Communications in Information and Systems
Volume 19, Number 4, 357–373, 2019

Genetic continuity in the last seven Millennia in human hepatitis B viruses

Xiaoyun Lei[*], Ye Zhang[*], and Shi Huang[†]

Hepatitis B virus (HBV) is a major human pathogen and yet the evolution history of HBV has largely remained uncertain. With a better theoretical understanding of genetic diversity, we here used a new method to examine the previously published ancient and present day HBV genomes. We identified an informative region in the HBV polymerase that is slow evolving and used it to study genetic distances among HBVs. Three ancient human HBV isolates from 4488–7074 years ago in Germany were identified as genotype G that is also presently common in the same country. We constructed a new phylogenetic tree of HBVs that placed genotype D as the most basal branch with an inferred age of ~20500 years, which is remarkably consistent with the worldwide distribution and a most parsimonious migration route of HBV genotypes today. These results help resolve the evolutionary history of HBV and provide a useful method for studying the phylogenetics of HBV and other viruses in general.

Keywords and phrases: Ancient DNA, hepatitis B virus HBV, slow clock.

1. Introduction

Hepatitis B virus (HBV) is a major cause of human hepatitis and related diseases (http://www.who.int/mediacentre/factsheets/fs204/en/). The origin and evolution of HBV has largely remained uncertain, like most viruses. HBV has a circular, partially double-stranded DNA genome of about 3.2kbp that encodes four overlapping open reading frames (P polymerase, pre-S/S envelope, pre-C/C core protein, and X). At least 8 genotypes (A–H) based on nucleotide sequence similarity are classified for human HBV and they have a heterogeneous global distribution (Castelhano et al. 2017). The putative basal genotypes F and H are found exclusively in the Americas, thus

[*] The first two authors contributed equally to this work.
[†] Corresponding author.

inconsistent with the notion that HBV co-evolved with modern humans as part of the Recent Out of Africa hypothesis. Yet, HBVs in non-human primates (NHP), such as chimpanzees and gorillas, are phylogenetically related to human HBV isolates, seemingly supporting the idea of an Africa origin of the virus (Locarnini et al. 2013; Souza et al. 2014).

Recently, a number of ancient HBV genomes have been uncovered from human skeletons found in Europe and Asia that are between approximately 500–7000 years ago (Kahila Bar-Gal et al. 2012; Krause-Kyora et al. 2018; Muhlemann et al. 2018). While most of the relatively younger HBV genomes (<4488 years ago) were closest to present day human HBVs, all three oldest HBV samples found in Germany (between 4488–7074 years ago) were unexpectedly closest to chimpanzee or gorilla HBVs and hence considered extinct today. The finding challenges expectations as HBV today must have an ancient ancestor which must have infected a large population in the past to have a chance to survive to the present. As large populations have a greater probability of having some of its remains discovered today, the probability of discovering an ancestor of today's human HBVs should be much greater than that of finding a now extinct ancient human HBV sample. Thus, the unusually high rate of discovering ancient human HBV samples that are now extinct (3 independent samples in 3 different archaeological sites) indicates potential flaws in the phylogenetic method employed, especially given that existing methods have yet to produce a consistent evolutionary history of the HBVs. Importantly, the theoretical framework underlying the existing methods, the neutral theory, has been widely known to be inadequate as an explanatory theory of the observed genetic diversity patterns (Kreitman 1996; Ohta, Gillespie 1996; Hahn 2008; Leffler et al. 2012; Hu et al. 2013; Kern, Hahn 2018). It is unfortunate that existing phylogenetic methods have relied heavily on the neutral theory being a valid interpretation of nature.

Different positions in a viral genome are known to have different mutation rates. The fast changing sites in influenza virus play adaptive roles in escaping host immune defense and undergo constant and quick turnovers (Shih et al. 2007). The antigenic sites in human influenza A virus mutate and turn over quickly (several times within a 30 year period), which is critical for the virus to escape host immune defense and hence for flu epidemics. In contrast, other sites stayed largely unchanged within the same period. The influenza results illustrate two general points with regard to evolutionary dynamics of a genome that have so far been overlooked. First, fast evolving or less conserved DNAs are also functional rather than neutral as they are essential for quick adaptive needs in response to fast changing environments. Second, fast evolving DNAs turn over quickly and can be shown to violate the infinite sites model. Hence, they cannot be used for phylogenetic inference involving evolutionary timescales. If one uses the fast changing sites in a flu virus to infer the phylogenetic relationship of the virus isolates responsible for different epidemics in a past period of say 10 years, one would have reached the erroneous conclusion that each epidemic was caused by a distinct type of flu virus with no genetic continuity among them rather than just minor variations of the same type.

