Ancient hepatitis B viruses from the Bronze Age to the Medieval period

Barbara Mühlemann^{1,29}, Terry C. Jones^{1,2,29}, Peter de Barros Damgaard^{3,29}, Morten E. Allentoft^{3,29}, Irina Shevnina⁴, Andrey Logvin⁴, Emma Usmanova⁵, Irina P. Panyushkina⁶, Bazartseren Boldgiv⁷, Tsevel Bazartseren⁸, Kadicha Tashbaeva⁹, Victor Merz¹⁰, Nina Lau¹¹, Václav Smrčka¹², Dmitry Voyakin¹³, Egor Kitov¹⁴, Andrey Epimakhov¹⁵, Dalia Pokutta¹⁶, Magdolna Vicze¹⁷, T. Douglas Price¹⁸, Vyacheslav Moiseyev¹⁹, Anders J. Hansen³, Ludovic Orlando^{3,20}, Simon Rasmussen²¹, Martin Sikora³, Lasse Vinner³, Albert D. M. E. Osterhaus²², Derek J. Smith¹, Dieter Glebe^{23,24}, Ron A. M. Fouchier²⁵, Christian Drosten^{2,26}, Karl-Göran Sjögren¹⁸, Kristian Kristiansen¹⁸ & Eske Willerslev^{3,27,28*}

Hepatitis B virus (HBV) is a major cause of human hepatitis. There is considerable uncertainty about the timescale of its evolution and its association with humans. Here we present 12 full or partial ancient HBV genomes that are between approximately 0.8 and 4.5 thousand years old. The ancient sequences group either within or in a sister relationship with extant human or other ape HBV clades. Generally, the genome properties follow those of modern HBV. The root of the HBV tree is projected to between 8.6 and 20.9 thousand years ago, and we estimate a substitution rate of 8.04×10^{-6} – 1.51×10^{-5} nucleotide substitutions per site per year. In several cases, the geographical locations of the ancient genotypes do not match present-day distributions. Genotypes that today are typical of Africa and Asia, and a subgenotype from India, are shown to have an early Eurasian presence. The geographical and temporal patterns that we observe in ancient and modern HBV genotypes are compatible with well-documented human migrations during the Bronze and Iron Ages^{1,2}. We provide evidence for the creation of HBV genotype A via recombination, and for a long-term association of modern HBV genotypes with humans, including the discovery of a human genotype that is now extinct. These data expose a complexity of HBV evolution that is not evident when considering modern sequences alone.

HBV is transmitted perinatally or horizontally via blood or genital fluids³. The estimated global prevalence is 3.6%, ranging from 0.01% (UK) to 22.38% (South Sudan)⁴. In high endemicity areas, in which prevalence is over 8%, 70–90% of the adult population show evidence of past or present infection⁵ (<http://www.who.int/mediacentre/factsheets/fs204/en/>). The young and the immunocompromised are most likely to develop chronic HBV infection, which can result in high viraemia over years to decades³. Approximately 257 million people are chronically infected and around 887,000 people died in 2015 owing to associated complications (<http://www.who.int/mediacentre/factsheets/fs204/en/>).

Despite the prevalence and public health impact of HBV, its origin and evolution remain unclear^{6,7}. Inference of HBV nucleotide substitution rates is complicated by the fact that the virus genome consists of four overlapping open reading frames⁸, and that mutation rates differ

between phases of chronic infection⁹. Studies based on heterochronous sequences, sampled over a relatively short time period, find higher substitution rates, whereas rates estimated using external calibrations tend to be lower, leading to a wide range of estimated HBV substitution rates (7.72×10^{-4} – 3.7×10^{-6} substitutions per site per year)^{10–12}. Human HBV is classified into at least nine genotypes (A–I) based on sequence similarity of at least 92.5% within genotypes¹³, with a heterogeneous global distribution^{7,8} (Fig. 1a). Attempts to explain the origin of genotypes using human migrations have been inconclusive. The hypothesis that HBV co-evolved with modern humans as they left Africa 60–100 thousand years ago (ka) has been contested owing to the basal phylogenetic position of genotypes F and H, which are found exclusively in the Americas⁶. HBV also infects non-human primates, and the human and other great ape HBVs are interspersed in the phylogenetic tree, possibly owing to cross-species transmission¹⁴. Given the variability of estimated substitution rates, the incongruence of the tree topology with some human migrations and the mixed topology of the non-human primate and human HBV sequences in the phylogenetic tree, there remains considerable uncertainty about the evolutionary history of HBV.

Recent advances in the sequencing of ancient DNA (aDNA) have yielded important insights into human evolution, past population dynamics¹⁵ and diseases^{16,17}. However, ancient sequences have been recovered for only a handful of exogenous human viruses, including influenza virus (sample approximately 100 years old)¹⁸, variola virus (sample approximately 350 years old)¹⁹ and HBV (samples approximately 340 and 450 years old)^{20,21}. The knowledge gained from these cases emphasizes the general importance of ancient sequences for the direct study of long-term viral evolution. HBV has several characteristics that make it a good candidate for detection in an aDNA virus study: its extended high viraemia during chronicity³, the relative stability of its virion²², and its small, circular and partially double-stranded DNA genome⁸.

Shotgun sequence data were previously generated from 167 Bronze Age¹ and 137 predominantly Iron Age² individuals from central to western Eurasia with a sample age range of approximately 7.1–0.2 thousand years (kyr) old. We identified reads that matched the HBV genome in

¹Center for Pathogen Evolution, Department of Zoology, University of Cambridge, Cambridge, UK. ²Institute of Virology, Charité, Universitätsmedizin Berlin, Berlin, Germany. ³Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Copenhagen, Denmark. ⁴Archaeological Laboratory, Faculty of History and Law, A. A. Baitursynov Kostanay State University, Kostanay, Kazakhstan. ⁵Saryarka Archaeological Institute, Karaganda State University, Karaganda, Kazakhstan. ⁶Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ, USA. ⁷Department of Biology, School of Arts and Sciences, National University of Mongolia, Ulaanbaatar, Mongolia. ⁸Laboratory of Virology, Institute of Veterinary Medicine, Mongolian University of Life Sciences, Ulaanbaatar, Mongolia. ⁹National Academy of Sciences, Bishkek, Kyrgyzstan. ¹⁰Pavlodar State University, Pavlodar, Kazakhstan. ¹¹Centre for Baltic and Scandinavian Archaeology, Schleswig, Germany. ¹²Institute for History of Medicine and Foreign Languages of the First Faculty of Medicine, Charles University, Prague, Czech Republic. ¹³Margulan Institute of Archaeology, Almaty, Kazakhstan. ¹⁴N. N. Miklouho-Maklay Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. ¹⁵South Ural Department, Institute of History and Archaeology UBRS, South Ural State University, Chelyabinsk, Russia. ¹⁶Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden. ¹⁷Matrica Museum, Százhalombatta, Hungary. ¹⁸Department of Historical Studies, University of Gothenburg, Gothenburg, Sweden. ¹⁹Department of Physical Anthropology, Peter the Great Museum of Anthropology and Ethnography, Saint-Petersburg, Russia. ²⁰Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, Toulouse, France. ²¹Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark. ²²Research Center for Emerging Infections and Zoonoses, University of Veterinary Medicine Hannover, Hannover, Germany. ²³Institute of Medical Virology, Justus Liebig University of Giessen, Giessen, Germany. ²⁴National Reference Centre for Hepatitis B and D Viruses, German Center for Infection Research (DZIF), Giessen, Germany. ²⁵Department of Viroscience, Erasmus Medical Centre, Rotterdam, The Netherlands. ²⁶German Center for Infection Research (DZIF), Braunschweig, Germany. ²⁷Cambridge GeoGenetics Group, Department of Zoology, University of Cambridge, Cambridge, UK. ²⁸Wellcome Trust Sanger Institute, Hinxton, UK. ²⁹These authors contributed equally: Barbara Mühlemann, Terry C. Jones, Peter de Barros Damgaard, Morten E. Allentoft. *e-mail: ewillerslev@snm.ku.dk

Fig. 1 | Geographical distribution of analysed samples and modern genotypes. **a**, Distribution of modern human HBV genotypes⁷. Genotypes relevant to this Letter are shown in colour. Coloured shapes indicate the locations of the HBV-positive samples included for further analysis. **b**, Locations of analysed Bronze Age samples¹ are shown as circles and Iron Age and later samples² are shown as triangles. Coloured markers indicate HBV-positive samples. Ancient genotype A samples are found in regions in which genotype D predominates today, and HBV-DA27 is of sub-genotype D5 which today is found almost exclusively in India.

25 samples (Table 1, Extended Data Table 1a and Supplementary Table 3), spanning a period of almost 4,000 years, from several different cultures and with a broad geographical range (Fig. 1b, Table 1, Extended Data Table 1a and Supplementary Table 3). Using TaqMan PCR, we tested two samples (DA195 and DA222) with high genome coverage and two samples (DA85 and DA89) with low genome coverage for the presence of HBV. The high-coverage samples tested positive, whereas the low-coverage samples tested negative (Extended Data Table 1b). This is consistent with shotgun sequencing being more effective than targeted PCR for analysing highly degraded DNA²³. On the basis of the availability of sample material, libraries from 14 samples were selected for targeted enrichment (capture) of HBV DNA fragments (Supplementary Tables 1, 2). This resulted in increased genome coverage and an average of a 2.4-fold increase in the number of HBV-positive reads (Extended Data Table 1a and Supplementary Table 3).

