzania, with ~71.6 % of all HBV/D sequences being obtained from the Lake zone.

Vaccine-escape substitutions

Specific amino acid substitutions, A128V, Q129H and M133T, within the 'a' determinant region of the S-protein associated with vaccine escape were identified in 3 of 671 HBV/A1 sequences (0.4 %), while HBV/D and HBV/E did not bear any of the known vaccine escape substitutions. The G145R mutation was not detected in any of the Tanzanian sequences. However, HBV/A1 yielded an additional nine rare mutations within the 'a' determinant – T126S (n=2), A128G (n=1), A128Y (n=1), M133I (n=1), D144E (n=1), D144N (n=1) and D144G (n=2) – while HBV/D yielded the A128G and Q129P mutations. The role of these mutations is not yet clearly understood.

DISCUSSION

This study describes the molecular epidemiological analysis of HBV genomic sequences from several geographical zones in Tanzania. Our data show that 86.9 % of HBV circulating in Tanzania belongs to HBV/A. This is consistent with a previous report in which 90.9 % of HBV-infected voluntary blood donors in Tanzania were infected with HBV/A [15]. It is obvious that, although the number of HBV isolates detected here does not come from the entire country of Tanzania, our finding is consistent with the previous report on HBV genotypes in the country [15]. All HBV/A strains reported here were classified into subtype HBV/A1 in concordance with other studies [15, 16]. It is important to note that HBV/A1 is predominant in Malawi, Uganda, Kenya [16–18] and Rwanda [19], the countries that share borders with Tanzania (Fig. 6). The Tanzanian HBV/A1 variants were found frequently at the long branches of the phylogenetic tree and completely intermixed with the global HBV/A1 sequences (Fig. 3), reflecting the long evolutionary history of this subtype in the country.

The Tanzanian HBV/A1 S-gene sequences are extraordinarily diverse. Among all of them, 84 % are unique, despite the extensive sampling, while a similarly extensive sampling from the USA showed the existence of large clusters of HBV/A2 strains sharing the S-gene sequences [20], with