For short term lineage divergence that has yet to reach saturation for the fast changing sites, both fast and slow changing sequences could be informative to phylogeny. However, for evolutionary timescale where divergence in fast changing sites have reached saturation, only the slow sites (the slow clock method) could be informative, as has been previously shown and explained by the maximum genetic diversity (MGD) hypothesis (Huang 2012; 2016; Hu et al. 2013; Yuan, Huang 2017; Yuan et al. 2017). The MGD hypothesis has recently solved the longstanding puzzle of genetic diversity (Huang 2009; 2016) and made it now possible for the first time to realistically infer phylogenetic relationships based on genetic diversity data. It has now been demonstrated that genetic diversities are mostly at saturation level (Yuan et al. 2012; 2014; Zhu et al. 2015a; 2015b; Gui, Lei, Huang 2017; He et al. 2017; Lei, Huang 2017; Lei et al. 2018; Teske et al. 2018), which therefore renders most of the past molecular results invalid since those results were based on mistreating saturated phases of genetic distance as linear phases. Only slow evolving nuclear sequences are still at linear phase and hence informative to phylogenetic relationships.

Here, we investigated the genetic relationships among the ancient HBVs and present day human HBVs using the slow clock method. We found that all three ancient HBV samples that were thought to group with NHP isolates in fact grouped with human HBVs. We also constructed a new phylogenetic tree of the human HBV genotypes, which is remarkably consistent with their distribution patterns.

2. Results

2.1. Identity analyses in nucleotide and amino acid sequences

We selected 3 ancient HBV genomes from Germany for analyses here, 7074 year old Karsdorf from LBK culture in Lower-Saxony, 5353 year old Sorsum from Funnelbeaker culture from Lower-Saxony, and 4488 year old RISE563 from Bell Beaker culture in Osterhofen-Altenmarkt (Krause-Kyora et al. 2018; Muhlemann et al. 2018). The other Bronze age samples were not

Table 1: Nucleotide identity among the three ancient HBV genomes

Samples	Nucleotide identity*		
	Karsdorf	Sorsum	RISE563
Karsdorf		94.8%	93.0%
Sorsum			95.1%
RISE563			

*Calculated by multiplying the number of matches in the pair of aligned sequences by 100 and dividing by the length of the aligned region, not including gaps.

studied as one (RISE254) was very close to RISE563 and the others had many sequence gaps. In nucleotide identity, the three ancient samples were all closer to each other than to any present day samples, and the highest identity was between Sorsum and RISE563 (Table 1).

We searched the Genbank protein database to identify the closest present day HBV genome to the ancient HBVs in amino acid identity in the polymerase, the largest open reading frame in the HBV genome (832–845 amino acids in length). Upon identifying the closest, we also examined its identity to the ancient HBV in other proteins, pre S, X, and core proteins (Table 2). Present day HBV isolates closest to the ancient HBVs in nucleotide identity were not found to be the closest in amino acid identity in the polymerase. For example, the Karsdorf sample was closest to a chimpanzee HBV (accession AB032433) in nucleotide sequence but a human HBV (HE981175, genotype G) in amino acid sequence in the polymerase. However, in other proteins, the closest to the ancient HBVs were all present day NHP HBVs (Table 2).

The 3 ancient HBV genomes were also closest to each other in polymerase amino acid identity than to any other present day samples (Table 2). However, different from the nucleotide result, the highest amino acid identity in polymerase was between Karsdorf and Sorsum. As Sorsum differs from RISE563 in both time periods and locations while only in time periods from Karsdorf, Sorsum is expected to be a closer relative of Karsdorf and hence to have more similarity in slow changing sites (amino acid) with Karsdorf. On the other hand, as Sorsum was 1721 years apart from Karsdorf, ~2 fold more than its time difference with RISE563 (865 years), Sorsum is expected to have more genetic distance from Karsdorf due to fast changing sites as may be reflected in the nucleotide sequence. Together, these results showed significant disconnect between amino acid and nucleotide sequences in revealing genetic relationships among HBVs, which is likely due to the generally faster evolutionary rates for nucleotide sequences relative to protein sequences.