We obtained 17.9–100% HBV genome coverage from the sequence data, with genomic depth ranging from $0.4\times$ to $89.2\times$ (Table 1 and Extended Data Table 1a). We selected 12 samples for phylogenetic analyses. Criteria for inclusion were at least 50% genome coverage and clear aDNA damage patterns after capture (Extended Data Fig. 1).

For an initial phylogenetic grouping, we estimated a maximum likelihood tree using the ancient HBV genomes together with modern human, non-human primate, rodent and bat HBV genomes (dataset 1, see Methods). All ancient viruses fell within the diversity of Old World primate HBV genotypes, which includes all human and other great ape genotypes with the exception of human genotypes F and H (Extended Data Fig. 2).

Recombination is known to occur in HBV²⁴. We found strong evidence that an ancient sequence (HBV-DA51) and an unknown parent recombined to form the ancient genotype A sequences. Although this cannot literally be the case owing to sample ages, the logical interpretation is that an ancestor of HBV-DA51 was involved in the recombination. The same recombination is also suggested for the two modern genotype A sequences that were included in the analysis. The ancient genotype B (HBV-DA45), a modern genotype B and two modern genotype C sequences were not similarly flagged, which suggests that the possible recombination occurred after genotypes A, B and C had diverged. The predicted recombination break points (Extended Data Table 2 and Extended Data Fig. 3) correspond closely to the polymerase gene. It is therefore possible that the polymerase from an unknown parent and the remainder of the genome from an HBV-DA51 ancestor recombined to form the now-ubiquitous genotype A about 7.4–9 ka (Fig. 2, Extended Data Table 3b and Methods). Similar recombination events that involved the creation of genotypes E, G and a currently circulating B/C recombinant have previously been identified²⁴.

For detailed phylogenetic analyses, we used a set of 112 reference human and non-human primate HBV sequences (dataset 2, see Methods). A maximum likelihood phylogenetic tree based on these reference sequences and the 12 ancient sequences was constructed (Extended Data Fig. 4). Regression of root-to-tip genetic distances against sampling dates, as well as date randomization tests, showed a clear temporal signal in the data (Extended Data Fig. 5 and Supplementary Figs. 1–3), suggesting that molecular clock models can be applied. A dated coalescent phylogeny was constructed using BEAST²⁵ (Fig. 2). The molecular clock was calibrated using tip dates. Strict and relaxed log-normal molecular clocks were tested with coalescent constant, exponential and Bayesian skyline population priors (Extended Data Table 3a). Model comparisons favoured a relaxed molecular clock model with log-normally distributed rate variation and a coalescent exponential population prior (Extended Data Table 3a). The median root age of the resulting tree is estimated to be 11.6 kyr (95% highest posterior density (HPD) interval: 8.6–15.3 kyr).

Table 1 | Overview of samples used for phylogenetic analyses

Sample	¹⁴ C age (s.d.)	Median cal. BP age, or estimate (years)	Approximate sample age (years)	Site	Culture or period	Sex	Reads included in consensus	Coverage	
								Consensus	Depth
RISE563	3,955 (35)	4,421	4,488	Osterhofen-Altenmarkt, Germany	Bell Beaker	M	4,383	100%	79.3×
DA222	ND	1,200–1,000	1,167	Butakty, Kazakhstan	Karluk	M	4,132	100%	89.2×
DA195	2,479 (35)	2,578	2,645	Sandorfalva-Eperjes, Hungary	Hungarian Scythian	F	1,445	99.9%	29.2×
DA51	2,220 (37)	2,230	2,297	Keden, Kyrgyzstan	Saka	M	712	99.2%	14.5×
RISE254	3,631 (29)	3,942	4,009	Százhalombatta-Földvár, Hungary	Vatya	M	1,491	99.0%	36.6×
DA119	ND	1,500	1,567	Poprad, Slovakia	North Carpathian	M	2,597	98.8%	53.1×
RISE386	3,758 (34)	4,121	4,188	Bulanovo, Russia	Sintashta	M	331	97.8%	7.0×
DA27	1,641 (33)	1,543	1,610	Halvay 3, Kazakhstan	Hun–Sarmatian	M	890	90.0%	14.3×
DA29	849 (25)	755	822	Karasyur, Kazakhstan	Medieval	M	222	87.5%	4.8×
DA45	2,083 (27)	2,053	2,120	Omnogobi, Mongolia	Xiongnu	M	215	87.2%	4.3×
RISE387	3,822 (33)	4,215	4,282	Bulanovo, Russia	Sintashta	ND	284	86.6%	6.2×
RISE154	3,522 (24)	3,784	3,851	Szczepankowice, Poland	Unetice	F	128	57.2%	2.0×

Samples included in phylogenetic analysis, by decreasing genome coverage. Criteria for inclusion were at least 50% genome coverage and clear aDNA damage patterns after capture. The read count indicates the number of reads used to make consensus sequences. Cal. BP, calibrated years before present (taken as AD 1950); ND, not determined (samples for which dating was not performed, or for which osteological sex was undetermined). See Methods for information on sequence matching, consensus making and sample dating.

Fig. 2 | Dated maximum clade credibility tree of HBV. A log-normal relaxed clock and coalescent exponential population prior were used. Grey horizontal bars indicate the 95% HPD interval of the age of the node. Larger numbers on the nodes indicate the median age and 95% HPD interval of the age (in parentheses) under a strict clock and Bayesian skyline tree prior. Clades of genotypes C (except clade C4), E, F, G and H are collapsed and shown as dots. The figure includes a possible tenth genotype, J, based on a single human isolate. Taxon names for ancient

samples indicate era (BA, Bronze Age; IA, Iron Age or later), sample name, sample age in years, ISO 3166 three-letter abbreviation of country of sequence origin, and region of sequence origin. Taxon names for modern samples indicate human genotype or subgenotype or host species if non-human, GenBank accession number, sample age in years, ISO 3166 three-letter abbreviation of country of sequence origin, and region of sequence origin.

and the median clock rate is 1.18×10^{-5} substitutions per site per year (95% HPD interval: 9.21×10^{-6} – 1.45×10^{-5} substitutions per site per year). Under a strict molecular clock, a coalescent Bayesian skyline population prior was favoured, in which case the median root age is

15.6 kyr (95% HPD interval: 13.7–17.8 kyr) and the median substitution rate is 9.48×10^{-6} substitutions per site per year (95% HPD interval: 8.3×10^{-6} – 1.07×10^{-5} substitutions per site per year) (Extended Data Table 3a–c).

Under all model parameterizations used here, the substitution rate that we find is lower than rates estimated from phylogenies built using either modern heterochronous sequences¹⁰ or sequences from mother-to-child transmissions²⁶ but higher than rates inferred using external calibrations based on human migrations¹¹. A lower rate is consistent with previous work²⁷ in which it was shown that, although mutation rates may be high, mutations within an individual often revert back to the genotype consensus and thus rarely lead to long-term sequence change. It is also consistent with the time-dependent rate phenomenon, observed for many viruses, which suggests that short-term evolutionary rates are higher than long-term rates²⁸.

The ancient HBV genome data enable us to formally evaluate hypotheses concerning HBV origins using path sampling of calibrated phylogenies based on appropriate external divergence date assumptions. We tested several calibration points that would be implied by a co-expansion of HBV with humans after leaving Africa for support of congruence between migrations and geographical locations of HBV clades¹¹. We find weak evidence for the split of the F and H clade occurring between 13.4 and 25.0 ka under a strict, but not a relaxed, clock model. We do not find support for the divergence of subgenotype C3 strains between 5.1 and 12.0 ka (hypothesized to have led to its distribution in different regions of Polynesia¹¹) or for divergence of Haitian A3 strains from other genotype A strains between 0.2 and 0.5 ka under either strict or relaxed clock models (Extended Data Table 3d).

In the dated coalescent phylogeny, four ancient sequences (from youngest to oldest: HBV-DA119, HBV-DA195, HBV-RISE386 and HBV-RISE387) group with genotype A. The first three are well within the 7.5% nucleotide divergence criterion that was used to delimit membership in HBV genotypes, and HBV-RISE387 is right on this limit (7.51%)¹³ (Extended Data Table 4a). The three oldest samples lack a six-nucleotide insertion at the carboxyl end of the core gene (C) that is present in all modern genotype A viruses⁶ (Table 2). HBV-RISE387 encodes a stop codon in its pre-core peptide that would have ablated the expression of the immune modulator HBe antigen (HBeAg), a phenomenon that is known to occur in modern HBV infections (Table 2). This characteristic viral mutant is usually found in chronic HBV carriers who seroconverted from HBeAg to anti-HBe. RISE386 and RISE387 have archaeologically dates of only about 100 years apart and both come from the Bulanovo site in Russia, but their viruses have only 93.34% sequence identity (Extended Data Table 4b), which indicates the existence of substantial localized HBV diversity about 4.2 ka.

The ancient sequence HBV-DA45 phylogenetically groups with genotype B and has 97.65% sequence identity with modern genotype B (Extended Data Table 4a).

Sequences HBV-DA27, HBV-DA29, HBV-DA51 and HBV-DA222 phylogenetically group with the modern genotype D. They have high sequence identity (96.99–98.74%) with modern genotype D sequences (Extended Data Table 4a), and have the typical 33-nucleotide deletion in the preS1 region of the S gene, encoding the three HBV surface proteins⁸ (Table 2).

