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

Mühlemann, Barbara; Jones, Terry C.; Damgaard, Peter de Barros; Allentoft, Morten E.; Shevnina, Irina; Logvin, Andrey; Usmanova, Emma; Panyushkina, Irina P.; Boldgiv, Bazartseren; Bazartseren, Tsevel

Published in:
Nature

Link to article, DOI:
10.1038/s41586-018-0097-z

Publication date:
2018

Document Version
Peer reviewed version

Ancient Hepatitis B viruses from the Bronze Age to the Medieval

Barbara Mühlemann[*,1], Terry C. Jones[*,1,2], Peter de Barros Damgaard[*,3], Morten E. Allentoft[*,3], Irina Shevnina[4], Andrey Logvin[4], Emma Usmanova[5], Irina P. Panyushkina[6], Bazartseren Boldgiv[7], Tsevel Bazartseren[8], Kadicha Tashbaeva[9], Victor Merz[10], Nina Lau[11], Václav Smrčka[12], Dmitry Voyakin[13], Egor Kitov[14], Andrey Epimakhov[15], Dalia Pokutta[16], Magdolna Vicze[17], T. Douglas Price[18], Vyacheslav Moiseyev[19], Anders J. Hansen[3], Ludovic Orlando[3,20], Simon Rasmussen[21], Martin Sikora[3], Lasse Vinner[3], Albert D. M. E. Osterhaus[22], Derek J. Smith[1], Dieter Glebe[23,24], Ron A. M. Fouchier[25], Christian Drosten[2,26], Karl-Göran Sjögren[18], Kristian Kristiansen[18], Eske Willerslev[§,3,27,28].

1. Center for Pathogen Evolution, Department of Zoology, University of Cambridge, Downing St., Cambridge CB2 3EJ, UK. 2. Institute of Virology, Charité, Universitätsmedizin Berlin, Campus Charité Mitte, Charitéplatz 1, 10117 Berlin, Germany. 3. Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark. 4. Archaeological Laboratory, Faculty of History and Law, A. A. Baitursynov Kostanay State University, 47 Baitursynov St., Kostanay, 110000, Kazakhstan 5. Karaganda State University, Saryarka Archaeological Institute, 28 Universitetskaya St., 100012 Karaganda, 100012, Kazakhstan. 6. Laboratory of Tree-Ring Research, University of Arizona, 1215 E. Lowell St., Tucson, AZ 85721, USA. 7. Department of Biology, School of Arts and Sciences, National University of Mongolia, Ulaanbaatar 14201, Mongolia. 8. Laboratory of Virology, Institute of Veterinary Medicine, Mongolian University of Life Sciences, Ulaanbaatar 17024, Mongolia. 9. National Academy of Sciences, 265a Chuy Ave., Bishkek, 720001, Kyrgyzstan. 10. Pavlodar State University, 64 Lomov Str., Pavlodar, 140008, Kazakhstan. 11. Centre for Baltic and Scandinavian Archaeology, Schloss Gottorf, D 24837 Schleswig, Germany. 12. Institute for History of Medicine and Foreign Languages of the First Faculty of Medicine, Charles University, 32 Kateřinská, Prague, 121 08, Czech Republic. 13. Margulan Institute of Archaeology, 71 Zenkov St., Almaty, 050010, Kazakhstan. 14. N. N. Miklouho-Maklay Institute of Ethnology and Anthropology, Russian Academy of Sciences, 32a Leninsky Ave., Moscow, 1199111, Russia. 15. South Ural Department, Institute of History and Archaeology UBRAS, South Ural State University, 76 Lenina St., Chelyabinsk, 454080, Russia. 16. Department of Archaeology and Classical Studies, Stockholm University, Universitetsvägen 10, Stockholm, 11418, Sweden. 17. Matrica Museum, 1-3 Gesztenyés St., Százhalombatta, 2440, Hungary. 18. Department of Historical Studies, University of Gothenburg, Eklandagatan 86, 412 61 Göteborg, Sweden. 19. Department of

Physical Anthropology, Peter the Great Museum of Anthropology and Ethnography, 2 Universitetskaya Naberazhnaya, Saint-Petersburg, 199034, Russia. 20. Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, CNRS UMR 5288, Toulouse, 31000, France. 21. Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kgs Lyngby, Denmark. 22. Research Center for Emerging Infections and Zoonoses, University of Veterinary Medicine Hannover, Bünteweg 17, Hannover, 30559, Germany. 23. Institute of Medical Virology, Justus Liebig University of Giessen, Schubertstrasse 81, Giessen, D 35392, Germany. 24. National Reference Centre for Hepatitis B and D Viruses, German Center for Infection Research (DZIF), Schubertstrasse 81, D 35392 Giessen, Germany. 25. Department of Viroscience, Erasmus Medical Centre, Wytemaweg 80, Rotterdam, 3015 CN, Netherlands. 26. German Center for Infection Research (DZIF), Inhoffenstraße 7, Braunschweig, 38124, Germany. 27. Cambridge GeoGenetics Group, Department of Zoology, Downing St., University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. 28. Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK.

* These authors contributed equally to this work.

§ To whom correspondence should be addressed.

Abstract

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 between ~0.8-4.5 thousand years old. The ancient sequences group either within or in sister relationship to 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 8.6-20.9 thousand years ago (kya), and estimate a substitution rate between $8.04x10^{-6}$-$1.51x10^{-5}$ nucleotide substitutions per site per year (s/s/y). In several cases, the geographic 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 geographic and temporal patterns we observe in ancient and modern HBV genotypes are compatible with well-documented human migrations during the Bronze and Iron Ages[1,2]. We show evidence for the creation of genotype A via recombination and a long-term association of modern HBV genotypes with humans, including the discovery of a human genotype that is now extinct. Taken together, the 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 fluids[3]. The estimated global prevalence is 3.6%, ranging from 0.01% (UK) to 22.38% (South Sudan)[4]. In high endemicity areas, where prevalence is > 8%, 70-90% of the adult population show evidence of past or present infection[5,6]. The young and the immunocompromised are most likely to develop chronic HBV infection, which can

result in high viremia over years to decades[3]. Approximately 257 million people are chronically infected[5] and around 887,000 died in 2015 due to associated complications[5].