Table 2: Amino acid identities in the polymerase among ancient and present day HBV genomes

Proteins	Species	Karsdorf (7074)		Sorsum (5353)		RISE563 (4488)	
		Access.	Identity	Access.	Identity	Access.	Identity
Polymerase	Human	HE981175	92.9%	EU239218	92.5%	EU239218	90.8%
	Chimp.	AB032433	92.6%	AB032433	92.5%	AB032433	90.9%
	Gorilla	AJ131567	90.6%	AJ131567	90.6%	AJ131567	89.4%
	Orangutan	AF193863	88.2%	AF193863	88.3%	AF193863	87.4%
	Gibbon	AJ131571	91.3%	AJ131571	89.8%	AJ131571	88.7%
	Karsdorf	Karsdorf		Karsdorf	95.5%	Karsdorf	93.5%
	Sorsum	Sorsum	95.5%	Sorsum		Sorsum	94.8%
	RISE563	LT992443	93.5%	LT992443	94.8%	LT992443	
Pre S	Human	HE981175	91.1%	EU239218	92.3%	EU239218	91.7%
	Ape	AB032433	92.6%	AB032433	95.6%	AJ131567	91.5%
X prot.	Human	HE981175	83.1%	EU239218	87.4%	EU239218	85.3%
	Ape	AB032433	84.4%	AB032433	90.9%	AJ131567	91.7%
Core prot.	Human	HE981175	94.2%	EU239218	91.7%	EU239218	92.6%
	Ape	AB032433	99.3%	AJ131567	96.1%	AJ131567	98.7%

Note: Approximate age in years of ancient HBV samples is indicated next to sample name. The closest human HBV to the ancient HBV in polymerase identity was selected for comparison. Gaps were excluded in the identity calculation. The underlined accession numbers indicate the closest modern HBV to the ancient HBV in nucleotide identity as previously published.

While results in Table 2 showed clear affinity of Karsdorf with human HBV, Sorsum showed equal affinity with human and chimpanzee HBVs and RISE563 showed slightly more affinity to chimpanzee than to human HBV, indicating some uncertainty regarding the informative nature of the full length polymerase protein. The polymerase is composed of 4 domains, terminal protein, non-conserved spacer, reverse transcriptase, and RNase H. Upon examining the HBVdb database (https://hbvdb.ibcp.fr/) (Hayer et al. 2013), together with our own alignment analyses, we found that the amino acid region corresponding to the reverse transcriptase and RNase H domains are more conserved or slow evolving (343–844 aa for genotype G starting with VNL). The 1–342 aa region shows ~83% identity between human and chimpanzee, significantly lower than the ~93% in the 343–844 aa region. We therefore tested this 501 aa region to see if it may show better results than the full length polymerase in linking ancient HBVs with human rather than NHP (Table 3). Again, all three ancient HBVs showed closer identity, but to a greater degree, to human HBVs than to NHP HBVs. While Karsdorf was again closest to Sorsum, it was closer to a present day human HBV (HE981175) than to the ancient HBV RISE563, indicating clear HBV genetic continuity from the time of Karsdorf to present time and the more

Table 3: Amino acid identity among ancient and present day HBV genomes in the slow region (501 aa)

Species	Karsdorf (7074)		Sorsum (5353)		RISE563 (4488)	
	Access.	Identity	Access.	Identity	Access.	Identity
Human	HE981175	96.0%	HE981175	95.4%	EU239218	93.9%
Chimp.	AB032433	93.8%	AB032433	95.2%	AB032433	93.5%
Gorilla	AJ131567	93.8%	AJ131567	95.2%	AJ131567	93.7%
Orangutan	AF193863	92.0%	AF193863	93.2%	AF193863	92.1%
Gibbon	AJ131571	93.2%	AJ131571	94.0%	AJ131571	92.2%
Karsdorf	Karsdorf		Karsdorf	97.4%	Karsdorf	94.6%
Sorsum	Sorsum	97.4%	Sorsum		Sorsum	96.2%
RISE563	LT992443	94.6%	LT992443	96.2%	LT992443	

HBV genomes selected for comparison were the same as those in Table 2. The underlined accession numbers indicate the closest modern HBV to the ancient HBV in nucleotide identity as previously published.

informative nature of the 501 aa slow region of the polymerase.

As slow evolving DNAs are more likely to be in linear phase and hence more informative to phylogenetic relationships as explained by the MGD theory, we examined whether the 501 aa slow region of the polymerase is the slowest evolving among the protein genes in the HBV genome by comparing amino acid identity between human and orangutan HBV proteins (Supplementary Table S1 http://intlpress.com/site/pub/files/_supp/cis/2019/0019/0004/CIS-2019-0019-0004-s001.zip). The 501 aa slow region of the polymerase was found to be the second most conserved, just slightly less conserved than the core protein. However, because the core protein was relatively short (178 aa), it is expected to be less informative than a longer protein with similar degree of conservation. Together with outgroup analyses (see below), we have found the 501 aa slow region of the polymerase to be the most informative to phylogenetic inferences of HBV strains.

2.2. Outgroup inferences based on amino acid mutations

The above results raise the important question of which type of sequences may be most informative to HBV phylogeny. For viruses, different hosts may confer different physiological selection pressures which may result in viruses from different hosts to have drastic or non-conservative amino acid changes. Taking into account of non-conservative changes may thus be informative to phylogenetic relationships where an outgroup NHP HBV to two sister strains of human HBV is expected to show more non-conservative amino acid changes from the human HBVs.