Sequences HBV-RISE154, HBV-RISE254 and HBV-RISE563 are in a sister relationship with the chimpanzee–gorilla HBV clade (Fig. 2). HBV-RISE254 and HBV-RISE563 have the same 33-nucleotide deletion in the preS1 sequence that is shared with non-human primate HBVs and human genotype D (Table 2). HBV-RISE563 does not encode a functional pre-core peptide (Table 2). On the basis of sequence similarity across the whole genome, HBV-RISE563 and HBV-RISE254 together might be classified as a new human HBV genotype that is extinct today, and HBV-RISE154 might possibly be classified as another (Extended Data Table 4). However, HBV-RISE154 has low genome coverage, which precludes an exact calculation. The sister relationship of these three sequences with modern chimpanzee and gorilla HBVs could be interpreted as a consequence of relatively recent transmission(s) of HBV from humans to non-human primates¹⁴. However, other scenarios and confounding factors are possible, as these lineages are deeply separated in the tree. Incomplete lineage sorting combined with viral extinction (possibly boosted by massive recent reductions in great ape populations) should be considered. More data on current and, if possible, ancient HBVs will be necessary to reach definitive conclusions.

The geographical locations of some of the ancient virus genotypes do not match the present-day genotype distribution, and also do not match dates and/or locations inferred in previous studies of HBV. Although the data presented here are limited, they provide important spatio-temporal reference points in the evolutionary history of HBV. Their synopsis suggests a more complicated ancestry of present-day genotypes than previously assumed, especially in light of recent insights into the history of human migration.

We find genotype A in south-western Russia by 4.3 ka (in samples RISE386 and RISE387) in individuals belonging to the Sintashta culture, and in a Hungarian sample (DA195) from the Scythian culture. The western Scythians are related to the Bronze Age cultures of western steppe populations² and their shared ancestry suggests that the modern genotype A may descend from this ancient Eurasian diversity and not, as previously hypothesized, from African ancestors^{29,30}. This is also consistent with the phylogeny (Fig. 2), as well as the fact that the three oldest ancient genotype A sequences (HBV-DA195, HBV-RISE386 and HBV-RISE387) lack the six-nucleotide insertion found in the youngest (HBV-DA119) and in all modern genotype A sequences. The ancestors of subgenotypes A1 and A3 could have been carried into Africa subsequently, via migration from western Eurasia³¹.

The ancient HBV genotype D sequences were all found in Central Asia. HBV-DA27, found in Kazakhstan and dated to 1.6 ka, falls basal to the modern subgenotype D5 sequences that today are found in the Paharia tribe from eastern India³². DA27 and the Paharia people in India are linked by their East Asian ancestry^{2,33}.

Based on the observation that genotypes go extinct and can be created by recombination, the ancient sequence data show that the diversity that we observe today is only a subset of the diversity that has ever existed. These data support a scenario in which all present-day

Table 2 | Genome properties of ancient sequences included in phylogenetic analyses

Sample	Genotype of closest modern sequence	Sequence identity to closest modern sequence	Genome length	Serotype	Insertions/deletions	Predicted HBeAg status
DA119	A3	97.8%	3,221	adw2	6-nt	Positive
DA195	A3	96.2%	3,215	adw2	None	Positive
RISE386	A	95.2%	3,215	adw2	None	Positive
RISE387	A	92.5%	3,215	adw2	None	Negative; PreC stop codon
DA45	B1	96.6%	3,215	ayw1	None	Positive
DA29	D3	98.5%	3,182	ayw2	33-nt	Positive
DA222	D3	98.7%	3,182	ayw2	33-nt	Positive
DA27	D1	97.2%	3,182	ayw2	33-nt	Positive
DA51	D1	96.7%	3,182	ayw2	33-nt	Positive
RISE154	Chimpanzee	92.5%	Ambiguous	adw2*	Ambiguous	Positive
RISE254	Chimpanzee	95.2%	3,182	adw2	33-nt	Positive
RISE563	Gorilla	92.7%	3,182	adw2	33-nt	Negative; PreC stop codon

Genotype groups are sorted by increasing sample age. 6-nt, 6-nucleotide insertion at the C terminus of the core region; 33-nt, 33-nucleotide deletion at the N terminus of the preS1 region.

*Serotype could not be determined unambiguously, owing to lack of coverage.

HBV diversity arose only after the split of the Old World and New World genotypes (25–13.4 ka). Any attempt to interpret the currently known HBV tree based on human migrations that happened before this event will necessarily result in anomalies that cannot be reconciled, such as the basal position of genotypes F and H and the apical position of subgenotype C4, which is found exclusively in indigenous Australians⁸. If HBV did co-evolve with ancient modern humans as they left Africa as previously proposed⁶, most of the pattern of earlier diversity has been replaced by changes that happened after the split of the Old and New World genotypes. Genotypes F and H would therefore be remnants of the earlier now-extinct diversity, and the arrival of subgenotype C4 in Australia would have taken place long after the split between Old and New World genotypes, as supported by the tree in Fig. 2. Alternatively, there could have been a New World origin of HBV or the virus could have been introduced into humans from a different host. Our data do not allow us to speculate either way.

To our knowledge, we report the oldest exogenous viral sequences recovered from DNA of humans or any vertebrate, and show that it is possible to recover viral sequences from samples of this age. We show that humans throughout Eurasia were widely infected with HBV for thousands of years. Despite the age of the samples and the imperfect diagnostic test, our dataset contained a high proportion of HBV-positive individuals. The actual ancient prevalence during the Bronze Age and thereafter might have been higher, reaching or exceeding the prevalence typically found in contemporary indigenous populations⁵. This clearly establishes the potential of HBV as powerful proxy tool for research into human spread and interactions. The data from ancient genomes reveal aspects of complexity in HBV evolution that are not apparent when only modern sequences are considered. They show the existence of ancient HBV genotypes in locations incongruent with their present-day distribution, contradicting previously suggested geographical or temporal origins of genotypes or sub-genotypes; evidence for the creation of genotype A via recombination and the emergence of the genotype outside Africa; at least one now-extinct human genotype; ancient genotype-level localized diversity; and demonstrate that the viral substitution rate obtained from modern heterochronously sampled sequences is probably misleading. Together, these findings suggest that the difficulty in formulating a coherent theory for the origin and spread of HBV may be due to genetic evidence of an earlier evolutionary scenario being overwritten by relatively recent alterations, as has previously been suggested in the context of recombination²⁴. The lack of ancient sequences limits our understanding of the evolution of HBV and very probably of other viruses. Discovery of additional ancient viral sequences may provide a clearer picture of the true origin and early diversification of HBV, enable us to address questions of palaeo-epidemiology, and broaden our understanding of the contributions of natural and cultural changes (including migrations and medical practices) to human disease burden and mortality.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0097-z>.

Received: 14 April 2017; Accepted: 6 April 2018;
Published online 9 May 2018.

1. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
2. Damgaard, P. D. B. et al. 137 ancient human genomes from across the Eurasian steppes. *Nature* <https://doi.org/10.1038/s41586-018-0094-2> (2018).
3. Lai, C. L., Ratzliff, V., Yuen, M.-F. & Poyndar, T. Viral hepatitis B. *Lancet* **362**, 2089–2094 (2003).
4. Schweitzer, A., Horn, J., Mikolajczyk, R. T., Krause, G. & Ott, J. J. Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *Lancet* **386**, 1546–1555 (2015).
5. Murhekar, M. V., Murhekar, K. M. & Sehgal, S. C. Epidemiology of hepatitis B virus infection among the tribes of Andaman and Nicobar Islands, India. *Trans. R. Soc. Trop. Med. Hyg.* **102**, 729–734 (2008).