Despite the prevalence and public health impact of HBV, its origin and evolution remain unclear[7,8]. Inference of HBV nucleotide substitution rates is complicated by the fact that the virus genome consists of four overlapping open reading frames[9], and mutation rates differ between phases of chronic infection[10]. 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 substitution rates ($7.72x10^{-4}$-$3.7x10^{-6}$) for HBV[11-13]. Human HBV is classified into at least nine genotypes, A-I, roughly corresponding to sequence similarity of at least 92.5% within genotypes[14], with a heterogeneous global distribution (Fig. 1a)[8,9]. Attempts to explain the origin of genotypes using human migrations have been inconclusive. The hypothesis that HBV co-evolved with ancient modern humans as they left Africa has been contested due to the basal phylogenetic position of genotypes F and H, found exclusively in the Americas[7]. HBV also infects non-human primates (NHP), and the human and other great ape HBV are interspersed in the phylogenetic tree, possibly due to cross-species transmission[15]. Given the variability of estimated substitution rates, the incongruence of the tree topology with some human migrations, and the mixed topology of the NHP and human HBV sequences in the phylogenetic tree, considerable uncertainty remains about the evolutionary history of HBV.

Recent advances in the sequencing of ancient DNA (aDNA) have yielded important insights into human evolution, past population dynamics[16], and diseases[17,18]. However, ancient sequences have been recovered for only a handful of exogenous human viruses, including influenza (~100 years)[19], variola (~350 years)[20], and HBV (~340 years and ~450 years)[21,22]. The knowledge gained from these few 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 viremia during chronicity[3], the relative stability of its virion[23], and its small, circular, and partially double-stranded DNA genome[9].

Shotgun sequence data were generated from 167 Bronze Age[1] and 137 predominantly Iron Age[2] individuals from Central to Western Eurasia with a sample age range of ~7.1-0.2 kya. We identified reads that matched the HBV genome in 25 samples (Table 1, Extended Data Table 1a, SI Table 3), spanning a period of almost 4000 years, from several different cultures and a broad geographical range (Fig. 1b, Table 1, Extended Data Table 1a, SI Table 3). Using TaqMan PCR, we tested two samples with high genome coverage (DA195, DA222) and two samples with low coverage (DA85, DA89) 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 DNA[24]. Based on availability of sample material, libraries from 14 samples were selected for targeted enrichment (capture) of HBV DNA fragments (SI Tables 1 and 2). This resulted in increased genome coverage and an average of a 2.4-fold increase in number of HBV positive reads (Extended Data Table

1a, SI Table 3). In total, we obtained 17.9 to 100% HBV genome coverage from the sequence data, with genomic depth ranging from 0.4x to 89.2x (Table 1, 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 (ML) tree using the ancient HBV genomes together with modern human, NHP, 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, except human genotypes F and H (Extended Data Fig. 2).

Recombination is known to occur in HBV[25]. We found strong evidence that an ancient sequence, HBV-DA51, and an unknown parent recombined to form the ancient genotype A sequences. Although that cannot literally be the case due 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, suggesting that the possible recombination occurred after genotypes A, B, and C had diverged. The predicted recombination break points (Extended Data Table 2, Extended Data Fig. 3) correspond closely to the polymerase gene. Thus, it is 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 kya (Fig. 2, Extended Data Table 3b, Methods). Similar

recombinations events, involving the creation of genotypes E and G and a currently circulating B/C recombinant, have also been identified[25].

For detailed phylogenetic analyses, we used a set of 112 reference human and NHP HBV sequences (Dataset 2, see Methods). An ML phylogenetic tree based on these reference sequences and all ancient sequences was constructed (Extended Data Fig. 4). Regression of root-to-tip genetic distances against sampling dates, as well as date randomisation tests, showed a clear temporal signal in the data (Extended Data Fig. 5, SI Figs. 1-3), suggesting that molecular clock models can be applied. A dated coalescent phylogeny was constructed using BEAST2[26] (Fig. 2). The molecular clock was calibrated using tip dates. Strict and relaxed lognormal 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 lognormally distributed rate variation and a coalescent exponential population prior (Extended Data Table 3a). The median root age of the resulting tree is estimated to 11.6 kya (95% Highest Posterior Density (HPD) interval: 8.6 to 15.3 kya) and the median clock rate is 1.18×10^{-5} s/s/y (95% HPD interval: 9.21×10^{-6} to 1.45×10^{-5} s/s/y). Under a strict molecular clock, a coalescent Bayesian skyline population prior was favoured, in which case the median root age is 15.6 kya (95% HPD interval: 13.7 to 17.8 kya) and the median substitution rate 9.48×10^{-6} s/s/y (95% HPD interval: 8.3×10^{-6} to 1.07×10^{-5} s/s/y) (Extended Data Tables 3a-c).

Under all model parameterisations used here, the substitution rate we find is lower than rates estimated from phylogenies built using either modern heterochronous sequences[11] or sequences from mother-to-child transmissions[27], but higher than rates

179 inferred using external calibrations based on human migrations[12]. A lower rate is
180 consistent with Tedder et al. (2013)[28], who found that although mutation rates may be
181 high, mutations within an individual often revert back to the genotype consensus, and
182 thus rarely lead to long-term sequence change. It is also consistent with the so-called
183 time-dependent rate phenomenon, observed for many viruses, which shows that short-
184 term evolutionary rates are higher than long-term rates[29].
185
186 The knowledge of ancient HBV genomes enables us to formally evaluate hypotheses
187 concerning HBV origins using path sampling of calibrated phylogenies based on
188 appropriate external divergence date assumptions. We tested several calibration points
189 implied by a co-expansion of HBV with humans after leaving Africa for support of
190 congruence between migrations and geographical locations of HBV clades[12]. We find
191 weak evidence for a split of the F/H clade between 13.4 and 25.0 kya under a strict,
192 but not a relaxed clock model. We do not find support for the divergence of
193 subgenotype C3 strains between 5.1-12.0 kya, leading to a distribution in different
194 regions of Polynesia, or for divergence of Haitian A3 strains from other genotype A
195 strains between 0.2-0.5 kya under either strict or relaxed clock models (Extended Data
196 Table 3d).
197
198 In the dated coalescent phylogeny, four ancient sequences (from youngest to oldest:
199 HBV-DA119, -DA195, -RISE386, and -RISE387) group with genotype A. The first
200 three fall well within the 7.5% nucleotide divergence criterion used to delimit
201 membership in HBV genotypes. HBV-RISE387 is right on this limit (Extended Data
202 Table 4a)[14]. The three oldest samples lack a six nucleotide insertion at the carboxyl
203 end of the Core gene that is present in all modern genotype A viruses (Table 2)[9].