Table 4: Outgroup inferences from non-conservative (drastic) amino acid changes

Outgroup		Polymerase				Slow region (aa 343–844)			
		Mutations		Fraction	P value	Mutations		Fraction	P value
Isolates	Accession	Drastic	All	Drastic	Hu. v Ape	Drastic	All	Drastic	Hu. v Ape
Karsdorf	NA	5	16	0.31	0.02	5	10	0.50	n.s.
Human	HE981175	9	24	0.38	0.02	4	9	0.44	n.s.
Chimpanzee	AB032433	16	22	0.73		11	15	0.73	
Sorsum	NA	7	16	0.44	n.s.	2	6	0.33	n.s.
Human	EU239218	18	39	0.46	n.s.	5	18	0.28	n.s.
Chimpanzee	AJ131575	26	39	0.67		4	7	0.57	
RISE563	LT992443	8	24	0.33	0.004	2	6	0.44	n.s.
Human	EU239218	16	32	0.50	n.s.	2	13	0.24	n.s.
Gorilla	AJ131567	27	39	0.69		12	16	0.75	
Sum									
Ancient human HBV		20	56	0.36	<0.0001	9	22	0.41	0.03
Present human HBV		43	95	0.45	0.0003	11	40	0.28	0.0002
Ape HBV		69	100	0.69		27	38	0.71	

Drastic or non-conservative and conservative changes were according to the designation by the blastp algorithm. Numbers of these changes can be found in the Supplementary Table S2. P value was from Fisher's test.

To confirm the human rather than NHP affinity of the ancient HBV isolates as shown by the polymerase, we therefore performed protein alignment involving 3 strains, an ancient HBV, its closest human HBV, and its closest NHP HBV. It is expected that an outgroup should have a higher fraction of non-conservative or drastic amino acid changes among all mutations that led to differences between the outgroup and the other two sister strains. We examined those positions where the two sisters had the same residue while the outgroup was different. We tested each of the three compared HBV viruses as the candidate outgroup and obtained the fraction of drastic changes among all positions where the two sisters were the same while the outgroup was different (Table 4 and Supplementary Materials for details of this analysis). For the full length polymerase protein, RISE563 as the outgroup had a significantly smaller fraction of drastic changes than the gorilla HBV as the outgroup (0.33 vs 0.69, P = 0.004). This indicates that RISE563 and present day human HBV (EU239218) were sister strains while the gorilla HBV was the outgroup. Similar analyses showed that for Karsdorf and Sorsum samples, the NHP HBVs all showed the highest fraction of drastic changes (Table 4). We also performed the combined analysis where we first add up all the drastic changes of an outgroup (with the outgroup being ancient HBV, present day human HBV, or NHP HBV) and then calculated the fraction of drastic changes. The fraction of drastic changes in the NHP HBV when tested as the outgroup was significantly higher than either that in the ancient HBVs or the present day human HBVs when they were tested as the outgroup. We also obtained similar results for the slow region of the polymerase (Table 4). These results confirmed that ancient HBVs isolates grouped with human rather than NHP HBVs.

In contrast, for the other three smaller size proteins of HBV genome, the Pre S protein, the X protein and the pre core protein, none was found informative in identifying an outgroup (Supplementary Table S2). When we did the same analysis by using these three proteins as concatenated single molecule, we also failed to identify any clear outgroup. The fractions of drastic changes in either ancient or present day human HBVs were similar to that of NHP and showed enrichment of non-conservative amino acid changes, which was unlike the case for the polymerase. Thus, the ancient HBVs did not group with either present day NHP or human in any of these proteins. That the observed changes were enriched for non-conservative amino acid mutations indicated functional adaptation or selection. Although ancient HBVs all showed slightly closer identities in these proteins to NHP HBVs, such weak affinity may be fortuitous.