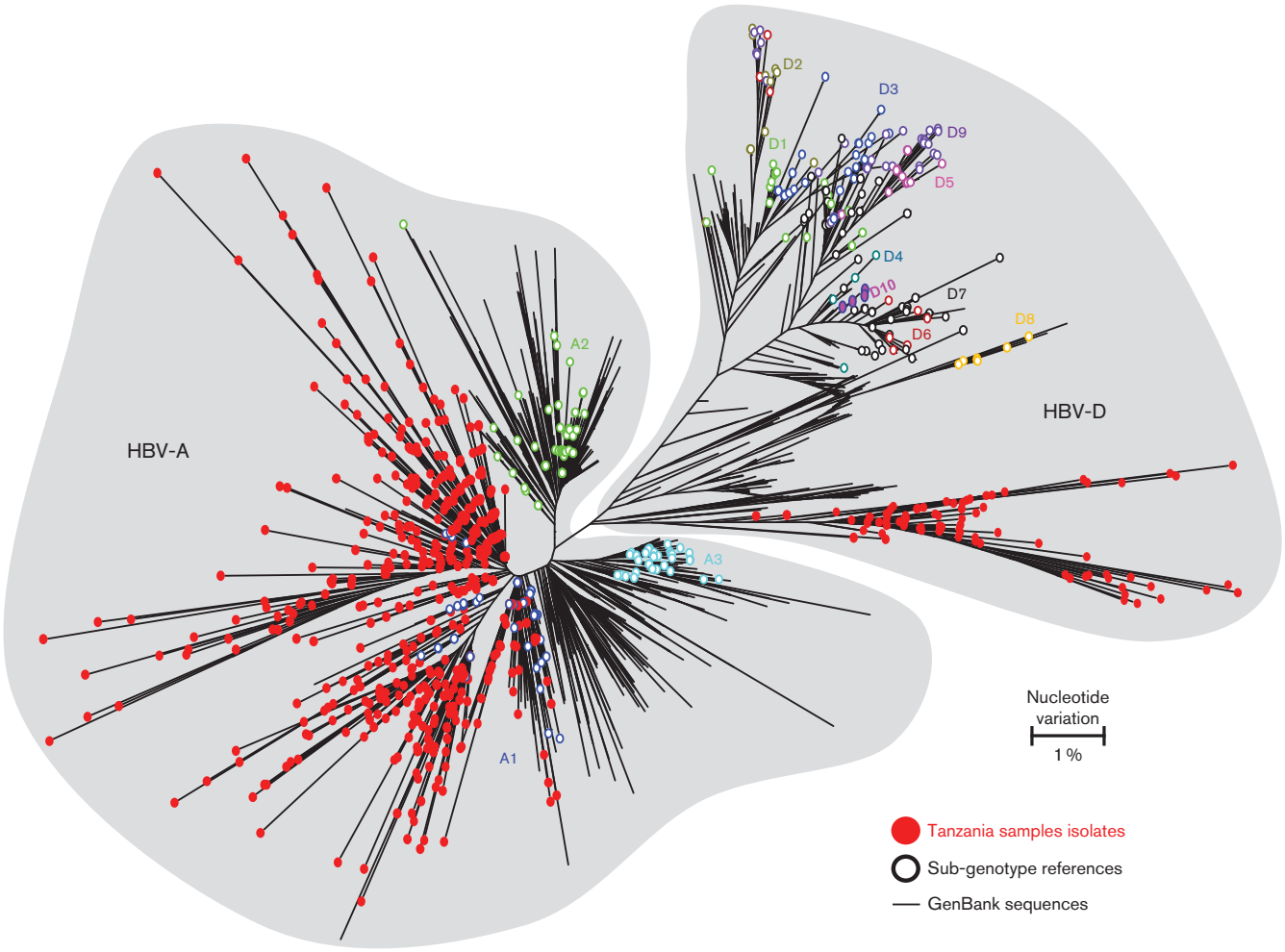

Fig. 3. Maximum-likelihood tree for the HBV/A S-gene sequences determined in this study and those available in GenBank: Tanzanian isolates (HBV/A1) and HBV/D are shown in red. GenBank references of subtypes A/1-3 and HBV/D1-9 are shown in other colours.

only 53 % of the sequences being unique. Recently, HBV/A1 strains were classified into Asian-American and African clusters, in which only 44 and 60 % of the S-gene sequences were unique, respectively [14]. The predominance of a diverse HBV/A1 population suggests that Tanzania is, at least in part, the geographic origin of the subtype, which is in concurrence with the hypothesis of the East African origin of HBV/A1 [16, 17, 21]. If so, it is interesting that there is a strong segregation of HBV/A1 clades of the S-gene sequences among geographic zones in Tanzania, which indicates either the introduction of the clades to different geographic zones in Tanzania from other East African countries or the independent origin of the clades in each zone. Either way, a strong association between the geographic distribution within the country and the phylogenetic relationships among HBV/A1 strains indicates a particular geographic distribution and further suggests a difference in the geographic origin for the HBV/A1 clades. It is likely that this geographic association plays a major role in the presentation of HBV genotypes in Tanzania. However, further studies are required to understand the role of geographic isolation in the partitioning of HBV genotypes seen in Tanzania and Africa as a whole.

HBV/D strains were identified in 12.3 % of the tested specimens from Tanzania. These HBV/D variants could not be further classified into subtypes, indicating that the S-gene sequences should be used cautiously for the subtyping of HBV strains. Although the Tanzanian HBV/A S-gene sequences were found to be faithfully segregated into subtypes, the HBV/D reference sequences of different subtypes were found to be completely intermixed (Fig. 3). Nevertheless, the Tanzanian HBV/D variants were found to be segregated into a separate cluster and were not intermixed with any known HBV/D subtype, suggesting their unique origin in Tanzania. A smaller sample size of HBV/D sequences only allowed for the statistically significant identification of phylogenetic associations for the Lake and East zones, which still indicates that, similar to HBV/A1, HBV/D strains have a specific distribution in the country.