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 known in modern HBV infections (Table 2). This characteristic viral mutant is usually found in chronic HBV carriers who seroconverted from HBeAg to anti-HBe. Interestingly, RISE386 and RISE387 are archaeologically dated only ~100 years apart and both come from the Bulanovo site in Russia, but their viruses show only 93.34% sequence identity (Extended Data Table 4b), indicating the existence of significant localized HBV diversity ~4.2 kya.

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-DA51, -DA27, -DA222, and -DA29 phylogenetically group with the modern genotype D. They have high sequence identity (96.99 to 98.74%) with modern genotype D sequences (Extended Data Table 4a), and have the typical 33 nucleotide deletion in the PreS1 sequence of the S-gene, encoding the three HBV surface proteins[9] (Table 2).

Sequences HBV-RISE154, -RISE254, and -RISE563 are in sister relationship to the Chimpanzee/Gorilla HBV clade (Fig. 2). HBV-RISE254 and -RISE563 have the same 33 nucleotide deletion in the PreS1 sequence that is shared with NHP HBV and human genotype D (Table 2). HBV-RISE563 does not encode a functional pre-Core peptide (Table 2). Based on sequence similarity across the whole genome, HBV-RISE563 and -RISE254 together might be classified as a new human genotype that is extinct today, and HBV-RISE154 as possibly another (Extended Data Table 4).

229 However, HBV-RISE154 has low genome coverage, which precludes an exact
230 calculation. The sister relationship of these three sequences with modern Chimpanzee
231 and Gorilla HBV could be interpreted as a consequence of relatively recent
232 transmission(s) of HBV from humans to NHPs[15]. However, other scenarios and
233 confounding factors are possible, as these lineages are deeply separated in the tree.
234 Incomplete lineage sorting combined with viral extinction (possibly boosted by
235 massive recent reductions in great ape populations) should be considered. More data
236 on current and, if possible, ancient NHP-associated HBV will be necessary to reach
237 definitive conclusions.
238
239 The geographic locations of some of the ancient virus genotypes do not match the
240 present-day genotype distribution, and also do not match dates and/or locations
241 inferred in previous studies of HBV. While it is important to keep in mind that the
242 data presented here are limited, they provide important spatiotemporal reference
243 points in the evolutionary history of HBV. Their synopsis suggests a more
244 complicated ancestry of present-day genotypes than previously assumed, especially in
245 light of recent insights into the history of human migration.
246
247 We find genotype A in South-Western Russia by 4.3 kya (RISE386, RISE387), in
248 individuals belonging to the Sintashta culture and in a sample (DA195) from the
249 Scythian culture. The western Scythians are related to the Bronze Age cultures of the
250 Western Steppe populations[2] and their shared ancestry suggests that the modern
251 genotype A may descend from this ancient Eurasian diversity and not, as previously
252 hypothesized, from African ancestors[30,31]. This is also consistent with the phylogeny
253 (Fig. 2), as well as the fact that the three oldest ancient genotype A sequences (HBV-

DA195, -RISE386, and -RISE387) lack the six nucleotide insertion found in the youngest (HBV-DA119), and all modern genotype A sequences. The ancestors of subgenotypes A1 and A3 could have been carried into Africa subsequently, via migration from western Eurasia[32].

The ancient HBV genotype D sequences were all found in Central Asia. HBV-DA27, found in Kazakhstan and dated to 1.6 kya, falls basal to the modern subgenotype D5 sequences that today are found in the Paharia tribe from eastern India[33]. DA27 and the Paharia people in India are linked by their Tibeto-Burman ancestry[2,34], possibly explaining the similar viruses.

Based on the observation that genotypes go extinct and can be created by recombination, our data show that the diversity we observe today is only a subset of the diversity that has ever existed. Our data support a scenario in which all present day HBV diversity arose only after the split of the Old World and New World genotypes (25-13.4 kya). 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/H and the apical position of subgenotype C4, which is exclusively found in indigenous Australians[9]. If HBV did co-evolve with ancient modern humans as they left Africa as proposed previously[7], 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 Old/New World split, as supported by our tree in Figure 2. Alternatively, there could have been

279 a New World origin of HBV or the introduction of HBV into humans from a different
280 host. Our data do not allow us to speculate either way.
281
282 To our knowledge, we report the oldest exogenous viral sequences recovered from
283 DNA of humans or any vertebrate. We show for the first time that is possible to
284 recover viral sequences from samples of this age. We show that humans throughout
285 Eurasia were widely infected with HBV for thousands of years. Despite the age of the
286 samples and the imperfect diagnostic test, our dataset contained a surprisingly high
287 proportion of HBV-positive individuals. The actual ancient prevalence during the
288 Bronze Age and thereafter might have been higher, reaching or exceeding the
289 prevalence typically found in contemporary indigenous populations[6]. This clearly
290 establishes the potential of HBV as powerful proxy tool for research into human
291 spread and interactions. The ancient data reveal aspects of complexity in HBV
292 evolution that are not apparent when only modern sequences are considered. They
293 show the existence of ancient HBV genotypes in locations incongruent with their
294 present-day distribution, contradicting previously suggested geographic or temporal
295 origins of genotypes or sub-genotypes; evidence for the creation of genotype A via
296 recombination and the emergence of the genotype outside Africa; at least one now-
297 extinct human genotype; ancient genotype-level localized diversity; and demonstrate
298 that the viral substitution rate obtained from modern heterochronously sampled
299 sequences is misleading. These suggest that the difficulty in formulating a coherent
300 theory for the origin and spread of HBV may be due to genetic evidence of an earlier
301 evolutionary scenario being overwritten by relatively recent alterations, as also
302 suggested by Simmonds et al., in the context of recombination[25]. The lack of ancient
303 sequences limits our understanding of the evolution of HBV and, very likely, 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 paleo-epidemiology, and broaden our understanding of the contributions of natural and cultural changes (including migrations and medical practices) to human disease burden and mortality.

References

1 Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-172, doi:10.1038/nature14507 (2015).

2 de Barros Damgaard, P. & Willerslev, E. 137 ancient human genomes from across the Eurasian steppe. *Nature* (2018, in principle accepted).