2.3. Phylogenetic relationships among HBV genotypes

The above results suggest that the slow region of the polymerase (aa 343–844) may be the most informative with regard to HBV phylogenetic relationships. We next used this region to reconstruct the phylogenetic tree for the 8 HBV genotypes by using the reference genomes for these genotypes (2 genomes for each genotype) as indexed by the HBVdb database. We first obtained the pairwise identities in the slow region among the 8 genotypes or 16 genomes and the average identity of each genotype to the other 7 genotypes (Table 5 and Supplementary Table S3). We also determined the lowest pairwise identity within each genotype by searching the Genbank database using the reference genomes and found D to have the lowest within genotype identity (461 aa), indicating that D has the largest within genotype genetic diversity among all genotypes and hence qualifies as the most basal genotype. We constructed a schematic diagram of the phylogenetic tree that best fits the data in Table 5 (Fig. 1A). Three pairs of closest genotypes were identified as GE, BC, and FH. Although B was slightly closer to G than to C, its average identity to GE was lower than to C. GE was grouped with A rather than with BC as A was closer to GE than C was, assuming convergent evolution may underlie the similarity between B and GE. FH was closer to

Table 5: Average pairwise identities in the slow region of the HBV polymerase (501 aa)

Genotype	Identical number of amino acids									
	A	B	C	D	E	F	G	H	All[1]	Lowest[2]
A		460.75	455.75	458.25	460.5	454	467	451.5	458.25	468
B	460.75		467.25	461.5	463	463.5	468.5	461.5	463.71	>467
C	455.75	467.25		458.25	461.5	455.5	465.5	454	459.68	>467
D	458.25	461.5	458.25		460.75	459	461.5	458	459.61	461
E	460.5	463	461.5	460.75		460	471.25	453.75	461.54	>471
F	454	463.5	455.5	459	460		461.25	474	461.04	>474
G	467	468.5	465.5	461.5	471.25	461.25		460.75	465.11	482
H	451.5	461.5	454	458	453.75	474	460.75		459.07	488

[1]Average identities of a genotype to all other genotypes. [2]Lowest pairwise identities within the genotype.

B while more distant to A, C, E, and G than D was, which makes it difficult to assign the basal branch based on identity alone. However, D had much greater within genotype diversity than F or H and hence was better qualified as the basal branch. Standard UPGMA (unweighted pair group method with arithmetic mean) software cannot take the greater genetic diversity of D into account and did not place D as the basal branch.

Relative to the existing tree (Fig. 1B), the new tree is more consistent with the present day distribution pattern of the HBV genotypes (Fig. 1C). In particular, consistent with expectations, the basal type D is the most widely distributed and most common in Northeast Asia and Russia/Siberia, and well positioned to split the first branch, genotype F and H, specific to the New World.

Using the slow region of the polymerase, we also examined the relationships of selected NHP HBVs with the human reference genomes and found closer relative affinity of African NHP HBVs with genotype G and of Asian NHP HBVs with genotype C (Table 6). Although we only looked at one HBV genome each for each NHP species, the closest chimpanzee and gorilla in nucleotide sequence to the ancient human HBVs and an arbitrarily chosen orangutan and gibbon HBV, the relative affinity should hold for other NHP HBVs. We also examined the identity of ancient HBVs with the human HBV reference genomes and found all three to be most related to genotype G, with Sorsum and RISE563 relatively closer to genotype E than Karsdorf. The results suggest that Sorsum and RISE563 may be on their way diverging from genotype G to E. Based on the amino acid difference between Karsdorf and genotype G ($501 - 479 = 22$ amino acid) and the age of Karsdorf, we inferred the mutation rate in the slow region

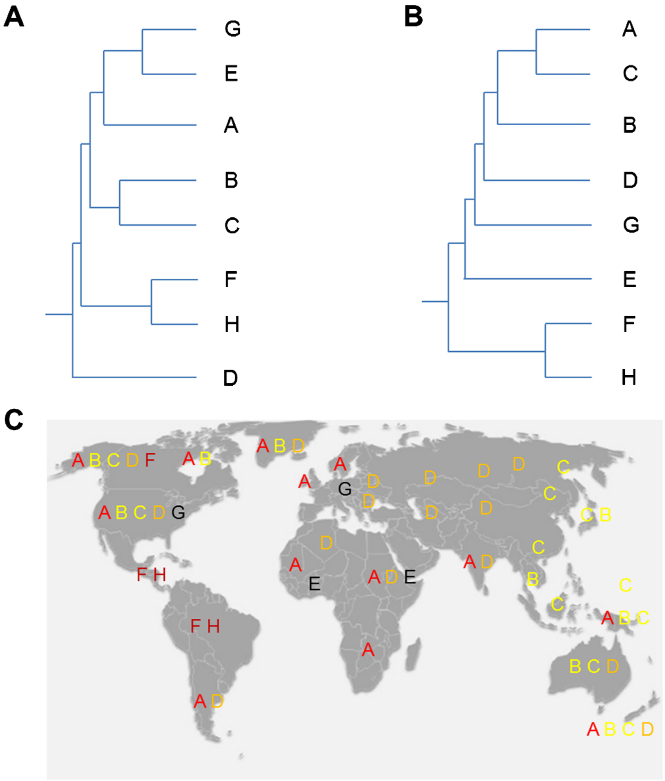

Figure 1: Phylogenetic tree of HBV genotypes. Branch lengths are relatively and roughly to scale. The tree is meant to be more of a schematic diagram. A. Tree built by using the slow region. B. Tree built by using nucleotide sequences as found in Muhlemann et al. (Muhlemann et al. 2018). C. Distribution map of HBV genotypes.