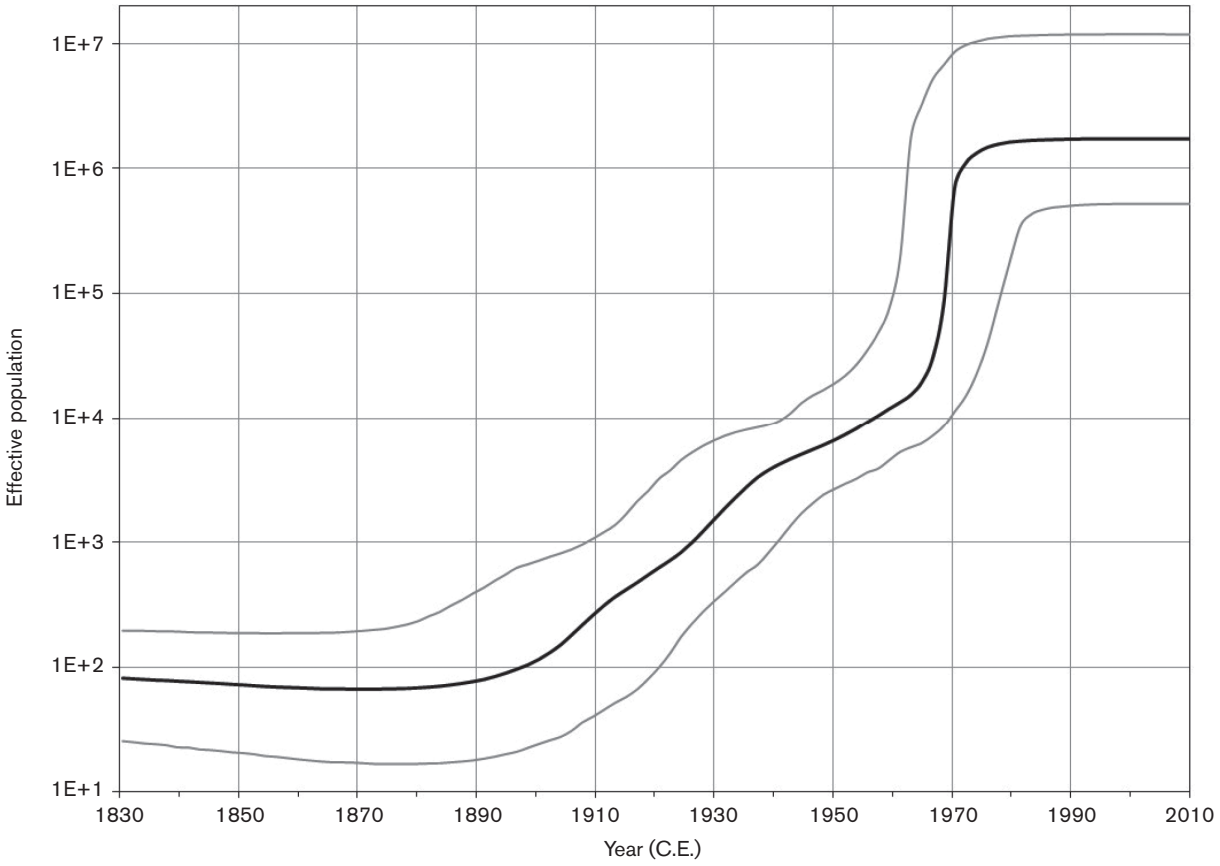

Fig. 4. Skyline plot for Tanzanian HBV genotype A sequences. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. Gray lines mark the limits of the 95 % HPD for the effective population and the solid black line is the mean of this distribution.

Only six HBV/E strains (0.8 %) were detected in this study. HBV/E circulates almost exclusively in West/Central Africa [22, 23]. The geographic origin of HBV/E isolated in this study remains unknown. However, low prevalence and genetic diversity indicate that this genotype was introduced to Tanzania. The six HBV/E strains probably originated from West Africa or have some links associated with transmission from West Africa, where HBV/E is prevalent [22, 23]. The travel history of the six individuals infected with HBV/E is also unknown.

The identification of many divergent strains of three genotypes (Figs 1 and 3) suggests a long evolutionary history of HBV in Tanzania. Time-scaled phylogenetic reconstruction indicates that the HBV/A tMRCA is ~188 years ago. The tMRCA of HBV/D was estimated to be ~127 years ago. Interestingly, this coincides with the tMRCA (128 years ago) of HBV/D from India [24]. This suggests that the common ancestor of the currently circulating HBV/D in Tanzania existed at the same time as those from the Indian sub-continent [24]. A skyline plot showed that the Tanzanian HBV/A experienced a rapid exponential expansion between 1960–1970, while HBV/D experienced such an expansion between 1970–1990. During the period of both rapid exponential expansions, Tanzania was involved in independence wars, the Tanganyika Rifles from 1961 to 1964 and the Uganda–Tanzania war in 1978–1979. The role played by the two Tanzanian wars in the expansion of HBV in Tanzania remains to be investigated, but their coincidence with rapid expansion of HBV/A and HBV/D may be indicative of their contribution to the expansion observed. The expansion phase of HBV/D coincided with the massive expansion of HBV/E in the African HBV/E crescent, probably due to massive public health interventions, as we reported previously [23]. It is likely that the expansion of HBV/D in Tanzania and HBV/E in Nigeria [23] was shaped by related mechanisms, such as unsafe injection practices [24], although the data presented here provide no direct evidence for such an association.

Analysis of the sequences obtained in this study from samples collected within a short period of time and using short genomic regions could result in inaccurate estimation of tMRCA. The timescale approximation for the HBV epidemic may vary considerably, depending on the estimation of evolutionary rate. The rate of 2.97×10^{-4} substitutions