3 Lai, C. L., Ratziu, V., Yuen, M.-F. & Poynard, T. Viral hepatitis B. *Lancet* **362**, 2089-2094, doi:10.1016/S0140-6736(03)15108-2 (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, doi:10.1016/S0140-6736(15)61412-X (2015).

5 *World Health Organization, Hepatitis B fact sheet*, <http://www.who.int/mediacentre/factsheets/fs204/en/> (2017).

6 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, doi:10.1016/j.trstmh.2008.04.044 (2008).

7 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, doi:10.1016/j.semcancer.2013.08.006 (2013).

8 Littlejohn, M., Locarnini, S. & Yuen, L. Origins and Evolution of Hepatitis B Virus and Hepatitis D Virus. *Cold Spring Harb. Perspect. Med.* **6**, a021360, doi:10.1101/cshperspect.a021360 (2016).

9 Kramvis, A. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **57**, 141-150, doi:10.1159/000360947 (2014).

10 Hannoun, C., Horal, P. & Lindh, M. Long-term mutation rates in the hepatitis B virus genome. *J. Gen. Virol.* **81**, 75-83, doi:10.1099/0022-1317-81-1-75 (2000).

11 Zhou, Y. & Holmes, E. C. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J. Mol. Evol.* **65**, 197-205, doi:10.1007/s00239-007-0054-1 (2007).

12 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, doi:10.1016/j.ympev.2015.07.010 (2015).

13 Zehender, G. Enigmatic origin of hepatitis B virus: An ancient travelling companion or a recent encounter? *World J. Gastroenterol.* **20**, 7622, doi:10.3748/wjg.v20.i24.7622 (2014).

14 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, doi:10.1002/jmv.21049 (2008).

15 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).

16 Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302-310, doi:10.1038/nature21347 (2017).

17 Rasmussen, S. *et al.* Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell* **163**, 571-582, doi:10.1016/j.cell.2015.10.009 (2015).

18 Feldman, M. *et al.* A High-Coverage Yersinia pestis Genome from a Sixth-Century Justinianic Plague Victim. *Mol. Biol. Evol.* **33**, 2911-2923, doi:10.1093/molbev/msw170 (2016).
19 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. U. S. A.* **96**, 1651-1656 (1999).
20 Duggan, A. T. *et al.* 17(th) Century Variola Virus Reveals the Recent History of Smallpox. *Curr. Biol.* **26**, 3407-3412, doi:10.1016/j.cub.2016.10.061 (2016).
21 Kahila Bar-Gal, G. *et al.* Tracing hepatitis B virus to the 16th century in a Korean mummy. *Hepatology* **56**, 1671-1680, doi:10.1002/hep.25852 (2012).
22 Patterson Ross, Z. *et al.* The paradox of HBV evolution as revealed from a 16th century mummy. *PLOS Pathogens* **14**, e1006750, doi:10.1371/journal.ppat.1006750 (2018).
23 Bond, W. W. *et al.* Survival of hepatitis B virus after drying and storage for one week. *Lancet* **1**, 550-551 (1981).
24 Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762, doi:10.1038/nature08835 (2010).
25 Simmonds, P. & Midgley, S. Recombination in the genesis and evolution of hepatitis B virus genotypes. *J. Virol.* **79**, 15467-15476, doi:10.1128/JVI.79.24.15467-15476.2005 (2005).
26 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537, doi:10.1371/journal.pcbi.1003537 (2014).
27 Simmonds, P. Reconstructing the origins of human hepatitis viruses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1013-1026, doi:10.1098/rstb.2001.0890 (2001).
28 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, doi:10.3851/imp2655 (2013).
29 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. *Proceedings of the Royal Society B: Biological Sciences* **281** (2014).
30 Zehender, G. *et al.* Reliable timescale inference of HBV genotype A origin and phylodynamics. *Infect. Genet. Evol.* **32**, 361-369, doi:10.1016/j.meegid.2015.03.009 (2015).
31 Hannoun, C. 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, doi:10.1099/vir.0.80972-0 (2005).
32 Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2632-2637, doi:10.1073/pnas.1313787111 (2014).
33 Ghosh, S. *et al.* Unique Hepatitis B Virus Subgenotype in a Primitive Tribal Community in Eastern India. *J. Clin. Microbiol.* **48**, 4063-4071, doi:10.1128/jcm.01174-10 (2010).
34 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. *Proceedings of the National Academy of Sciences* **113**, 1594-1599, doi:10.1073/pnas.1513197113 (2016).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements BB dedicates this work to his late mother, D. Tserendulam. We thank Stuart 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), 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, Germany. BB was supported by Taylor Famil-Asia Foundation Endowed Chair in Ecology and Conservation Biology.

Author Contributions

All authors contributed to the interpretation of the results.

BM, TJ, PD, MA, SR, MS, LO, LV, DS, DG, RF, CD, EW wrote the paper.

BM, TJ: screened and analysed data, created display items.

PD, MA: conducted sampling and generated sequence data.

IS, AL, EU, IP, BB, TB, KT, VM, NL, DV, EK, AE, DP, MV, TDP, VM, VS: excavated, curated, and analysed samples and archaeological context.

AH: designed virus capture probes.

LV: designed virus capture probes, performed TaqMan PCR and target enrichment experiments.

AO: initiated and provided critical input on the development of NGS bioinformatics tools.

DS, DG, RF: computational analysis.

CD: analysed data, PCR probe design.

KS, KK: conducted sampling and archaeological background.

EW: initiated the work, led sampling and generation of the sequence data.

Author Information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E. W. (ewillerslev@snm.ku.dk).