to be 2.0×10^{-6} aa per aa per year and the age of G at ~14500 years ($14500 \times 501 \times 2.0 \times 10^{-6} + 7500 \times 501 \times 2.0 \times 10^{-6} = 22$ aa). Given the distance between the basal genotype D and the other genotypes to be ~41 as well as the largest within genotype distance of 40 for D, we inferred an origination of human HBV at ~20500 years ago followed soon by the divergence of genotype F and H as people crossed the Bering Strait into the New World.

3. Discussion

Our results suggest genetic continuity in the last seven Millennia for human HBV and establish a more informative method for building phylogenetic

Table 6: Identities between NHP and ancient HBV isolates with the reference genomes of human HBV genotypes

	Identical number of amino acids						
Genotype	AB032433	AJ131567	AF193863	AJ131571			
	Chimp.	Gorilla	Orangutan	Gibbon	Karsdorf	Sorsum	RISE563
A	458	459	449	452	470	467	463
B	463	461.5	457	462	467	473.5	468.5
C	462.5	459	463	461.5	463.5	470	467
D	459.5	460	454.5	455.5	463	468.5	461
E	463.5	461	455.5	456.5	461	469	470.5
F	458.5	459	454	458	461.5	466.5	471.5
G	468	466	458	465	479	476	475.5
H	454	454.5	452.5	462	455.5	460.5	466

Values represent the average of two genomes of each genotype.

trees for HBVs. Our findings on the ancient HBV isolates here, rather than the conclusions from previous analyses (Krause-Kyora et al. 2018; Muhlemann et al. 2018), appear more consistent with a priori expectation on the ancient ancestors. That the oldest Karsdorf sample (7074 years ago) found in Germany was closest to present day genotype G that is also common in Germany, France and the United States is also consistent with expectation and further supports the conclusion of genetic continuity in HBVs in the last 7000 years (Sunbul 2014). That Sorsum and RISE563 samples were closely related to Karsdorf indicates also an affinity of these two isolates with genotype G although they were relatively more related to genotype E than Karsdorf. E is the most closely related to G among present HBV isolates. As E is found today only in Africa (Andernach et al. 2009), our finding here is in line with the known migration of Europeans into Africa during the last 5000 year period (Fregel et al. 2018).

The closer nucleotide distance of Sorsum with RISE563 than with Karsdorf may reflect linear phase distance whereas the closer amino acid distance of Sorsum with Karsdorf may reflect continuity in housekeeping functions. That Karsdorf was most similar to gorilla HBV in nucleotide sequence but to present day human HBV in amino acid sequence in the polymerase or the slow region may reflect continuation in the housekeeping physiology of the viruses and merely fortuitous similarities in nucleotide sequences reflecting some shared environmental adaptation. The closer similarity of ancient HBVs with NHP HBVs in non-polymerase proteins may reflect non-housekeeping adaptive roles in these proteins: the ancient HBV non-polymerase proteins in the ancient human hosts are expected to be more similar to NHP proteins given that the primitive life style of ancient humans may be less different from that of NHP.

The three ancient samples differ from genotype G however in that they lacked the 36 nucleotide insertion in the core gene region that is specific to HBV-G. This indicates that the insertion of this 36 bp sequence may have taken place later than 7000 years ago.

Genotype D is known to have the largest worldwide distribution, which is consistent with its basal position in the new tree here and its widest within genotype genetic diversity as found here. D shares the 33 nucleotide deletion in the pre S region with the NHP HBVs and the ancient HBVs, again indicating its more primitive nature. The insertion of 33 nucleotide region in other genotypes may have taken place when they first diverged from genotype D. As African NHP HBVs were closer to G while Asian NHP HBVs were closer to C, there was likely transmission of human HBVs into NHPs via migration of G from Europe into Africa and migration of C in the Asia mainland where it is most common today into the islands in Southeast Asia.

The new HBV phylogenetic tree appears more consistent with the reality of genotype distribution around the globe than the existing ones and has implications for the origin of the virus. That the first split was between D and the rest followed by the split of F and H indicates that HBV may have originated in a region covering Northeast Asia and Siberia where D may have originated. It then went to either the West to give rise to G, E, and A or the South to give rise to B and C. The estimated age of the human HBV of ∼20500 years corresponds to the period of Last Glacial Maximum (LGM), indicating that human population expansion post LGM may have played a role in the origin of HBV (Tallavaara et al. 2015).