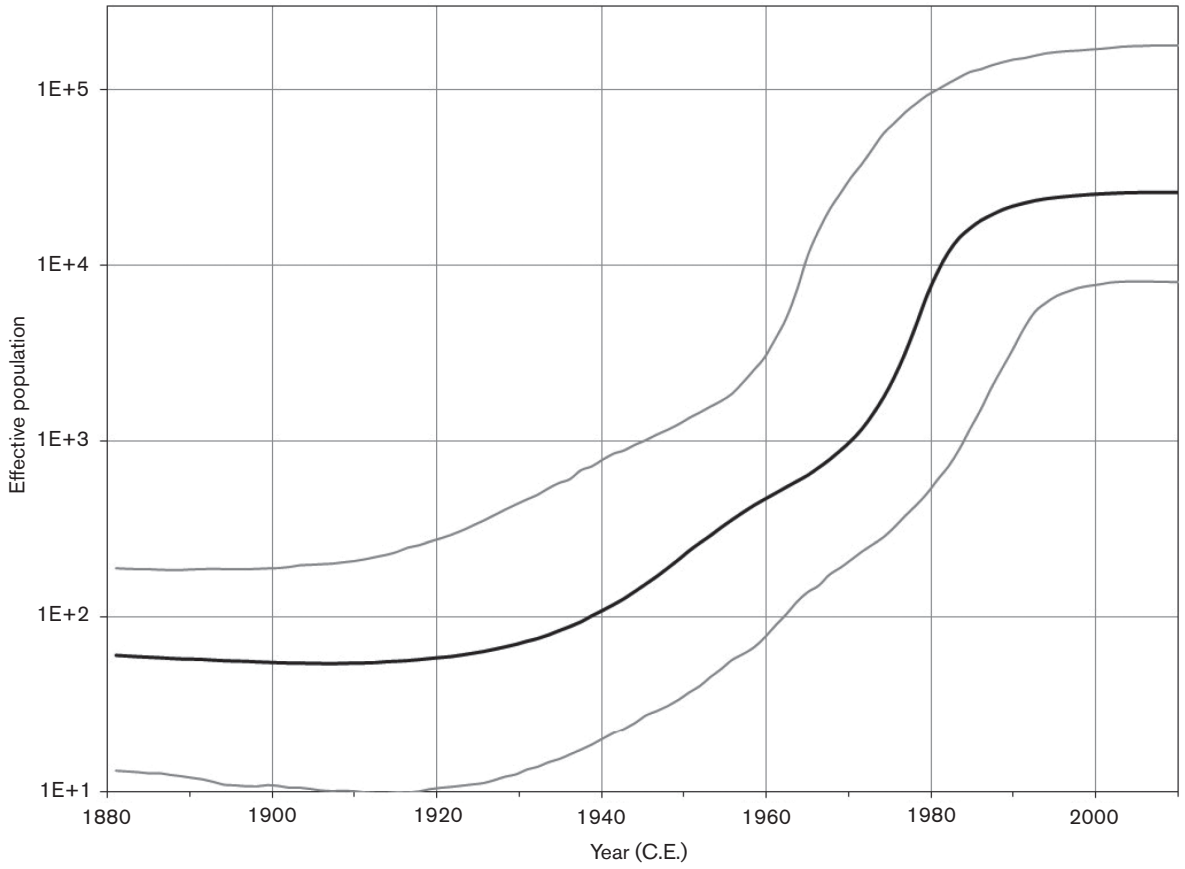

Fig. 5. Skyline plot for Tanzanian HBV genotype D sequences. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. Grey lines mark the limits of the 95 % HPD for the effective population and the solid black line is the mean of this distribution.

per site per year used in the present analysis represents the best estimate of HBV evolutionary rates [25]. Another possible limitation of the study is that quasispecies analysis of HBV strains was not performed to evaluate the extent of intra and inter-host heterogeneity.

Only 0.4 % of HBV/A1 strains and none of HBV/D and HBV/E carry known substitutions that confer immune escape properties. This finding presents an opportunity and a potential advantage to public health interventions in Tanzania. Since only a small fraction of HBV strains is potentially less preventable through vaccination programs [26], the current HBV vaccine should be effective in preventing infection. Additionally, conventional diagnostic assays for HBsAg [27–30] should be effective in detecting infections in this country. Hepatitis B vaccine was introduced in Tanzania in 2002 and has been shown to be effective [31], as further supported by the data presented in the present study. Surprisingly, G145R, a common vaccine escape substitution in the 'a' determinant [32, 33], was not detected in any of the 772 individuals in this study. Finding remarkably low levels of vaccine escape substitution for a period of ~10 years from the start of vaccination to the time these samples were collected (2011) argues well for the effectiveness of vaccination in Tanzania. These findings suggest that Tanzania is at an advantage in HBV infection control efforts in comparison to other HBV-hyperendemic countries in Africa [17].

In conclusion, the data obtained in this study suggest that Tanzania is, at least in part, the geographic origin of the HBV/A1 and a new clade of HBV/D. Although the HBV epidemic in Tanzania is characterized by a recent increase in the transmission rate and significant HBV genetic diversity, current testing and vaccination efforts should be effective as part of control strategies.