Tables

Table 1: Overview of samples used for phylogenetic analyses

Sample	^{14}C age (standard deviation)	Median cal BP age, or estimate (years)	Approx. sample age (years)	Site	Culture or period	Sex	Reads included in consensus	Coverage	
								Consensus	Depth
RISE563	3955 (35)	4421	4488	Osterhofen-Altenmarkt, Germany	Bell Beaker	M	4383	100%	79.3x
DA222	N/D	1200-1000	1167	Butakty, Kazakhstan	Karluk	M	4132	100%	89.2x
DA195	2479 (35)	2578	2645	Sandorfalva-Eperjes, Hungary	Hungarian Scythian	F	1445	99.9%	29.2x
DA51	2220 (37)	2230	2297	Keden, Kyrgyzstan	Saka	M	712	99.2%	14.5x
RISE254	3631 (29)	3942	4009	Szàzhalombatta-Földvàr, Hungary	Vatya	M	1491	99.0%	36.6x
DA119	N/D	1500	1567	Poprad, Slovakia	North Carpathian	M	2597	98.8%	53.1x
RISE386	3758 (34)	4121	4188	Bulanovo, Russia	Sintashta	M	331	97.8%	7.0x
DA27	1641 (33)	1543	1610	Halvay 3, Kazakhstan	Hun-Sarmatian	M	890	90.0%	14.3x
DA29	849 (25)	755	822	Karasyur, Kazakhstan	Medieval	M	222	87.5%	4.8x
DA45	2083 (27)	2053	2120	Omnogobi, Mongolia	Xiongnu	M	215	87.2%	4.3x
RISE387	3822 (33)	4215	4282	Bulanovo, Russia	Sintashta	N/D	284	86.6%	6.2x
RISE154	3522 (24)	3784	3851	Szczepankowice, Poland	Unetice	F	128	57.2%	2.0x

Samples included in phylogenetic analysis, by decreasing genome coverage. Criteria for inclusion were at least 50% genome coverage and sufficient aDNA damage patterns after capture. The read count indicates the number of reads used to make consensus sequences. N/D (not determined) indicates samples where dating was not performed or where osteological sex was undetermined. See Methods for information on sequence matching, consensus making, and sample dating.

458 **Table 2: Genome properties of ancient sequences included in phylogenetic analyses**

Sample	Genotype of closest sequence	Sequence identity to closest sequence	Genome length	Sero-type	Insertions / deletions	Predicted HBeAg status
DA119	A3	97.8%	3221	*adw2*	6nt insert at the C-terminus of core region	Positive
DA195	A3	96.2%	3215	*adw2*	None	Positive
RISE386	A	95.2%	3215	*adw2*	None	Positive
RISE387	A	92.5%	3215	*adw2*	None	Negative PreC stop codon
DA45	B1	96.6%	3215	*ayw1*	None	Positive
DA29	D3	98.5%	3182	*ayw2*	33nt deletion at the N-terminus of the preS1 region	Positive
DA222	D3	98.7%	3182	*ayw2*	33nt deletion at the N-terminus of the preS1 region	Positive
DA27	D1	97.2%	3182	*ayw2*	33nt deletion at the N-terminus of the preS1 region	Positive
DA51	D1	96.7%	3182	*ayw2*	33nt deletion at the N-terminus of the preS1 region	Positive
RISE154	Chimp.	92.5%	Ambiguous	*adw2*[*]	Ambiguous	Positive
RISE254	Chimp.	95.2%	3182	*adw2*	33nt deletion at the N-terminus of the preS1 region	Positive
RISE563	Gorilla	92.7%	3182	*adw2*	33nt deletion at the N-terminus of the preS1 region	Negative PreC stop codon

459 Genotype groups are sorted by increasing sample age. [*] Serotype could not be determined unambiguously, due
460 to lack of coverage.

461

Figure legends

Figure 1: Geographic distribution of analysed samples and modern genotypes

a, Distribution of modern human HBV genotypes[8]. Genotypes relevant to the manuscript are shown in colour. Coloured shapes indicate the locations of the HBV-positive samples included for further analysis, as in panel **b**. **b,** Locations of analysed Bronze Age samples[1] are shown as circles, Iron Age and later samples[2], as triangles. Coloured markers indicate HBV-positive samples. Ancient genotype A samples are found in regions where genotype D predominates today, and DA27 is of sub-genotype D5 which today is found almost exclusively in India.

Figure 2: Dated maximum clade credibility tree of HBV

A lognormal 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 age and 95% HPD interval of the age under a strict clock and Bayesian skyline tree prior. Clades of genotypes C (except clade C4), F, and H are collapsed and shown as dots. Taxon names indicate: genotype / subgenotype, accession number, sample age, country abbreviation of sequence origin, region of sequence origin, host species, and optional additional remarks.

Methods

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: 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, Horseshoe bat, and four sequences from Roundleaf bats and three from Tent-making bats, largely following Drexler et al.[35]

Dataset 2: 124 HBV genomes, from humans and NHP. This set contains 92 sequences from Paraskevis et al.[12] (excluding their incomplete sequences), 7 additional genotype D sequences, the Korean mummy genotype C sequence[21], the 12 ancient sequences from the present study, and 12 full genomes selected from a set of 9066 full HBV genomes downloaded from NCBI[36] on 2017-08-24 (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 where 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: 124 HBV genomes, from humans, NHP, and a variety of other Orthohepadnavirus host species, including Woolly monkey, Roundleaf and Tent-making bat, Ground and Arctic ground squirrel, Woodchuck, and Snow goose. This set contains 113 sequences that are the

union of a selection of 91 sequences from Paraskevis et al.[12] and 29 from Drexler et al.[35], and 11 additional sequences.

Dataset 4: 3505 HBV genomes. 3384 are from Bell et al., (2016)[37], 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 2017-01-18, resulting in 14 genome categories.

Dating of ancient samples

Sample ages were determined by direct ^{14}C-dating. These ages were calibrated using OxCal[38] (version 4.3) using the IntCal13 curve[39]. Table 1 shows the ^{14}C age and standard deviation for each sample. This is followed by the median probability calibrated age before present (cal BP), where "present" is defined as 1950. RISE386 was ^{14}C dated twice, with ages (standard deviation) of 3740 (33) and 3775 (34), so a rounded mean of 3758 (34) was used for its calibration. DA29 was dated at 822 years using ^{14}C and also at ~700 years using multi-proxy methods, the former 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.

AdapterRemoval[40] (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[41]) using BWA[42] (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 2016-08-31) using DIAMOND[43] (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 HBV-positive samples that had more than three reads matching HBV using DIAMOND were selected for a subsequent BLAST[44] (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 described by Drosten et al.,[45] which amplifies 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 ul, including 5 ul template. Using the Roche LC480 or Agilent Mx3006p instruments, PCRs were incubated for 15 min. at 95°C followed by 45 cycles of 15 seconds at 94°C and 60 seconds 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 ancient DNA clean lab facilities, in which no prior targeted work has been carried out on HBV. Ancient DNA extracts and non-template controls (NTC) were

added into PCR reactions in this location too, which were not subsequently opened. Positive control material was handled in labs in a physically separated building. Here, standard material, diluted to 5-50 copies/reaction, was added to duplicate PCR reactions along with additional NTCs.