6. Locarnini, S., Littlejohn, M., Aziz, M. N. & Yuen, L. Possible origins and evolution of the hepatitis B virus (HBV). *Semin. Cancer Biol.* **23**, 561–575 (2013).
7. Littlejohn, M., Locarnini, S. & Yuen, L. Origins and evolution of hepatitis B virus and hepatitis D virus. *Cold Spring Harb. Perspect. Med.* **6**, a021360 (2016).
8. Kramvis, A. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **57**, 141–150 (2014).
9. Hannoun, C., Horal, P. & Lindh, M. Long-term mutation rates in the hepatitis B virus genome. *J. Gen. Virol.* **81**, 75–83 (2000).
10. Zhou, Y. & Holmes, E. C. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J. Mol. Evol.* **65**, 197–205 (2007).
11. Paraskevis, D. et al. Dating the origin of hepatitis B virus reveals higher substitution rate and adaptation on the branch leading to F/H genotypes. *Mol. Phylogenet. Evol.* **93**, 44–54 (2015).
12. Zehender, G. et al. Enigmatic origin of hepatitis B virus: an ancient travelling companion or a recent encounter? *World J. Gastroenterol.* **20**, 7622–7634 (2014).
13. Kramvis, A. et al. Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. *J. Med. Virol.* **80**, 27–46 (2008).
14. MacDonald, D. M., Holmes, E. C., Lewis, J. C. & Simmonds, P. Detection of hepatitis B virus infection in wild-born chimpanzees (*Pan troglodytes verus*): phylogenetic relationships with human and other primate genotypes. *J. Virol.* **74**, 4253–4257 (2000).
15. Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
16. Rasmussen, S. et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571–582 (2015).
17. Feldman, M. et al. A high-coverage *Yersinia pestis* genome from a sixth-century Justinianic plague victim. *Mol. Biol. Evol.* **33**, 2911–2923 (2016).
18. Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. Origin and evolution of the 1918 “Spanish” influenza virus hemagglutinin gene. *Proc. Natl Acad. Sci. USA* **96**, 1651–1656 (1999).
19. Duggan, A. T. et al. 17th century variola virus reveals the recent history of smallpox. *Curr. Biol.* **26**, 3407–3412 (2016).
20. Kahila Bar-Gal, G. et al. Tracing hepatitis B virus to the 16th century in a Korean mummy. *Hepatology* **56**, 1671–1680 (2012).
21. Patterson Ross, Z. et al. The paradox of HBV evolution as revealed from a 16th century mummy. *PLoS Pathog.* **14**, e1006750 (2018).
22. Bond, W. W. et al. Survival of hepatitis B virus after drying and storage for one week. *Lancet* **317**, 550–551 (1981).
23. Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
24. Simmonds, P. & Midgley, S. Recombination in the genesis and evolution of hepatitis B virus genotypes. *J. Virol.* **79**, 15467–15476 (2005).
25. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
26. Simmonds, P. Reconstructing the origins of human hepatitis viruses. *Phil. Trans. R. Soc. Lond. B* **356**, 1013–1026 (2001).
27. Tedder, R. S., Bissett, S. L., Myers, R. & Ijaz, S. The ‘Red Queen’ dilemma—running to stay in the same place: reflections on the evolutionary vector of HBV in humans. *Antivir. Ther.* **18**, 489–496 (2013).
28. Duchêne, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. Lond. B* **281**, 20140732 (2014).
29. Zehender, G. et al. Reliable timescale inference of HBV genotype A origin and phylogenetics. *Infect. Genet. Evol.* **32**, 361–369 (2015).
30. 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.* **86**, 2163–2167 (2005).
31. Pickrell, J. K. et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl Acad. Sci. USA* **111**, 2632–2637 (2014).
32. Ghosh, S. et al. Unique hepatitis B virus subgenotype in a primitive tribal community in eastern India. *J. Clin. Microbiol.* **48**, 4063–4071 (2010).
33. Basu, A., Sarkar-Roy, N. & Majumder, P. P. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl Acad. Sci. USA* **113**, 1594–1599 (2016).

Acknowledgements B.B. thanks D. Tserendulam for help, wisdom and guidance. We thank S. Rankin and the staff of the University of Cambridge High Performance Computing service and the National High-throughput Sequencing Centre (Copenhagen). This work was supported by: The Danish National Research Foundation, The Danish National Advanced Technology Foundation (The Genome Denmark platform, grant 019-2011-2), The Villum Kann Rasmussen Foundation, KU2016, European Union FP7 programme ANTIGONE (grant agreement No. 278976), and European Union Horizon 2020 research and innovation programmes, COMPARE (grant agreement No. 643476), VIROGENESIS (grant agreement No. 634650). The National Reference Center for Hepatitis B and D Viruses is supported by the German Ministry of Health via the Robert Koch Institute (Berlin). B.B. was supported by Taylor Family-Asia Foundation Endowed Chair in Ecology and Conservation Biology. A.D.M.E.O. was supported by N-RENNT of the Ministry of Science and Culture of Lower Saxony, Germany.

Reviewer information Nature thanks P. Simmonds, B. Shapiro, C. Pepperell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed to the interpretation of the results. B.M., T.C.J., P.d.B.D., M.E.A., S.R., M.S., L.O., L.V., D.J.S., D.G., R.A.M.F., C.D. and E.W. wrote the paper. B.M. and T.C.J. screened and analysed data, and created display items. P.d.B.D. and M.E.A. conducted sampling and generated sequence data. I.S., A.L., E.U., I.P.P., B.B., T.B., K.T., V.M., N.L., V.S., D.V., E.K., A.E., D.P., M.V., T.D.P. and V.M. excavated, curated, and analysed samples and archaeological contexts. A.J.H. designed virus capture probes. L.V. designed virus capture probes, performed TaqMan PCR and target enrichment experiments. A.D.M.E.O. initiated and provided critical input on the development of next-generation sequencing bioinformatics tools. B.M., T.C.J., D.J.S., D.G., R.A.M.F. and C.D. performed recombination analysis. C.D. analysed data and designed PCR probes. K.-G.S. and K.K. conducted sampling and provided archaeological background. E.W. initiated the work, and led sampling and generation of the sequence data.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0097-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0097-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

HBV datasets. The following HBV datasets were used in the present study. Full listings of accession numbers are given in the Supplementary Methods.

Dataset 1. Dataset 1 comprises 26 HBV genomes, covering all species in the *Orthohepadnaviridae*. This includes one sequence each from the human HBV genotypes (A–J), orangutan, chimpanzee, gorilla, gibbon, woolly monkey, woodchuck, ground squirrel, Arctic ground squirrel and horseshoe bat, four sequences from roundleaf bats, and three sequences from tent-making bats, largely following a previous publication³⁴.

Dataset 2. Dataset 2 comprises 124 HBV genomes, from humans and non-human primates. This set contains 92 sequences from a previous publication¹¹ (excluding their incomplete sequences), 7 additional genotype D sequences, the Korean mummy genotype C sequence²⁰, the 12 ancient sequences from the present study and 12 full genomes selected from a set of 9,066 full HBV genomes downloaded from NCBI³⁵ on 24 August 2017 (Entrez query: hepatitis b virus[organism] not rna[title] not clone[title] not clonal[title] not patent[title] not recombinant[title] not recombination[title] and 3000:4000[sequence length]) corresponding to the closest, non-artificial match for each of the ancient sequences. Dates for these sequences were acquired by looking for a date of sample collection in the NCBI entry, or the paper in which the sequence was first published. If a range of dates was mentioned, the mean was used. If no date of sample collection was found in this way, either the year of the publication of the paper, or the year of addition of the sequence to GenBank was used, whichever was earlier.

Dataset 3. Dataset 3 comprises 124 HBV genomes, from humans, non-human primates and a variety of other *Orthohepadnaviridae* host species, including woolly monkey, roundleaf bat, tent-making bat, ground squirrel, Arctic ground squirrel, woodchuck and snow goose. This set contains 113 sequences that were obtained from the union of 91 sequences from Paraskevis et al.¹¹ and 29 from Drexler et al.³⁴, plus 11 additional sequences (giving 124 sequences in total).

Dataset 4. Dataset 4 comprises 3,505 HBV genomes. Of these, 3,384 are from a previous publication³⁶, divided into ten human genotypes. To these, we added 17 chimpanzee, 56 gorilla, 12 gibbon and 36 orangutan full HBV genome sequences downloaded from NCBI on 18 January 2017, resulting in 14 genome categories.

Dating of ancient samples. Sample ages were determined by direct ¹⁴C dating. These ages were calibrated using OxCal³⁷ (version 4.3) using the IntCal13 curve³⁸. Table 1 shows the ¹⁴C age and standard deviation for each sample. This is followed by the median probability calibrated age before present (cal. BP). RISE386 was ¹⁴C dated twice, with ages (standard deviation) of 3,740 (33) and 3,775 (34); a rounded mean of 3,758 (34) was used for its calibration. DA29 was dated at 822 years using ¹⁴C and also at about 700 years using multi-proxy methods: the former date was used for consistency. The dates for DA119, DA222, RISE548, RISE556, RISE568 and RISE597 are best estimates, based on sample context.

Data and data processing. We analysed 101 Bronze Age samples published in Allentoft et al.¹, 137 predominantly Iron Age samples published in Damgaard et al.² and 66 additional samples from the Bronze Age. A total of 114.58×10^9 Illumina HiSeq 2500 sequencing reads were processed.

AdapterRemoval³⁹ (version 2.1.7) was used with its default settings to remove adaptors from all sequences, to trim N bases from the ends of reads and to trim bases with quality ≤ 2 . Reads were aligned against a human genome (GRCh38, <https://www.ncbi.nlm.nih.gov/grc/human>) using BWA⁴⁰ (version 0.7.15-r1140, mem algorithm). Reads that did not match the human genome were then mapped against the NCBI viral protein reference database containing 274,038 viral protein sequences (downloaded on 31 August 2016) using DIAMOND⁴¹ (version 0.8.25). Protein matches were grouped into their corresponding viruses. Reads matching HBV were found in 25 samples.

The non-human reads from the samples that had more than three reads matching HBV using DIAMOND were selected for a subsequent BLAST⁴² (version 2.4.0) analysis. A BLAST database was made from dataset 3, and samples were matched using BLASTn (with arguments -task blastn -evalue 0.01). Matching reads with bit scores greater than 50 for all samples (except DA222 (70) and DA45 (55)) were selected for subsequent processing. The number of reads selected from the BLAST matches, per sample, is shown in Table 1, with additional detail in Extended Data Table 1. Across all samples 11,149 reads matched against HBV sequences.

PCR confirmation. Real-time PCR was established using primers and TaqMan probes as previously described⁴³, which were used to amplify a 91-base-pair amplicon of the HBV genome. Primers and probe were added to QuantiTect PCR mix (Qiagen #204343) in a final concentration of 400 nM or 200 nM, respectively, in a total reaction volume of 25 μ l, including 5 μ l template. Using the Roche LC480 or Agilent Mx3006p instruments, PCRs were incubated for 15 min at 95 °C followed by 45 cycles of 15 s at 94 °C and 60 s at 60 °C, measuring fluorescence from the

6-carboxy-fluorescein/BHQ1-labelled probe and the passive dye (ROX) at the end of each cycle.