It remains uncertain whether the sequence identity between the ancient HBV and human HBV is significantly greater than that between the ancient HBV and NHP HBV. The limited number of ancient samples makes the ideal statistical test not realistic here: if one has 10 ancient samples and all of them are closer to human than to NHP (even though only slightly closer as found here), then it is statistically significant. Future sampling of more ancient samples would help resolve this issue.

Our slow clock method of phylogenetic inferences here used only identity scores or distance matrix and we believe that one should not use amino acid variation matrix and evolutionary models in phylogenetic inferences. The other commonly used methods are Maximum Likelihood (ML) methods and related Bayesian methods, which all require evolutionary models. In the case of amino acid substitution models, such as giving high probability for R-K change than for R-L change, the probability matrix of change were derived from observing large numbers of protein alignments. Mismatches in such alignments are largely due to functional selection rather than neutral drift according to the MGD theory (R-K change being more common than R-L change is indication of functional selection). One simply cannot use evolutionary models as no one really knows the chain of events from the ancestor to the extant species, which can be extremely hard to model realistically. That leaves us only with the distance methods, which when used properly can qualify for the job. First, using slow evolving genes still at the linear phase of change will solve the issue of saturation, and the field needs to get rid of the approaches involving unreliable corrections on saturated distances in fast evolving genes (most such saturations are in fact maximum, which means that it is simply impossible to correct). Second, changes in such slow evolving genes qualify as neutral, because they are under neither positive nor negative selection. They are too slow to meet adaptive needs to be positively selected and their occurrences per se indicate lack of strong negative selection. Hence, an amino acid may truly have near equal probability of changing to any other. Finally, slow evolving genes may vary less dramatically among species. We can always restrict our analyses on closely related species (do human-ape tree, followed by ape-monkey tree, followed by monkey-prosimian tree).

The study here demonstrates the informative nature of the slow clock method in resolving the longstanding uncertainty on HBV origin and evolution. While it is widely known that different regions of a virus genome may produce different phylogenetic trees, it remains uncertain as to which region or whether the whole genome is the most informative to phylogenetic inferences. The MGD theory and the study here illustrate the informative nature of the slow evolving regions. The method should be generally applicable to evolutionary studies of other viruses.

4. Materials and methods

Present day HBV sequences were retrieved from Genbank and HBVdb database. Ancient HBV genomes were downloaded using previously published accession numbers. Alignments in nucleotide and amino acid sequences were done by blast. Fisher's exact test was used to estimate p values, 2 tailed.

Acknowledgements

We thank Ben Krause-Kyora and Julian Susat for sharing the sequences of ancient human HBV samples. This work is supported by the National

Natural Science Foundation of China grant 81171880, the National Basic Research Program of China grant 2011CB51001, and the Furong Scholars program (S. H.).

Author contributions

XL, YZ, SH designed the study and performed data analyses. SH wrote the first draft and all authors contributed to the final version.

Financial interest statements

The authors declare that they have no competing interests that might be perceived to influence the results and/or discussion reported in this paper.

References

Andernach, IE, JM Hubschen, CP Muller. 2009. Hepatitis B virus: the genotype E puzzle. Rev Med Virol 19:231–240.

Castelhano, N, NM Araujo, M Arenas. 2017. Heterogeneous recombination among Hepatitis B virus genotypes. Infect Genet Evol 54:486–490.

Fregel, R, FL Mendez, Y Bokbot, et al. 2018. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. Proc Natl Acad Sci USA 115:6774–6779.

Gui, Y, X Lei, S Huang. 2017. Collective effects of common SNPs and genetic risk prediction in type 1 diabetes. Clin Genet.

Hahn, MW. 2008. Toward a selection theory of molecular evolution. Evolution 62:255–265.

Hayer, J, F Jadeau, G Deleage, A Kay, F Zoulim, C Combet. 2013. HBVdb: a knowledge database for Hepatitis B Virus. Nucleic Acids Res 41:D566–570.

He, P, X Lei, D Yuan, Z Zhu, S Huang. 2017. Accumulation of minor alleles and risk prediction in schizophrenia. Sci Rep 7:11661.

Hu, T, M Long, D Yuan, Z Zhu, Y Huang, S Huang. 2013. The genetic equidistance result, misreading by the molecular clock and neutral theory and reinterpretation nearly half of a century later. Sci China Life Sci 56:254–261.

Huang, S. 2009. Inverse relationship between genetic diversity and epigenetic complexity. Preprint available at Nature Precedings: doi: doi.org/10.1038/npre.2009.1751.1032.

Huang, S. 2012. Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci 55:709–725.

Huang, S. 2016. New thoughts on an old riddle: What determines genetic diversity within and between species? Genomics 108:3–10.