Virus capture

14 samples with sufficient sample material were selected for virus capture (DA27, DA29, DA45, DA51, DA85, DA89, DA119, DA195, DA222, RISE254, RISE386, RISE416, RISE568, RISE556). The viral reference genomes for probes were selected as follows. The International Committee for Taxonomy of Viruses (ICTV) 2012 listed 2618 viral species. As many had no associated reference genomes or merely partial sequence information, we selected 2599 sequences of full-length viral genomes, available from GenBank (June 2014), representing viral species found in vertebrates excluding fish. Sequences <1000 nt were discarded. Sequences with identical length and organism ID 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 (SI Tables 1 and 2). For example, among 72 available Hepatitis C virus genome sequences, we selected one genome per subtype (1a-c, g; 2a-c, i, k; 3a, b, i, k; 4a-d, f, g, k-r, t; 5a; 6a-u; 7a). Likewise, 12 HIV-1 genomes were selected representing groups 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 encoding 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 segments sequences. All available segments were included from genomes belonging to *Arenaviridae*, *Bunyaviridae*, and *Reoviridae*. For members of *Poxvirinae* for which full genomes were unavailable (Skunk-, Racoon-, and Volepox virus) sequences representing the conserved gene encoding the DNA-dependent RNA polymerase

were included (n=22). In addition, 2 partial genomes of Squirrelpox virus were included. By mistake 2 and 9 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 (Madison, USA) based on the resulting reference sequences.

Capture was performed on double-indexed libraries prepared from ancient DNA, following the manufacturer's protocol (version 4.3) with the following modifications. Briefly, 1.8 to 2.2 µg of pooled libraries were hybridized at 47°C for 65-70 hours with low complexity C_0T-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 sec denaturation at 98°C, 15 sec annealing at 65°C and 30 sec elongation at 72 °C, ending with 5 min at 72°C. The amplified captured libraries were purified using AMPureXP beads (Agencourt).

Shotgun sequencing data was generated as described in Allentoft et al. (2015)[1]. 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, 6757 reads matched HBV after capture.

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 ancient DNA were applied[46].

(2) Sequences were checked for typical ancient DNA damage patterns using mapDamage[47] (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 ancient DNA[48] (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)[49]. After USER treatment (3h at 37°C) of the aDNA extract, the damage pattern is eliminated (Extended Data Fig. 1b).

(3) As the ancient viruses are from three different HBV genotypes (A, B, D) and a clade in sister relationship to NHP viruses, any argument that samples were contaminated would have to account for this diversity as well as the sequence novelty.

(4) HBV sequences were identified in 25 of 305 analysed samples (Table 1), showing that the findings cannot be due to a ubiquitous laboratory contaminant.

(5) 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 ancient DNA 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, above), 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.

(6) HBV is a blood-borne virus that is mainly transmitted by exposure to infectious blood and that does not occur in the environment[3], making contamination during archaeological excavation extremely unlikely.

Consensus sequences

Reads from the original sequencing and from the capture were aligned to a reference genome (SI Table 3) in Geneious[50] (version 9) using Medium Sensitivity / Fast and Iterate up to 5 times. Because aDNA damage often clusters towards read termini[48], 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 (nt) database (downloaded December 28, 2016) using BLAST. 97.5% of the reads had HBV as their top match. All ancient consensus sequences were matched against the full HBV genomes of Dataset 4 with the Needleman-Wunsch algorithm[51], as implemented in EMBOSS[52] (version 6.6.0.0). For each ancient sequence, the percent sequence identity for each modern genotype and four NHP 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 Program[53], version 4 (RDP4) was used to search for evidence of recombination within the 12 ancient sequences and a selection of 15 modern human and NHP 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 (RDP[54], GENECONV[55], BootScan[56], MaxChi[57], Chimaera[58], SiScan[59], and 3Seq[60]) 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.

Phylogenetic analysis

Initial maximum likelihood phylogenies

An initial Maximum Likelihood (ML) tree was generated to ascertain that the ancient sequences fall within the primate HBV clades. Dataset 1 and the ancient sequences were aligned in MAFFT[61] (version 7). The ML tree was constructed using PhyML[62] (version 20160116), optimizing topology, branch lengths, and rates. We used a GTR substitution model, with base frequencies determined by ML, and an ML-estimated proportion of invariant sites and 100 bootstraps (Extended Data Fig. 2). Furthermore, an ML tree (Extended Data Fig. 4) was generated based on a MAFFT alignment of Dataset 2 and the ancient sequences, using the same parameters as outlined above. The final trees show nodes with support values less than 70 as polytomies.

Dated coalescent phylogenies

In order to check for a temporal signal in the data, a root-to-tip regression and date randomisation 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

673 Accession Number: AF046996) as an outgroup. Three phylogenetic algorithms were used,
674 Neighbour Joining, ML (PhyML), and Bayesian (MrBayes[63] (version 3.2.5)) methods (SI
675 Figs. 1-3). Root-to-tip distances were extracted using TempEst[64] (version 1.5). For ML and
676 Bayesian, root distances for tip taxa (in substitutions per site) were extracted from optimized
677 tree topologies (ML and Maximum Clade Credibility trees, respectively). For NJ, root-to-tip
678 distances were averaged over 1000 bootstrap replicates. Regression analyses were performed
679 with Scipy[65] (version 0.16.0). For the date randomisation tests, we used three different
680 approaches to randomise tip dates: First, tip dates were randomised between all sequences in
681 the phylogeny. Second, tip dates were randomised only among the ancient sequences
682 presented in this paper, as well as the Korean mummy sequence (accession number
683 JN315779). The modern sequences retained their correct ages. Third, dates were randomised
684 within a clade. For each of the three approaches, we performed three independent
685 randomisations. This resulted in a total of nine analyses, which were run for 100,000,000
686 generations each, under the relaxed lognormal clock model and coalescent exponential tree
687 prior. We also ran the same analyses under a strict clock and coalescent Bayesian skyline tree
688 prior, which were run for 20,000,000 generations. We used a GTR substitution model with
689 unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate
690 variation, and estimated proportion of invariant sites, as found by bModelTest[66] (version
691 1.0.4). None of the analyses using the relaxed clock converged (Estimated Sample Size (ESS)
692 < 200). This is most likely because the mis-specification of the dates leads to an incongruence
693 between the sequence and time information. Under the strict clock model, all runs converged,
694 and none of the 95% HPD intervals of the tree height overlapped between the randomised
695 and the non-randomised runs, fulfilling the criteria for evidence of a temporal signal[67].
696 Dated phylogenies were estimated using BEAST2[26] (version 2.4.4, prerelease). We used a
697 MAFFT alignment of Dataset 2. Using bModelTest[66], 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 strict or relaxed lognormal 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. According to Kass and Raftery[68], a Bayes factor in the range of 3-20 implies positive support, 20-150 strong support, and >150 overwhelming support. The relaxed lognormal clock model in combination with a coalescent exponential tree prior was favoured. For the final tree, a Markov chain Monte Carlo analysis was run until parameters reached an ESS > 200, sampling every 2000 generations. Convergence and mixing were assessed using Tracer[69] (version 1.6). The final tree files were subsampled to contain 10,000 or 10710 (for the relaxed lognormal clock, coalescent exponential tree prior) trees, with the first 25% of samples discarded as burn-in. Maximum clade credibility trees were made using TreeAnnotator[26] (version 2.4.4 prerelease).