Careful precautions were taken to prevent PCR contamination. PCR mastermixes were prepared in dedicated aDNA clean laboratory facilities, in which no prior targeted work has been carried out on HBV. aDNA extracts and non-template controls (NTCs) were added into PCR reactions in this location too, and were not subsequently opened. Positive control material was handled in laboratories in a physically separated building. Here, standard material, diluted to 5–50 copies per reaction, was added to duplicate PCR reactions along with additional NTCs.

Virus capture. Fourteen samples with sufficient sample material were selected for virus capture (DA27, DA29, DA45, DA51, DA85, DA89, DA119, DA195, DA222, RISE254, RISE386, RISE416, RISE568 and RISE556). The viral reference genomes for probes were selected as follows. The International Committee for Taxonomy of Viruses (ICTV) 2012 listed 2,618 viral species. As many had no associated reference genomes or merely partial sequence information, we selected 2,599 sequences of full-length viral genomes, available from GenBank (June 2014), representing viral species found in vertebrates excluding fish. Sequences < 1,000 nucleotides were discarded. Sequences with identical length and organism identification were regarded as duplicates and thus reduced to 1. For a number of specific viral taxa for which a large number of similar reference sequences are available, we manually selected representative genomes or genome segments (Supplementary Tables 1, 2). For example, among 72 available hepatitis C virus genome sequences, we selected one genome per subtype (subtypes 1a–1c, 1g; 2a–c, 2i, 2k; 3a, 3b, 3i, 3k; 4a–4d, 4f, 4g, 4k–4r, 4t; 5a; 6a–6u; 7a). Likewise, 12 HIV-1 genomes were selected to represent group M (subtypes A–D, F1, F2, H, J, K, N, O and P). For influenza A virus, we included only sequences from segment 7 and segment 5 that encode the conserved matrix proteins M1/M2 and the nucleocapsid protein NP, respectively. We selected 82 M1/M2 segments and 115 NP segments among the available segment sequences. All available segments were included from genomes belonging to *Arenaviridae*, *Bunyavirales* and *Reoviridae*. For members of *Poxviridae* for which full genomes were unavailable (skunkpox, raccoonpox and volepox viruses) sequences representing the conserved gene encoding the DNA-dependent RNA polymerase were included ($n = 22$). In addition, two partial genomes of squirrelpox virus were included. By mistake, two and nine partial sequences were included from *Iridoviridae* (1.5–2.5 kb) and *Coronaviridae* (1.3–14.5 kb), respectively, already represented by full genomes. Likewise, sequences representing Merkel cell polyomavirus and KI polyomavirus were not included among the reference genomes used for probe design. SeqCap EZ hybridization probes were designed and synthesized by Roche NimbleGen based on the resulting reference sequences.

Capture was performed on double-indexed libraries prepared from aDNA, following the manufacturer's protocol (version 4.3) with the following modifications. In brief, 1.8 to 2.2 μ g of pooled libraries were hybridized at 47 °C for 65–70 h with low complexity C₆T-1 DNA, specific P5/P7 adaptor-blocking oligonucleotides each containing a hexamer motif of inosine nucleotides to match individually indexed adapters, hybridization buffer containing 10% formamide, and the capture probes. Dynabeads M-270 (Invitrogen) were used to recover the hybridized library fragments. After washing and eluting the libraries, the post-capture PCR amplification was performed with KAPA uracil + polymerase (Kapa Biosystems). PCR cycling conditions were as follows: 1 cycle of 3 min at 95 °C, followed by 14 cycles of: 20 s denaturation at 98 °C, 15 s annealing at 65 °C and 30 s elongation at 72 °C, ending with 5 min at 72 °C. The amplified captured libraries were purified using AMPureXP beads (Agencourt).

Shotgun sequencing data were generated as previously described¹. Sequencing of target-enriched libraries was performed on Illumina HiSeq2500 SR80bp, V4 chemistry.

The resulting reads were compared to dataset 2 using BLASTn (with arguments -task blastn -evalue 0.01). Matching reads with bit scores greater than 50 for all samples (except DA222 (70) and DA45 (55)) were selected for subsequent processing. In total, 6,757 reads matched HBV in the capture data.

Sequence authenticity. The following evidence leads us to believe that the ancient HBV sequences are authentic and that the possibility of contamination can be excluded.

(1) Standard precautions for working with aDNA were applied⁴⁴.

(2) Sequences were checked for typical aDNA damage patterns using mapDamage⁴⁵ (version 2.0.6). Whenever sufficient amounts of data were available (> 200 HBV reads), we found C > T mutations at the 5' end, typical of aDNA⁴⁶ (see Extended Data Fig. 1a, c).

(3) Capture was performed on sample DA222 DNA extracts with and without pre-treatment by uracil-specific excision reagent (USER)⁴⁷. After USER treatment (3 h at 37 °C) of the aDNA extract, the damage pattern is eliminated (Extended Data Fig. 1b).

(4) As the ancient viruses are from three different HBV genotypes (A, B and D) and a clade in sister relationship to chimpanzee and gorilla HBVs, any argument

that samples were contaminated would have to account for this diversity as well as the sequence novelty.

(5) HBV sequences were identified in 25 of 304 analysed samples (Table 1), showing that the findings cannot be due to a ubiquitous laboratory contaminant.

(6) Despite the low frequency of positive samples, we sequenced extraction blanks to provide additional evidence against the possibility that the HBV sequences stemmed from sporadic incorporation, amplification and sequencing of background reagent contaminants into the aDNA libraries. The negative extraction controls were amplified for 40 PCR cycles, and BLAST was used to match the read sequences against dataset 3, with the same parameters used for the ancient samples. Because the ancient HBV-positive reads used to assemble genomes all had bit scores of at least 50 (see 'Data and data processing'), we filtered the negative extraction control BLAST output for reads with a bit score ≥ 45 . No reads (out of 23 million) matched any HBV genome at that level.

(7) HBV is a blood-borne virus that is mainly transmitted by exposure to infectious blood and that does not occur in the environment³, making contamination during archaeological excavation extremely unlikely.

Consensus sequences. Reads from the original sequencing and from the capture were aligned to a reference genome (Supplementary Table 3) in Geneious⁴⁸ (version 9) using medium sensitivity/fast and iterate up to 5 times. Because aDNA damage often clusters towards read termini⁴⁶, the resulting alignments were carefully curated by hand to remove non-matching termini of reads if the majority of the read showed a very good match with the reference sequence.

Genotyping. All reads used to construct the ancient HBV consensus sequences were matched against the full NCBI nucleotide database (downloaded 28 December 2016) using BLAST. Of these reads, 97.5% had HBV as their top match. All ancient consensus sequences were matched against the full HBV genomes of dataset 4 with the Needleman–Wunsch algorithm⁴⁹, as implemented in EMBOSS⁵⁰ (version 6.6.0.0). For each ancient sequence, the percentage of sequence identity with the most similar representative of each modern genotype and four non-human primate species is listed in Extended Data Table 4a. The Needleman–Wunsch algorithm was also used to calculate the pairwise sequence similarity between all ancient sequences (Extended Data Table 4b).

Recombination analysis. The recombination detection program⁵¹ version 4 (RDP4) was used to search for evidence of recombination within the 12 ancient sequences and a selection of 15 modern human and non-human primate sequences (Supplementary Methods). Recombination with HBV-RISE387 as the recombinant and HBV-DA51 as one parent, was suggested at positions 1567–2256, by seven recombination methods (RDP⁵², GENECONV⁵³, BootScan⁵⁴, MaxChi⁵⁵, Chimaera⁵⁶, SiScan⁵⁷ and 3Seq⁵⁸) with *P* values from 1.179×10^{-6} to 5.336×10^{-11} (Extended Data Table 2). The same recombination was suggested for all 4 ancient genotype A and two modern genotype A sequences. Graphical evidence of the recombination and the predicted break point distribution for sequences HBV-RISE386 and HBV-RISE387 from three methods (MaxChi, Bootscan and RDP) is shown in Extended Data Fig. 3.

Initial maximum likelihood phylogenies. An initial maximum likelihood tree was generated to ascertain whether the ancient sequences fall within the primate HBV clades. Dataset 1 and the ancient sequences were aligned in MAFFT⁵⁹ (version 7). The maximum likelihood tree was constructed using PhyML⁶⁰ (version 20160116), optimizing topology, branch lengths and rates. We used a general time reversible (GTR) substitution model, with base frequencies determined by maximum likelihood, and a maximum likelihood-estimated proportion of invariant sites and 100 bootstraps (Extended Data Fig. 2). Furthermore, a maximum likelihood tree (Extended Data Fig. 4) was generated based on a MAFFT alignment of dataset 2 and the ancient sequences, using the same parameters as above.