METHODS

Study population and ethics statement

In Tanzania, per the National Blood Transfusion Service testing algorithm, all donor blood units are screened for HBV with the Genedia ELISA 3.0 (Green Cross Life Sciences Corporation, Korea), a qualitative test designed for the detection of hepatitis B surface antigen (HBsAg) [34]. Samples from blood donor units screening positive by the

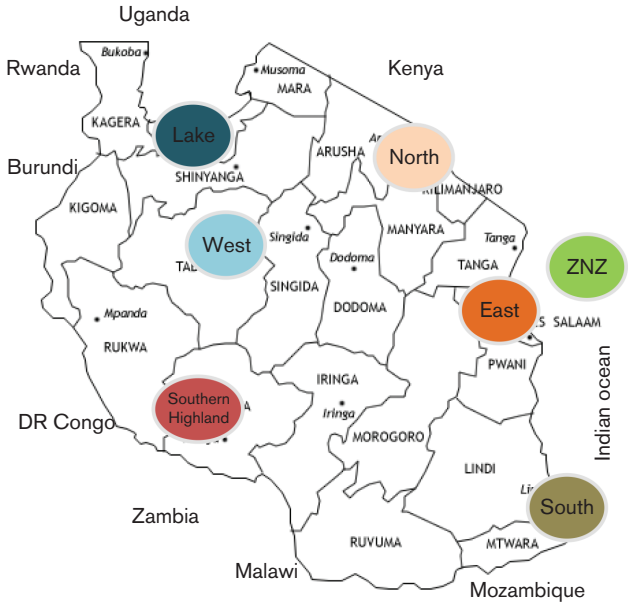

Fig. 6. Outline map of Tanzania showing the seven geographical regions from which the samples were collected, designated in circles with the names embedded.

national algorithm were sent to the Division of Viral Hepatitis Laboratory, Centers for Disease Control and Prevention (CDC), Atlanta, GA, for further testing and characterization of HBV isolates. A total of 772 distinct HBV isolates responsible for HBV infections identified among blood donor units in Tanzania collected during June–September 2011 were included in this study. The samples were collected from seven different geographical regions or zones of the country (East, Lake, North, Southern Highland, South, West and Zanzibar). The locations from which the samples originated are shown on the outline map of Tanzania in Fig. 6. Plasma samples from these blood donor units were stored at −70 °C until use. Ethical approval for this research was given by the Tanzania Ministry of Health and Social Welfare and informed consent was also obtained from all individuals who participated in the study. A non-research determination to test the anonymized samples for HBV was obtained from the CDC.

DNA extraction, amplification and sequencing

Total nucleic acid was extracted from plasma samples using the Roche MagNA Pure LC instrument and MagNA Pure LC Total Nucleic Acid Isolation kit (Roche Diagnostics Corporation, Indianapolis, IN) and eluted with 50 µl elution buffer according to the manufacturer's instructions. The HBV S-gene was amplified by two rounds of real-time PCR on the LightCycler 480 Instrument (Roche Diagnostics Corporation, Indianapolis, IN) using primer combinations and standardized protocols, as described previously [17, 23]. The PCR yielded a product size of ~403 bp. DNA sequencing was carried out in both directions using the Big Dye-Terminator version 3.1 cycling methodology (Applied Biosystems, CA) and an automated sequencer (ABI 3130xl Genetic Analyzer; Applied Biosystems, CA). The sequencing PCR involved 25 cycles, with each cycle consisting of 96 °C for 10 s, 50 °C for 5 s and 60 °C for 4 min.

Sequence alignment and phylogenetic analysis

Sequences were initially cleaned and assembled using the SeqMan and MegAlign programs from Lasergene DNA and Protein analysis software (version 10.1.2, DNASTAR, Madison, WI). All of the S-gene sequences were aligned using Geneious

Table 1. Befi-BaTS results for genotype HBV/A sequences from Tanzania

The null distribution and its confidence intervals were created in Befi-BaTS and the observed mean, confidence intervals and *P* values were calculated from the null distribution.

	Observed	Confidence interval		Null	Confidence interval		*P* value
	Mean	Lower 95 %	Upper 95 %	Mean	Lower 95 %	Upper 95 %	
AI	28.97	26.36	31.46	52.58	50.81	54.44	0
PS	252.73	239	266	404.94	397.53	411.81	0
NTI	11 992.08	10 779.64	13 282.41	13 274.86	13 169.9	13 367.58	0
NRI	2 242 084.25	1 835 544.75	2 761 449.75	2 376 957.25	2 373 950	2 380 052.5	0
PD	49 796.8	44 608.91	55 155.54	53 982.77	53 703.35	54 184.18	0
MC (Northern zone)	2.26	2	3	1.17	1.01	1.74	0.01
MC (Eastern zone)	5.43	4	8	3.34	2.84	4.11	0.01
MC (Southern zone)	6.05	4	8	2.29	2.04	2.76	0.01
MC (Western zone)	1	1	1	1.01	1	1.03	1
MC (Zanzibar zone)	3.5	3	5	1.4	1.04	2	0.01
MC (Lake zone)	19.11	11	27	4.81	4.26	5.65	0.01
MC (Highland zone)	2.2	1	4	1.39	1.05	1.99	0.02

AI, association index; PS, parsimony score; NTI, nearest taxa index; NRI, net relatedness index; PD, phylogenetic diversity; MC, maximum exclusive single-state clade size.