In order to formally test the Out of Africa hypothesis, calibration points were tested using path sampling as implemented in BEAST2. Calibration points were constrained as follows. Split of genotypes F and H: The 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[70,71]. Split of subgenotype A3 in Haiti: The MRCA of FJ692598 and FJ692611 was constrained using a uniform(200: 500) distribution, due to the timing of the slave trade to Haiti[72]. Split of C3 in Polynesia: The MRCA of X75656 and X75665 was constrained using a uniform(5,100: 12,000) distribution, due to the range of estimates for the MRCA of Polynesian populations[12,73]. Calibration points were tested under both a relaxed

722 lognormal clock, coalescent exponential tree prior, and a strict clock, Bayesian skyline tree
723 prior.
724

725 ## Data availability

726 The complete sequences in this study have been deposited in the European Nucleotide
727 Archive under sample accession numbers ERS2295383-ERS2295394.
728
729
730

References

35 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. U. S. A.* **110**, 16151-16156, doi:10.1073/pnas.1308049110 (2013).

36 Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* **38**, D492-496, doi:10.1093/nar/gkp858 (2010).

37 Bell, T. G., Yousif, M. & Kramvis, A. Bioinformatic curation and alignment of genotyped hepatitis B virus (HBV) sequence data from the GenBank public database. *Springerplus* **5**, 1896, doi:10.1186/s40064-016-3312-0 (2016).

38 Bronk Ramsey, C. & Ramsey, C. B. Bayesian Analysis of Radiocarbon Dates. *Radiocarbon* **51**, 337-360, doi:10.1017/s0033822200033865 (2009).

39 Reimer, P. J. *et al.* IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. *Radiocarbon* **55**, 1869-1887, doi:10.2458/azu_js_rc.55.16947 (2013).

40 Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337, doi:10.1186/1756-0500-5-337 (2012).

41 *Human Genome Overview - Genome Reference Consortium*, <https://www.ncbi.nlm.nih.gov/grc/human> (2017).

42 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

43 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

44 Camacho, C. *et al.* BLAST : architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

45 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).

46 Willerslev, E. & Cooper, A. Review Paper. Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences* **272**, 3-16, doi:10.1098/rspb.2004.2813 (2005).

47 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, doi:10.1093/bioinformatics/btt193 (2013).

48 Orlando, L., Gilbert, M. T. P. & Willerslev, E. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* **16**, 395-408, doi:10.1038/nrg3935 (2015).

49 Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87-e87, doi:10.1093/nar/gkp1163 (2010).

50 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, doi:10.1093/bioinformatics/bts199 (2012).

51 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).

52 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

53 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, doi:10.1093/ve/vev003 (2015).

54 Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563, doi:10.1093/bioinformatics/16.6.562 (2000).

55 Padidam, M., Sawyer, S. & Fauquet, C. M. Possible Emergence of New Geminiviruses by Frequent Recombination. *Virology* **265**, 218-225, doi:10.1006/viro.1999.0056 (1999).

784 56 Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. A Modified Bootscan
785 Algorithm for Automated Identification of Recombinant Sequences and
786 Recombination Breakpoints. *AIDS Res. Hum. Retroviruses* **21**, 98-102,
787 doi:10.1089/aid.2005.21.98 (2005).
788 57 Smith, J. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**,
789 doi:10.1007/bf00182389 (1992).
790 58 Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from
791 DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13757-
792 13762, doi:10.1073/pnas.241370698 (2001).
793 59 Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-Scanning: a Monte Carlo
794 procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-
795 582, doi:10.1093/bioinformatics/16.7.573 (2000).
796 60 Boni, M. F., Posada, D. & Feldman, M. W. An Exact Nonparametric Method for
797 Inferring Mosaic Structure in Sequence Triplets. *Genetics* **176**, 1035-1047,
798 doi:10.1534/genetics.106.068874 (2006).
799 61 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
800 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780,
801 doi:10.1093/molbev/mst010 (2013).
802 62 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood
803 Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321,
804 doi:10.1093/sysbio/syq010 (2010).
805 63 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference
806 under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
807 64 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal
808 structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus*
809 *Evol* **2**, vew007, doi:10.1093/ve/vew007 (2016).
810 65 *SciPy*, <http://www.scipy.org> (2017).
811 66 Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model
812 averaging and model comparison. *BMC Evol. Biol.* **17**, 42, doi:10.1186/s12862-017-
813 0890-6 (2017).
814 67 Duchêne, S., Duchêne, D., Holmes, E. C. & Ho, S. Y. W. The Performance of the
815 Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data.
816 *Mol. Biol. Evol.* **32**, 1895-1906, doi:10.1093/molbev/msv056 (2015).
817 68 Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773,
818 doi:10.2307/2291091 (1995).
819 69 Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. *Tracer v1.6*,
820 <http://beast.community/tracer> (2017).
821 70 Sanchez, G. *et al.* Human (Clovis)-gomphothere (Cuvieronius sp.) association ~
822 13,390 calibrated yBP in Sonora, Mexico. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10972-
823 10977, doi:10.1073/pnas.1404546111 (2014).
824 71 Bourgeon, L., Burke, A. & Higham, T. Earliest Human Presence in North America
825 Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves,
826 Canada. *PLoS One* **12**, e0169486, doi:10.1371/journal.pone.0169486 (2017).
827 72 Andernach, I. E., Nolte, C., Pape, J. W. & Muller, C. P. Slave trade and hepatitis B
828 virus genotypes and subgenotypes in Haiti and Africa. *Emerg. Infect. Dis.* **15**, 1222-
829 1228, doi:10.3201/eid1508.081642 (2009).
830 73 Kayser, M. *et al.* Melanesian and Asian origins of Polynesians: mtDNA and Y
831 chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23**, 2234-2244,
832 doi:10.1093/molbev/msl093 (2006).