Dated coalescent phylogenies. To check for a temporal signal in the data, root-to-tip regressions and date randomization tests were performed. For the root-to-tip regression, input trees were calculated using dataset 2 with the addition of a woolly monkey sequence (GenBank Accession Number: AF046996) as an outgroup. Three phylogenetic algorithms were used; neighbour joining, maximum likelihood (PhyML), and Bayesian (MrBayes⁶¹ (version 3.2.5)) methods (Supplementary Figs. 1–3). Root-to-tip distances were extracted using TempEst⁶² (version 1.5). For maximum likelihood and Bayesian methods, root-to-tip distances (in substitutions per site) were extracted from optimized tree topologies (maximum likelihood and maximum clade credibility trees, respectively). For the neighbour joining method, root-to-tip distances were averaged over 1,000 bootstrap replicates. Regression analyses were performed with Scipy (version 0.16.0; <http://www.scipy.org>). For the date randomization tests, we used three different approaches to randomize tip dates. First, tip dates were randomized between all sequences in the phylogeny. Second, tip dates were randomized only among the ancient sequences presented in this Letter, as well as the Korean mummy sequence (GenBank accession number JN315779), while the modern sequences retained their correct ages. Third, dates were randomized within a clade. For each of the three approaches, we performed three independent randomizations. This resulted

in a total of nine analyses, which were run for 100,000,000 generations each, under the relaxed log-normal clock model and coalescent exponential tree prior. We also ran the same analyses under a strict clock and coalescent Bayesian skyline tree prior, which were run for 20,000,000 generations. We used a GTR substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation and estimated proportion of invariant sites, as found by bModelTest⁶³ (version 1.0.4). None of the analyses using the relaxed clock converged (estimated sample size < 200). This is most probably because the mis-specification of the dates leads to incongruence between the sequence and time information. Under the strict clock model, all runs converged, and none of the 95% HPD intervals of the root age overlapped between the randomized and the non-randomized runs, fulfilling the criteria for evidence of a temporal signal⁶⁴.

Dated phylogenies were estimated using BEAST2²⁵ (version 2.4.4, prerelease). We used a MAFFT alignment of dataset 2. Using bModelTest⁶³, we selected a GTR substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation and estimated proportion of invariant sites. Proper priors were used throughout. Path sampling, as implemented in BEAST2, was performed to select between a strict or relaxed log-normal clock and a coalescent constant, exponential or coalescent Bayesian skyline tree prior (Extended Data Table 3a). Likelihood values were compared using a Bayes factor test. A Bayes factor in the range of 3–20 implies positive support, 20–150 strong support and > 150 overwhelming support⁶⁵. The relaxed log-normal clock model in combination with a coalescent exponential tree prior was favoured (Extended Data Table 3a). For the final tree, a Markov chain Monte Carlo analysis was run until parameters reached an estimated sample size > 200, sampling every 2,000 generations. Convergence and mixing were assessed using Tracer⁶⁶ (version 1.6). The final tree files were subsampled to contain 10,000 or 10,710 (for the relaxed log-normal clock, coalescent exponential tree prior) trees, with the first 25% of samples discarded as burn-in. Maximum clade credibility trees were made using TreeAnnotator²⁵ (version 2.4.4 prerelease).

To formally test the 'Out of Africa' hypothesis of HBV evolution, calibration points were tested using path sampling as implemented in BEAST2. Calibration points were constrained as follows. For the split of genotypes F and H, the most-recent common ancestor (MRCA) of all genotype F and H sequences was constrained using a uniform (13,400:25,000) distribution, as this is the range of estimates for when the Americas were first colonized^{67,68}. For the split of subgenotype A3 in Haiti, the MRCA of FJ692598 and FJ692611 was constrained using a uniform (200:500) distribution, owing to the timing of the slave trade to Haiti⁶⁹. For the split of C3 in Polynesia, the MRCA of X75656 and X75665 was constrained using a uniform (5,100:2,000) distribution, owing to the range of estimates for the MRCA of Polynesian populations^{11,70}. Calibration points were tested under both a relaxed log-normal clock, coalescent exponential tree prior, and a strict clock, Bayesian skyline tree prior.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The complete sequences in this study have been deposited in the European Nucleotide Archive under sample accession numbers ERS2295383–ERS2295394. All other data are available from the corresponding author upon reasonable request.

34. Drexler, J. F. et al. Bats carry pathogenic hepadnaviruses antigenically related to hepatitis B virus and capable of infecting human hepatocytes. *Proc. Natl Acad. Sci. USA* **110**, 16151–16156 (2013).
35. Geer, L. Y. et al. The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492–D496 (2010).
36. Bell, T. G., Younis, M. & Kramvis, A. Bioinformatic curation and alignment of genotyped hepatitis B virus (HBV) sequence data from the GenBank public database. *Springerplus* **5**, 1896 (2016).
37. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
38. Reimer, P. J. et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
39. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337 (2012).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
42. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
43. Drosten, C., Weber, M., Seifried, E. & Roth, W. K. Evaluation of a new PCR assay with competitive internal control sequence for blood donor screening. *Transfusion* **40**, 718–724 (2000).
44. Willerslev, E. & Cooper, A. Review Paper. Ancient DNA. *Proc. R. Soc. Lond. B* **272**, 3–16 (2005).
45. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).

46. Orlando, L., Gilbert, M. T. P. & Willerslev, E. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* **16**, 395–408 (2015).
47. Briggs, A. W. et al. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
48. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
49. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
50. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
51. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).
52. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563 (2000).
53. Padidam, M., Sawyer, S. & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225 (1999).
54. Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* **21**, 98–102 (2005).
55. Smith, J. M. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129 (1992).
56. Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl Acad. Sci. USA* **98**, 13757–13762 (2001).
57. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582 (2000).
58. Boni, M. F., Posada, D. & Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047 (2007).
59. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
60. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
61. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
62. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
63. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17**, 42 (2017).
64. Duchêne, S., Duchêne, D., Holmes, E. C. & Ho, S. Y. W. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* **32**, 1895–1906 (2015).
65. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
66. Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. Tracer v1.6. <https://github.com/beast-dev/tracer/releases/tag/v1.6> (2017).
67. Sanchez, G. et al. Human (Clovis)–gomphothere (*Cuvieronius* sp.) association ~ 13,390 calibrated yBP in Sonora, Mexico. *Proc. Natl Acad. Sci. USA* **111**, 10972–10977 (2014).
68. Bourgeon, L., Burke, A. & Higham, T. Earliest human presence in North America dated to the Last Glacial Maximum: new radiocarbon dates from Bluefish Caves, Canada. *PLoS ONE* **12**, e0169486 (2017).
69. Andernach, I. E., Nolte, C., Pape, J. W. & Muller, C. P. Slave trade and hepatitis B virus genotypes and subgenotypes in Haiti and Africa. *Emerg. Infect. Dis.* **15**, 1222–1228 (2009).
70. Kayser, M. et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23**, 2234–2244 (2006).

Extended Data Fig. 1 | aDNA damage patterns. The frequencies of the mismatches observed between the HBV reference sequences (Supplementary Table 3) and the reads are shown as a function of distance from the 5' end. C > T (5') and G > A (3') mutations are shown in red and blue, respectively. All other possible mismatches are shown in grey. Insertions are shown in purple, deletions in green and clippings in orange. The count of reads matching HBV for each sample is shown in

parentheses. **a**, Damage patterns for RISE563, DA222, DA119, RISE254, DA195, DA27, DA51, RISE386, RISE387, DA29, DA45, RISE416 and RISE154. **b**, Damage patterns for DA222 without (left) and with (right) USER treatment. **c**, Damage patterns with 10, 20, 50, 100, 200, 500 and 1,000 reads sampled from RISE563, in which each opaque line corresponds to one replicate set of reads.

Extended Data Fig. 2 | *Hepadnaviridae* maximum likelihood tree. This figure shows 26 *Orthohepadnaviridae* sequences (dataset 1, see Methods), including the ancient HBV sequences. Ancient genotype A sequences are shown in red, the ancient genotype B sequence in orange, ancient

genotype D sequences in blue and novel genotype sequences in green. The tree was constructed in PhyML⁶⁰, optimizing for topology, branch lengths and rates, with 100 bootstraps (see Methods). Internal nodes with < 70% bootstrap support are shown as polytomies.

Extended Data Fig. 3 | Genotype A recombination break-point evidence. RDP4⁵¹ was used to analyse the set of 12 ancient sequences plus a representative set of 15 modern human and non-human primate sequences (see Methods). The seven recombination programs used by RDP4 suggested that all genotype A sequences are recombinants, with the genotype D sequence HBV-DA51 as the minor parent and an unknown major parent. The obvious interpretation is that recombination formed an ancestor of the oldest sequences, evidence of which is still present in the less-ancient and the modern representatives. The figure shows the graphical evidence and predicted recombination break-point distribution

for the two oldest genotype A sequences, HBV-RISE386 and HBV-RISE387, according to three of the RDP4 methods (MaxChi, Bootscan and RDP). In all subplots, the predicted location of the break points is shown as a dashed vertical line and the surrounding grey area shows the 99% confidence interval for the break point. Subplots on the same row share their y axis and those in the same column share their x axis. **a**, HBV-RISE386 analysed by MaxChi. **b**, HBV-RISE386 analysed by Bootscan. **c**, HBV-RISE386 analysed by RDP. **d**, HBV-RISE387 analysed by MaxChi. **e**, HBV-RISE387 analysed by Bootscan. **f**, HBV-RISE387 analysed by RDP.