Table 2. Befi-BaTS results for genotype HBV/D sequences from Tanzania

	Observed	Confidence interval		Null	Confidence interval		
	Mean	Lower 95 %	Upper 95 %	Mean	Lower 95 %	Upper 95 %	*P* value
AI	3	2.28	3.73	4.32	3.65	4.97	0.01
PS	24.56	21	28	31.89	29.58	33.74	0
NTI	1 156.53	921.07	1 436.56	1 213.99	1 180.24	1 250.03	0
NRI	59 952.31	46 254.51	76 594.34	58 684.3	58 021.18	59 355.39	0.6
PD	3 780.02	3 044.43	4 622.78	3 934.54	3 883.73	3 990.96	0
MC (Lake zone)	16.44	9	23	6.74	5.08	9.76	0.01
MC (Eastern zone)	4.19	3	5	1.95	1.43	2.61	0.01
MC (Highland zone)	1	1	1	1.01	1	1.03	1
MC (Northern zone)	1	1	1	1	1	1	1
MC (Western zone)	1	1	1	1	1	1	1

AI, association index; PS, parsimony score; NTI, nearest taxa index; NRI, net relatedness index; PD, phylogenetic diversity; MC, maximum exclusive single-state clade size.

Pro version 5.5.8 (Biomatters, New Zealand) and also MUSCLE, as implemented in MEGA5 [35]. Phylogenetic analysis was performed by using the maximum likelihood method of the MEGA5 software [35]. Genotypes and subtypes were classified by comparing sequences obtained in this study with representative sequences of HBV/A–HBV/H as well as GenBank sequences representing subtypes HBV/A1–3 and subtypes HBV/D1–9. The S-gene sequences for comparative phylogenetic analysis were obtained from the HBV whole-genome sequences [14]. The nucleotide diversity (mean pairwise genetic distance) of each group of identified genotypes was calculated separately using maximum-likelihood correction and gamma-distributed rate variation among the sites. These analyses were performed in MEGA5 [35].

Bayesian posterior sample tree

The 772 nucleotide sequences were segregated by genotypes. All genotype A and D sequences were selected for analysis. The HBV/E isolates were not subjected to Bayesian analysis due to the small sample size. BEAST (version 1.7.5) [36] was used to create 10 000 Bayesian trees using the Hasegawa–Kishino–Yan (HKY) substitution model with four gamma-rate categories and invariant sites with an uncorrelated lognormal molecular clock and a coalescent constant population prior [37]. A fixed substitution rate of 2.97×10^{-4} mutations per site per year was used [25]. Using a strict as well as a relaxed clock for the initial estimate of the substitution rate, none of the models tested was found to be superior, indicating that the conclusions are not affected by the application of different models [23].

Phylogeographic association using Befi-BaTS

To analyse whether the phylogenetic trait of the region of origin for the sequences generated in this study was randomly associated with the tips of the phylogenetic tree, Befi-BaTS analysis was employed. Befi-BaTS uses a posterior sample of trees to approximate the true posterior distribution of phylogenies given the sequences. Randomization of the taxon–geographic location associations allows for the creation of a null distribution, providing a statistical significance test of the null hypothesis that geographic locations are randomly associated with phylogeny tips. All the methods in Befi-BaTS examine the tree as a whole, except for the monophyletic clade size, which can be used to examine geographic locations individually. After posterior samples were created for each genotype, an in-house Perl script was used to remove a burn-in of 75 % for the HBV/D sequences and 80 % for the genotype A sequences. There was a larger burn-in for the HBV/A sequences because of the size of their tree file. The resulting tree files were reformatted and analysed with Befi-BaTS (version 0.1.1) [38] using 100 repetitions to create the null distribution. The null distribution and its confidence intervals were created in Befi-BaTS and the observed mean, confidence intervals and *P* values were calculated from the null distribution: AI, association index [39]; PS, parsimony score [40]; NTI, nearest taxa index [41]; NRI, net relatedness index [41]; PD, phylogenetic diversity [42] and MC, monophyletic clade size (location) [38]. Significance is reached if $P\leq0.05$.

Dated phylogenies

Dated phylogenies were obtained using BEAST (version 1.7.5) [37] to estimate the tMRCA and to create skyline plots for each genotype independently [43]. The HKY substitution model was used with four gamma-rate categories and invariant sites. The uncorrelated lognormal molecular clock with a substitution rate of 2.97×10^{-4} substitutions per site per year calculated by Zhou and Holmes [44] was used for these calculations. A piecewise constant coalescent Bayesian skyline prior was used with 10 groups. The burn-in was 10 % for both genotypes. The population parameter returned by the skyline plot was $N_e\tau$, where N_e is the effective population and τ is the generation length in years. Because τ is constant, the use of $N_e\tau$ allows us to evaluate the relative changes in the effective population of HBV in this setting.