833

834

835

Extended Data table and figure titles and legends

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

a, Extended overview of samples with reads matching HBV. Rows are sorted by decreasing consensus coverage. Explanation of column titles, from left to right starting from the second column: ^{14}C age and standard deviation; Median cal BP age or estimate (in years); approximate sample age in years; site; culture or period; gender; number of sequencing reads that matched HBV using DIAMOND[43]; number of HBV proteins matched by those reads; number of sequencing reads that matched HBV using a BLASTn[44] database built from Dataset 3 (see Methods); the number of reads from the capture that matched HBV using BLASTn (as above); the bit score cut-off above which matching reads were used to form consensus sequences; the percentage of the consensus genome covered by matching reads; average depth of coverage across the reference genome, as reported by Geneious[50]. When reading sample information across a row, an empty cell will be encountered when processing on that sample was concluded, either (in column 6) due to too few matching reads or (penultimate column) 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 non-template controls.

Extended Data Table 2 | Genotype A predicted recombination break points and p-values

a, The p-values assigned to the predicted genotype A recombination by the seven methods used by RDP4[53], in the order given by RDP. The number of sequences in which the

recombination was predicted is always 6, corresponding to the 4 ancient and two modern genotype A sequences. **b,** The predicted start and end break points for each of the 6 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-RISE387/6 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

Models were compared using Path Sampling, as implemented in BEAST2[28]. Likelihood values were compared using a Bayes factor test. A positive value for the Bayes factor implies support for model 1, a negative value support for model 2. According to Kass and Raftery[68], 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 columns model, negative number for the rows model. **b,** MRCA age of individual nodes under a strict clock and Bayesian skyline tree prior or under a relaxed lognormal clock and coalescent exponential tree prior. **c,** 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 lognormal clock and coalescent exponential tree prior.

Extended Data Table 4 | Consensus sequence identity

a, Best consensus sequence identity with 14 groups of HBV full genomes. The Needleman-Wunsch algorithm (as implemented in EMBOSS[52]) was used to globally align each sample consensus sequence against each of the 3384 full HBV genomes of Dataset 4 (see

Methods). The table shows the best nucleotide (nt) similarity percentage for each sample consensus against 14 genome groups from the full set of HBV genomes. In cases where the consensus length is less than the genome length, the given figure is the percentage of identical nucleotides (nts) in the matching region, not counting any alignment gaps or ambiguous consensus nts. 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 sample consensus sequences against one another. The table shows the nt identity percentage for each alignment. In cases where the consensus lengths were unequal, the given figure is the percentage of identical nts in the matching region, not counting any alignment gaps or ambiguous consensus nts.

Extended Data Figure 1 | Ancient DNA damage patterns

The frequencies of the mismatches observed between the HBV reference sequences (Extended Data Table 1) 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 reported in gray. 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**,** RISE154. **b,** Damage patterns for DA222 without (left) and with (right) USER treatment. **c,** Damage patterns with 10, 20, 50, 100, 200, 500 and 1000 reads, where each opaque line corresponds to one replicate set of reads.

Extended Data Figure 2 | Hepadnavirus Maximum Likelihood tree

Shows 26 sequences from the Orthohepadnavirus species (Dataset 1, see Methods) including the ancient HBV sequences. 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. The tree was constructed in PhyML[62], optimizing for topology,

branch lengths, and rates, with 100 bootstraps (see Methods). Internal nodes with <70% bootstrap support are shown as polytomies.

Extended Data Figure 3 | Genotype A recombination break point evidence

RDP4[53] was used to analyse the set of 12 ancient sequences plus a representative set of 15 modern human and NHP 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 panel 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 sub-plots the predicted location of the break points is shown by a dashed vertical line and the surrounding gray area shows the 99% confidence interval for the break point. Sub-plots 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 Figure 4 | HBV Maximum likelihood tree

The sequences from Dataset 2 (see Methods) and the ancient sequences were aligned in MAFFT[61]. The tree was constructed in PhyML[62], 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. Letters on internal branches indicate the genotype. Taxon names indicate: genotype / subgenotype, GenBank accession number, age, country abbreviation of

946 sequence origin, region of sequence origin, host species, and optional additional remarks.
947 Note that the ML tree shows topological uncertainty (polytomies) in areas where the
948 BEAST2[26] tree (Figure 2) is well resolved. This is the case for two reasons. Firstly, BEAST2
949 always produces a fully-resolved binary topology without polytomies. Second, and more
950 important, BEAST2 creates a time tree and uses tip dates to constrain the possible
951 topologies under consideration. Thus BEAST2 can know that certain topologies are unlikely
952 or impossible, whereas ML cannot and thus inherently has greater uncertainty regarding tree
953 topology.
954
955 **Extended Data Figure 5 | Root-to-tip regression and date randomisation tests**
956 **a,** Regression of root-to-tip distances and ages performed in Scipy[65]. 124 branch lengths
957 were extracted using TempEst[64] from trees inferred using neighbour joining (NJ), ML, and
958 Bayesian methods. Shaded areas show 95% confidence intervals. Slopes: 1.01E-05, 1.20E-
959 05, 4.21E-06. Correlation coefficients: 0.45 (R2=0.2), 0.36 (R2=0.13), 0.51 (R2=0.26) for
960 ML, Bayesian, and NJ trees, respectively. **b,** Date randomisation tests under the strict clock
961 model. The median and 95% HPD interval for the substitution rates are given. The rate for
962 the correctly dated tree is shown in red. Dates were randomised within all sequences, within
963 the ancient sequences only, and within each genotype. We performed three replicates of
964 each. None of the 95% HPD intervals for the randomised runs overlap with the 95% HPD
965 intervals for the correctly dated runs, suggesting the presence of a temporal signal in the
966 data.
967