Extended Data Fig. 4 | HBV maximum likelihood tree. The sequences from dataset 2 (see Methods) and the ancient sequences were aligned in MAFFT⁵⁹. The tree was constructed in PhyML⁶⁰, optimizing for topology, branch lengths and rates, with 100 bootstraps (see Methods). Internal nodes with < 70% bootstrap support are shown as polytomies. Ancient genotype A sequences are shown in red, ancient genotype B sequences in orange, ancient genotype D sequences in blue and novel genotype sequences in green. Taxon names indicate: genotype or subgenotype, GenBank accession number, age, abbreviation of country of sequence origin, region of sequence origin, host species and optional additional

remarks. Note that the maximum likelihood tree shows topological uncertainty (polytomies) in areas where the BEAST2²⁵ tree (Fig. 2) is well resolved. This is the case for two reasons. First, BEAST2 always produces a fully resolved binary topology without polytomies. Second, and more important, BEAST2 creates a time tree and uses tip dates to constrain the possible topologies under consideration. Thus, BEAST2 can know that certain topologies are unlikely or impossible, whereas maximum likelihood cannot and thus inherently has greater uncertainty regarding tree topology.

Extended Data Fig. 5 | Root-to-tip regression and date randomization tests. **a**, Regression of root-to-tip distances and ages performed in Scipy (<http://www.scipy.org>). One hundred and twenty-four branch lengths were extracted using TempEst⁶² from trees inferred using neighbour joining, maximum likelihood and Bayesian methods. Shaded areas show 95% confidence intervals. Slopes are 1.01×10^{-5} , 1.20×10^{-5} and 4.21×10^{-6} , and correlation coefficients are 0.45 ($R^2=0.2$), 0.36 ($R^2=0.13$) and 0.51 ($R^2=0.26$), for maximum likelihood, Bayesian and neighbour joining

trees, respectively. **b**, Date randomization tests under the strict clock model. The median and 95% HPD interval for the substitution rates are given. The rate for the correctly dated tree is shown in red. Dates were randomized within all sequences, within the ancient sequences only, and within each genotype. We performed three replicates of each. None of the 95% HPD intervals for the randomized runs overlaps with the 95% HPD intervals for the correctly dated runs, suggesting the presence of a temporal signal in the data.

Extended Data Table 1 | Extended overview of samples with reads matching HBV and PCR results

a

Sample	¹⁴ C age (standard deviation)	Median cal BP age, or estimate (in years)	Approx. sample age (in years)	Site	Culture or period	Sex	Original reads hit DIA- MOND	Proteins hit DIA- MOND	Original reads hit BLASTn	Captured reads hit BLASTn	Bit score cut-off	Consen- sus	Coverage Depth
RISE563	3955 (35)	4421	4488	Osterhofen- Altenmarkt, GER	Bell Beaker	M	1499	7	4817	N/D	50	100%	79.3x
DA222	N/D	1200-1000	1167	Butakty, KAZ	Karluk	M	2269	7	2855	1318	70	100%	89.2x
DA195	2479 (35)	2578	2645	Sandorfalva- Eperjes, HUN	Hungarian Scythian	F	353	7	694	774	50	99.9%	29.2x
DA51	2220 (37)	2230	2297	Keden, KYR	Saka	M	54	7	127	595	50	99.2%	14.5x
RISE254	3631 (29)	3942	4009	Százhalombatta- Földvár, HUN	Vatya	M	158	7	211	1304	50	99.0%	36.6x
DA119	N/D	1500	1567	Poprad, SVK	North Carpathian	M	155	7	385	2272	50	98.8%	53.1x
RISE386	3758 (34)	4121	4188	Bulanovo, RUS	Sintashta	M	94	7	240	140	50	97.8%	7.0x
DA27	1641 (33)	1543	1610	Halvay 3, KAZ	Hun-Sarmatian	M	309	7	890	8	50	90.0%	14.3x
DA29	849 (25)	755	822	Karasyur, KAZ	Golden Horde	M	149	7	179	44	50	87.5%	4.8x
DA45	2083 (27)	2053	2120	Omnogobi, MNG	Xiongnu	M	29	4	110	118	55	87.2%	4.3x
RISE387	3822 (33)	4215	4282	Bulanovo, RUS	Sintashta	N/D	154	7	344	N/D	50	86.6%	6.2x
RISE416	3259 (40)	3488	3555	Nerquin Getashen, ARM	Middle Bronze Age	M	4	6	14	137	50	75.8%	3.5x
RISE154	3522 (24)	3784	3851	Szczepankowice, POL	Unetice	F	17	7	137	N/D	50	57.2%	2.0x
DA85	1781 (46)	1702	1769	Japyrik, KYR	Hun	M	14	4	27	6	50	35.8%	0.5x
RISE391	3612 (34)	3921	3988	Tanabergen II, KAZ	Sintashta	F	18	7	42	N/D	50	34.6%	0.9x
RISE568	N/D	1350-1150	1317	Brandýsek, CZE	Early Slav	F	6	7	7	9	50	33.9%	0.4x
RISE556	N/D	4500-4000	4317	Manching- Oberstimm, GER	Bell Beaker	F	2	2	N/D	32	50	28.0%	0.6x
DA89	1315 (45)	1247	1314	Berygavoya, KAZ	Turk	M	5	2	51	0	50	17.9%	1.0x
RISE548	N/D	5000-4300	4717	Temrta IV, RUS	Yamnaya	M	5	5	19	N/D	50	17.9%	0.5x
RISE533	3657 (37)	3980	4047	Arano, ITA	Early Bronze Age	N/D	3	6	N/D	N/D	N/D	N/D	N/D
RISE99	3713 (30)	4045	4112	L Bedinge, SWE	Nordic Late Neolithic	M	2	2	N/D	N/D	N/D	N/D	N/D
RISE478	3164 (34)	3391	3458	Schöngesing, GER	Bronze Age	M	2	3	N/D	N/D	N/D	N/D	N/D
RISE509	4186 (27)	4728	4795	Bateni, RUS	Afanasievo	F	2	2	N/D	N/D	N/D	N/D	N/D
RISE554	2782 (30)	2879	2946	Afontova Gora, RUS	Late Bronze Age hunter-gatherer	M	1	2	N/D	N/D	N/D	N/D	N/D
RISE597	N/D	3500-3000	3317	Turlojske, LTU	Late Bronze Age	M	1	3	N/D	N/D	N/D	N/D	N/D

b

Sample	% of HBV reads per sample	C _t -value	HBV DNA qPCR conclusion	Remarks
DA222	7.45E-05	31	Positive	Male
DA195	2.82E-05	40.52	Positive	Female
DA85	6.41E-07	>45	Negative	
DA89	2.85E-06	>45	Negative	
DA351 (neg.)	NA	>45	Negative	
Negative ancient samples (n=4)	NA	>45	Negative	
NTC (n=8)	NA	>45	Negative	
Enriched HBVpos library pool 1	NA	>40	Positive	
Enriched HBVpos library pool 2	NA	35.39	Positive	
Enriched HBVneg library pool 3	NA	>45	Negative	
Enriched HBVneg library pool 4	NA	>45	Negative	

a, Extended overview of samples with reads matching HBV. Rows are sorted by decreasing consensus coverage. Original reads hit DIAMOND, number of sequencing reads that matched HBV using DIAMOND⁴¹; Proteins hit DIAMOND, number of HBV proteins matched by HBV sequencing reads using DIAMOND; Original reads hit BLASTn, number of sequencing reads that matched HBV using a BLASTn⁴² database built from dataset 3 (see Methods); Captured reads hit BLASTn, the number of reads from the capture that matched HBV using BLASTn (as above); Bit score cut-off, the bit score cut-off above which matching reads were used to form consensus sequences; coverage consensus and depth, the percentage of the consensus genome covered by matching reads and average depth of coverage across the reference genome as reported by Geneious⁴⁸. When reading sample information across a row, an N/D (not determined) cell will be encountered when processing on that sample was concluded, either owing to too few matching reads or consensus coverage less than 50%. **b**, TaqMan PCR results. Four extracts from samples with HBV reads were selected for TaqMan PCR confirmation: two with a large proportion of HBV reads (DA222 and DA195), two with a small proportion of HBV reads (DA85 and DA89) and one with no HBV reads (DA351). HBV was detected in extracts from DA222 and DA195, whereas the three low- and zero-read samples were negative, as were all no-template controls.

Extended Data Table 2 | Genotype A predicted recombination break points and *P* values**a**

Method	Seqs	Av. <i>P</i> -value
RDP	6	1.641E-06
GENECONV	6	2.450E-06
BootScan	6	1.179E-06
MaxChi	6	3.494E-06
Chimaera	6	2.332E-06
SiScan	6	5.336E-11
3Seq	6	7.319E-08

b

Sample	Begin	Begin 99% CI	End	End 99% CI
RISE387	1567	1498-1638	2256	2130-2937
RISE386	1593	1498-1638	2248	2130-2937
DA195	1633	1498-1638	2247	2130-2937
DA119	1633	1498-1638	2247	2130-2937
KJ854705	1642	1498-1638	2247	2130-2937
LC074724	1622	1498-1638	2256	2130-2937

a, The *P* values assigned to the predicted genotype A recombination by the seven methods used by RDP4⁵¹, in the order given by RDP4. The number of sequences in which the recombination was predicted is always six, corresponding to the four ancient and two modern genotype A sequences. **b**, The predicted start and end break points for each of the six genotype A sequences. Sequences are ordered from oldest to youngest. The 99% confidence intervals for the start and end points are shown ($n = 15$ sequences analysed in all cases), and are identical for all sequences. The predicted break points are close to the boundaries of the polymerase. For example, for the modern genotype A sequence LC074724, the polymerase is found in regions 1–1623 and 2307–3221 and the predicted break points are 1622 and 2256. If recombination formed an HBV-RISE387/HBV-RISE386 ancestor, it is possible that the entire polymerase gene was contributed by one parent.