Skyline subset analysis

A Perl script was used to create random sequence data sets with 10, 30 or 50 % of the sequences from the complete HBV/A dataset. A minimum of five sets were created for each percentage. Bayesian skyline plots were created for each data set using the method described above. The estimates for the root height were averaged for each percentage and the significance of these means was determined using a *t*-test for two means while assuming unequal variances.

Detection of mutations within the 'a' determinant of the HBsAg sequences

Specific known mutations associated with immune escape located within the HBsAg immunodorminant 'a' determinant region (residues 124–147; I/T126A/N, A128V, Q129H/R, G130N, M133L/T, K141E, D144A/H and G145R) [33, 45] were searched for manually. The first and most common 'vaccine escape mutant' described [33] is the substitution of Gly at position 145 by Arg (G145R).

Nucleotide sequence accession numbers

Nucleotide sequences for the HBV strains have been deposited in GenBank under accession numbers KU594654–KU595425.

Funding information
Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

Acknowledgements
Disclaimer: this information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry. It does not represent and should not be construed to represent any agency determination or policy.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
Ethical approval for this research was given by the Tanzania Ministry of Health and Social Welfare and informed consent was also obtained from all individuals who participated in the study. A non-research determination to test the anonymized samples for HBV was obtained from the CDC.

References

1. **Lavanchy D.** Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *J Viral Hepat* 2004;11:97–107.
2. **Goldstein ST, Zhou F, Hadler SC, Bell BP, Mast EE *et al.*** A mathematical model to estimate global hepatitis B disease burden and vaccination impact. *Int J Epidemiol* 2005;34:1329–1339.
3. **WHO.** 2012. *Hepatitis B Geneva*. Switzerland: WHO. www.who.int/mediacentre/factsheets/fs204/en/index.html. [updated July 2012. Fact sheet N°204].
4. **Puoti M, Manno D, Nasta P, Carosi G.** Hepatitis B virus and HIV coinfection in low-income countries: unmet needs. *Clin Infect Dis* 2008;46:367–369.
5. **Nagu TJ, Bakari M, Matee M.** Hepatitis A, B and C viral co-infections among HIV-infected adults presenting for care and treatment at Muhimbili national hospital in Dar es Salaam, Tanzania. *BMC Public Health* 2008;8:416.
6. **Johnston LG, Holman A, Dahoma M, Miller LA, Kim E *et al.*** HIV risk and the overlap of injecting drug use and high-risk sexual behaviours among men who have sex with men in Zanzibar (Unguja), Tanzania. *Int J Drug Policy* 2010;21:485–492.
7. **van Houdt R, Bruisten SM, Speksnijder AG, Prins M.** Unexpectedly high proportion of drug users and men having sex with men who develop chronic hepatitis B infection. *J Hepatol* 2012;57:529–533.
8. **Palumbo E.** Hepatitis B genotypes and response to antiviral therapy: a review. *Am J Ther* 2007;14:306–309.
9. **Ganem D, Schneider RJ.** Hepadnaviridae: the viruses and their replication. In: Knipe DM and Howley PM (editors). *Fields Virology*. Philadelphia: Lippincott Williams & Wilkins; 2001. pp. 2923–2969.
10. **Lin CL, Liu CH, Chen W, Huang WL, Chen PJ *et al.*** Association of pre-S deletion mutant of hepatitis B virus with risk of hepatocellular carcinoma. *J Gastroenterol Hepatol* 2007;22:1098–1103.
11. **Cento V, Mirabelli C, Dimonte S, Salpini R, Han Y *et al.*** Overlapping structure of hepatitis B virus (HBV) genome and immune selection pressure are critical forces modulating HBV evolution. *J Gen Virol* 2013;94:143–149.
12. **Mahtab MA, Rahman S, Khan M, Karim F.** Hepatitis B virus genotypes: an overview. *Hepatobiliary Pancreat Dis Int* 2008;7:457–464.
13. **Ganova-Raeva L, Ramachandran S, Honisch C, Forbi JC, Zhai X *et al.*** Robust hepatitis B virus genotyping by mass spectrometry. *J Clin Microbiol* 2010;48:4161–4168.
14. **Lago BV, Mello FC, Kramvis A, Niel C, Gomes SA.** Hepatitis B virus subgenotype A1: evolutionary relationships between Brazilian, African and Asian isolates. *PLoS One* 2014;9:e105317.
15. **Hasegawa I, Tanaka Y, Kurbanov F, Yoshihara N, El-Gohary A *et al.*** Molecular epidemiology of hepatitis B virus in the United Republic of Tanzania. *J Med Virol* 2006;78:1035–1042.
16. **Kramvis A, Kew MC.** Molecular characterization of subgenotype A1 (subgroup Aa) of hepatitis B virus. *Hepatol Res* 2007;37:S27–S32.
17. **Forbi JC, Ben-Ayed Y, Xia GL, Vaughan G, Drobeniuc J *et al.*** Disparate distribution of hepatitis B virus genotypes in four sub-Saharan African countries. *J Clin Virol* 2013;58:59–66.
18. **Kwange SO, Budambula NL, Kiptoo MK, Okoth F, Ochwoto M *et al.*** Hepatitis B virus subgenotype A1, occurrence of subgenotype D4, and S gene mutations among voluntary blood donors in Kenya. *Virus Genes* 2013;47:448–455.
19. **Hübschen JM, Mugabo J, Peltier CA, Karasi JC, Sausy A *et al.*** Exceptional genetic variability of hepatitis B virus indicates that Rwanda is east of an emerging African genotype E/A1 divide. *J Med Virol* 2009;81:435–440.
20. **Ramachandran S, Purdy MA, Xia GL, Campo DS, Dimitrova ZE *et al.*** Recent population expansions of hepatitis B virus in the United States. *J Virol* 2014;88:13971–13980.
21. **Hannoun C, Söderström A, Norkrans G, Lindh M.** Phylogeny of African complete genomes reveals a West African genotype A subtype of hepatitis B virus and relatedness between Somali and Asian A1 sequences. *J Gen Virol* 2005;86:2163–2167.
22. **Andernach IE, Hübschen JM, Muller CP.** Hepatitis B virus: the genotype E puzzle. *Rev Med Virol* 2009;19:231–240.
23. **Forbi JC, Vaughan G, Purdy MA, Campo DS, Xia GL *et al.*** Epidemic history and evolutionary dynamics of hepatitis B virus infection in two remote communities in rural Nigeria. *PLoS One* 2010;5:e11615.
24. **Zehender G, Ebranati E, Gabanelli E, Shkjezi R, Lai A *et al.*** Spatial and temporal dynamics of hepatitis B virus D genotype in Europe and the Mediterranean Basin. *PLoS One* 2012;7:e37198.
25. **Zhou Y, Holmes EC.** Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J Mol Evol* 2007;65:197–205.
26. **Zuckerman AJ.** Effect of hepatitis B virus mutants on efficacy of vaccination. *Lancet* 2000;355:1382–1384.
27. **Grethe S, Monazahian M, Böhme I, Thomssen R.** Characterization of unusual escape variants of hepatitis B virus isolated from a hepatitis B surface antigen-negative subject. *J Virol* 1998;72:7692–7696.

28. **Carman WF, Korula J, Wallace L, Macphee R, Mimms L *et al.*** Fulminant reactivation of hepatitis B due to envelope protein mutant that escaped detection by monoclonal HBsAg ELISA. *Lancet* 1995; 345:1406–1407.
29. **Karthigesu VD, Allison LM, Fortuin M, Mendy M, Whittle HC *et al.*** A novel hepatitis B virus variant in the sera of immunized children. *J Gen Virol* 1994;75:443–448.
30. **Purdy MA.** Hepatitis B virus S gene escape mutants. *Asian J Transfus Sci* 2007;1:62–70.
31. **Metodi J, Aboud S, Mpembeni R, Munubhi E.** Immunity to hepatitis B vaccine in Tanzanian under-5 children. *Ann Trop Paediatr* 2010; 30:129–136.
32. **Salisse J, Sureau C.** A function essential to viral entry underlies the hepatitis B virus "a" determinant. *J Virol* 2009;83:9321–9328.
33. **Carman WF, Zanetti AR, Karayiannis P, Waters J, Manzillo G *et al.*** Vaccine-induced escape mutant of hepatitis B virus. *Lancet* 1990; 336:325–329.
34. **Ej O, Jang HS, Lee HI, Park YJ, Kim BK.** Evaluation of Genedia ELISA kit for hepatitis B (Anti-HBs, HBeAg/Anti-Hbe). *J Clin Pathol Qual Control* 1999;21:371–377.
35. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M *et al.*** MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;28:2731–2739.
36. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012;29: 1969–1973.
37. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012;29: 1969–1973.
38. **Parker J, Rambaut A, Pybus OG.** Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* 2008;8:239–246.
39. **Wang TH, Donaldson YK, Brettle RP, Bell JE, Simmonds P.** Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J Virol* 2001;75:11686–11699.
40. **Slatkin M, Maddison WP.** A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* 1989;123:603–613.
41. **Webb CO.** Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am Nat* 2000;156:145–155.
42. **Faith DP.** Conservation evaluation and phylogenetic diversity. *Biol Conserv* 1992;61:1–10.
43. **Drummond AJ, Rambaut A, Shapiro B, Pybus OG.** Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005;22:1185–1192.
44. **Zhou Y, Holmes EC.** Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J Mol Evol* 2007;65:197–205.
45. **Cooreman MP, Leroux-Roels G, Paulij WP.** Vaccine- and hepatitis B immune globulin-induced escape mutations of hepatitis B virus surface antigen. *J Biomed Sci* 2001;8:237–247.