Kahila Bar-Gal, G, MJ Kim, A Klein, et al. 2012. Tracing hepatitis B virus to the 16th century in a Korean mummy. Hepatology 56:1671–1680.

Kern, AD, MW Hahn. 2018. The Neutral Theory in Light of Natural Selection. Mol Biol Evol 35:1366–1371.

Krause-Kyora, B, J Susat, FM Key, et al. 2018. Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. Elife 7.

Kreitman, M. 1996. The neutral theory is dead. Long live the neutral theory. Bioessays 18:678–683; discussion 683.

Leffler, EM, K Bullaughey, DR Matute, WK Meyer, L Segurel, A Venkat, P Andolfatto, M Przeworski. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol 10:e1001388.

Lei, X, S Huang. 2017. Enrichment of minor allele of SNPs and genetic prediction of type 2 diabetes risk in British population. PLoS ONE 12:e0187644.

Lei, X, J Yuan, Z Zhu, S Huang. 2018. Collective effects of common SNPs and risk prediction in lung cancer. Heredity: doi: 10.1038/s41437-41018-40063-41434.

Locarnini, S, M Littlejohn, MN Aziz, L Yuen. 2013. Possible origins and evolution of the hepatitis B virus (HBV). Semin Cancer Biol 23:561–575.

Muhlemann, B, TC Jones, PB Damgaard, et al. 2018. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. Nature 557:418–423.

Ohta, T, JH Gillespie. 1996. Development of Neutral and Nearly Neutral Theories. Theor Popul Biol 49:128–142.

Shih, AC, TC Hsiao, MS Ho, WH Li. 2007. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc Natl Acad Sci USA 104:6283–6288.

Souza, BF, JF Drexler, RS Lima, O Rosario Mde, EM Netto. 2014. Theories about evolutionary origins of human hepatitis B virus in primates and humans. Braz J Infect Dis 18:535–543.

Sunbul, M. 2014. Hepatitis B virus genotypes: global distribution and clinical importance. World J Gastroenterol 20:5427–5434.

Tallavaara, M, M Luoto, N Korhonen, H Jarvinen, H Seppa. 2015. Human population dynamics in Europe over the Last Glacial Maximum. Proc Natl Acad Sci USA 112:8232–8237.

Teske, PR, TR Golla, J Sandoval-Castillo, A Emami-Khoyi, CD van der Lingen, S von der Heyden, B Chiazzari, B Jansen van Vuuren, LB Beheregaray. 2018. Mitochondrial DNA is unsuitable to test for isolation by distance. Sci Rep 8:8448.

Yuan, D, S Huang. 2017. On the peopling of the Americas: molecular evidence for the Paleoamerican and the Solutrean models. bioRxiv 130989; doi: https://doi.org/10.1101/130989.

Yuan, D, X Lei, Y Gui, Z Zhu, D Wang, J Yu, S Huang. 2017. Modern human origins: multiregional evolution of autosomes and East Asia origin of Y and mtDNA. bioRxiv 106864; doi: https://doi.org/10.1101/106864.

Yuan, D, Z Zhu, X Tan, et al. 2012. Minor alleles of common SNPs quantitatively affect traits/diseases and are under both positive and negative selection. arXiv:1209.2911.

Yuan, D, Z Zhu, X Tan, et al. 2014. Scoring the collective effects of SNPs: association of minor alleles with complex traits in model organisms. Sci China Life Sci 57:876–888.

Zhu, Z, Q Lu, J Wang, S Huang. 2015a. Collective effects of common SNPs in foraging decisions in Caenorhabditis elegans and an integrative method of identification of candidate genes. Sci. Rep.: doi: 10.1038/srep16904.

Zhu, Z, D Yuan, D Luo, X Lu, S Huang. 2015b. Enrichment of Minor Alleles of Common SNPs and Improved Risk Prediction for Parkinson's Disease. PLoS ONE 10:e0133421.

Xiaoyun Lei
Center for Medical Genetics
School of Biological Sciences
Central South University
110 Xiangya Road, Changsha
Hunan 410078
China
Institute of Molecular Precision Medicine
Xiangya Hospital
Central South University
110 Xiangya Road, Changsha
Hunan 410078
China
E-mail address: leixiaoyun@sklmg.edu.cn

Ye Zhang
Center for Medical Genetics
School of Biological Sciences
Central South University
110 Xiangya Road, Changsha
Hunan 410078
China
E-mail address: zhangye@sklmg.edu.cn

Shi Huang
Center for Medical Genetics
School of Biological Sciences
Central South University
110 Xiangya Road, Changsha
Hunan 410078
China
E-mail address: huangshi@sklmg.edu.cn

Received September 21, 2019