Extended Data Table 3 | Model testing and inferred age of genotypes

a

	Model (clock, coalescent)	Strict, Constant	Strict, Exponential	Strict, Skyline	Relaxed-log., Constant	Relaxed-log., Exponential	Relaxed-log., Skyline
Model (clock, coalescent)	Likelihood	-50556.67815	-50503.87888	-50499.63034	-50500.28013	-50324.79502	-50334.43366
Strict, Constant	-50556.67815	0	52.79926971	57.04780837	56.39801939	231.8831256	222.2444808
Strict, Exponential	-50503.87888	-52.79926971	0	4.248538656	3.598749675	179.0838559	169.4452111
Strict, Skyline	-50499.63034	-57.04780837	-4.248538656	0	-0.649788981	174.8353173	165.1966724
Relaxed-log., Constant	-50500.28013	-56.39801939	-3.598749675	0.649788981	0	175.4851062	165.8464614
Relaxed-log., Exponential	-50324.79502	-231.8831256	-179.0838559	-174.8353173	-175.4851062	0	-9.638644846
Relaxed-log., Skyline	-50334.43366	-222.2444808	-169.4452111	-165.1966724	-165.8464614	9.638644846	0

b

	Median MRCA in years (95% HPD interval)	
Clade	Strict, Skyline	Relaxed-log., Exponential
A	8221 (7438, 9076)	6699 (5566, 8148)
B	4883 (4264, 5556)	4590 (3468, 6025)
C	5501 (4782, 6356)	4886 (3499, 6534)
D	4158 (3669, 4683)	3921 (3077, 5044)
E	1016 (712, 1358)	857 (331, 1670)
G	181 (96, 287)	173 (70, 334)
F	4737 (4053, 5538)	4068 (2698, 5679)
H	1713 (1365, 2109)	1539 (832, 2498)
F/H	7341 (6224, 8603)	6040 (4001, 8637)
Chimp./Gorilla	5281 (4469, 6146)	4523 (2653, 6508)
Gibbon/J/ Orang.	7391 (6384, 8484)	5604 (3852, 7700)
RISE154/RISE254 /RISE563	8239 (7502, 9141)	6716 (5626, 8078)

c

Model (clock, coalescent)	Median root age in years (95% HPD interval)	Median substitution rate (substitutions/site/year)
Strict, Constant	16038 (14023, 18140)	9.27E-6 (8.14E-6, 1.05E-5)
Strict, Exponential	15840 (13969, 18056)	9.24E-6 (8.04E-6, 1.04E-5)
Strict, Skyline	15593 (13688, 17795)	9.48E-6 (8.3E-6, 1.07E-5)
Relaxed-log. Constant	14561 (9551, 20940)	1.16E-5 (8.63E-6, 1.45E-5)
Relaxed-log. Exponential	11623 (8613, 15275)	1.18E-5 (9.21E-6, 1.45E-5)
Relaxed-log. Skyline	12923 (8762, 18415)	1.23E-5 (9.68E-6, 1.51E-5)

d

	Strict, Skyline		Relaxed-log., Exponential	
Hypothesis	Likelihood	Bayes factor	Likelihood	Bayes factor
A3	-50502.9358	3.305462052	-50388.15232	63.35730355
C3	-50532.75485	33.12451278	-50388.13331	63.33829434
F/H	-50493.15396	-6.476375522	-50388.046	63.25097997

Models were compared using path sampling, as implemented in BEAST2²⁵. Likelihood values were compared using a Bayes factor test. According to previous work⁶⁵, a Bayes factor in the range of 3–20 implies positive support, 20–150 strong support and > 150 overwhelming support. **a**, Results of testing different clock models and population assumptions to be used for dated phylogenies. Positive numbers indicate support for the column model, and negative numbers indicate support for the row model. **b**, Median MRCA age of individual nodes under a strict clock and Bayesian skyline tree prior or under a relaxed log-normal clock and coalescent exponential tree prior. **c**, Median root age and substitution rates under different clock models and tree priors. **d**, Results of testing different calibration point hypotheses under a strict clock and Bayesian skyline tree prior or under a relaxed log-normal clock and coalescent exponential tree prior. The extreme maximum and minimum values for root age and substitution rate across all models are used for the numbers given in the abstract of this Letter.

Extended Data Table 4 | Consensus sequence identity**a**

	A	B	C	D	E	F	G	H	I	J	Chimp	Gibbon	Gorilla	Orang Utan
DA119	97.27	92.60	92.94	92.38	92.05	88.42	90.62	88.24	93.32	89.78	92.10	91.15	92.95	90.98
DA195	96.11	92.16	92.59	92.06	92.00	88.24	90.32	87.68	93.22	89.37	92.13	91.65	92.32	90.99
RISE386	95.32	92.55	92.96	92.83	91.87	88.57	90.22	87.94	93.79	89.60	92.33	91.43	92.64	91.34
RISE387	92.49	90.70	91.70	91.74	91.06	88.57	91.42	88.60	91.54	88.96	91.55	90.75	91.37	90.65
DA45	91.89	97.65	96.04	90.47	90.55	87.70	88.11	88.06	91.07	89.40	91.50	91.79	91.58	90.53
DA27	91.67	91.28	92.96	97.21	93.15	88.77	89.93	87.83	91.08	89.63	92.43	92.24	92.57	90.31
DA29	91.46	90.50	92.79	98.64	93.21	87.82	89.41	87.64	90.82	88.66	91.10	90.74	91.94	89.66
DA51	92.10	91.32	93.24	96.99	93.47	88.48	90.28	88.04	91.11	89.30	91.55	91.54	92.26	90.40
DA222	92.29	91.61	93.88	98.74	93.55	88.66	90.01	88.23	91.45	89.23	92.16	91.99	93.29	90.63
RISE154	90.86	91.07	92.11	91.76	91.81	89.20	91.26	88.91	91.62	90.05	92.54	92.36	92.26	92.36
RISE254	92.04	91.66	92.32	92.56	92.80	89.25	91.20	88.86	91.49	90.47	94.03	93.25	93.93	92.96
RISE563	90.63	90.43	90.70	91.41	91.27	88.83	90.51	88.75	90.27	89.16	92.30	91.94	92.66	91.08

b

	DA195	RISE386	RISE387	DA45	DA27	DA29	DA51	DA222	RISE154	RISE254	RISE563
DA119	96.79	95.99	92.69	92.08	91.87	91.97	92.37	92.41	91.39	92.23	90.78
DA195		95.86	92.72	91.83	92.15	91.50	92.47	91.83	91.34	91.91	90.36
RISE386			93.34	92.34	92.74	92.55	92.97	92.64	91.66	92.63	91.07
RISE387				90.60	91.50	91.33	91.89	91.75	91.78	92.12	91.26
DA45					89.73	89.39	90.75	90.38	90.43	90.97	89.66
DA27						96.42	97.24	97.31	91.45	92.49	91.39
DA29							96.62	98.06	90.72	91.52	90.48
DA51								97.12	91.83	92.66	91.44
DA222									91.57	92.53	91.28
RISE154										94.39	93.02
RISE254											95.77

a. Best consensus sequence identity with 14 groups of HBV full genomes. The Needleman–Wunsch algorithm (as implemented in EMBOSS⁵⁰) was used to globally align each sample consensus sequence against each of the 3,384 full HBV genomes of dataset 4 (see Methods). The table shows the best nucleotide similarity percentage for each sample consensus against 14 genome groups from the full set of HBV genomes. In cases in which the consensus length is less than the genome length, the given figure is the percentage of identical nucleotides in the matching region, not counting any alignment gaps or ambiguous consensus nucleotides. For each sample, the genome group with the highest identity is highlighted in bold. **b.** Inter-consensus sequence identity. The Needleman–Wunsch algorithm was used to globally align all ancient sample consensus sequences against one another. The table shows the nucleotide identity percentage for each alignment. In cases in which the consensus lengths were unequal, the given figure is the percentage of identical nucleotides in the matching region, not counting any alignment gaps or ambiguous consensus nucleotides.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection.

Data analysis

TempEst (version 1.5), Scipy (version 0.16.0), bModelTest (version 1.0.4), BEAST2 (version 2.4.4, prerelease), Geneious (version 9), EMBOSS (version 6.6.0.0), Recombination Detection Program (version 4), mapDamage (version 2.0.6), OxCal (version 4.3), AdapterRemoval (version 2.1.7), BWA (version 0.7.15-r1140), MAFFT (version 7), PhyML (version 20160116), MrBayes (version 3.2.5), Tracer (version 1.6), TreeAnnotator (version 2.4.4 prerelease), DIAMOND (version 0.8.25), BLAST (version 2.4.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The complete sequences in this study have been deposited in the European Nucleotide Archive under sample accession numbers ERS2295383-ERS2295394.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was required. All available human DNA samples were analyzed. Phylogenetic analyses include 124 sequences in total. Those were selected to represent the diversity of available HBV sequences.
Data exclusions	Samples with HBV genome coverage below 50% were excluded from detailed analysis. This coverage level was pre-established.
Replication	Phylogenetic trees were reproduced with several algorithms (maximum likelihood, Bayes, neighbour joining). The temporal signal in the data was found consistently using trees from all three of these algorithms and also a date randomization test. Reproducibility was successful in all cases.
Randomization	N/A (no experimental groups were used).
Blinding	N/A (as above).

Materials & experimental systems

Policy information about [availability of materials](#)

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging