**Supplementary information**

**Title**

**Neolithic and Medieval virus genomes reveal complex evolution of Hepatitis B**

**Authors**

Ben Krause-Kyora†, Julian Susat, Felix M. Key, Denise Kühnert, Esther Bosse, Alexander Immel, Christoph Rinne, Sabin-Christin Kornell, Diego Yepes, Sören Franzenburg, Henrike O. Heyne, Thomas Meier, Sandra Lösch, Harald Meller, Susanne Friederich, Nicole Nicklisch, Kurt Werner Alt, Stefan Schreiber, Andreas Tholey, Alexander Herbig, Almut Nebel, Johannes Krause†

†Corresponding author E-mails: b.krause-kyora@ikmb.uni-kiel.de, krause@shh.mpg.de

***Clip and merge***

The datasets produced for all ancient samples contained paired-end reads with varying numbers of overlapping nucleotides as well as artificial adapter sequences. We used ClipAndMerge version 1.7.3, a module of the EAGER pipeline, to clip adapter sequences, merge corresponding paired-end reads in overlapping regions and to trim the resulting reads. We used the default options with the following command:

java -jar ClipAndMerge.jar -in1 $FASTQ1 -in2 $FASTQ2 \

-f AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \

-r AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA \

-l 25 -qt -q 20 -o $output\_file

where $FASTQ1 and $FASTQ2 are the two gzipped FASTQ input files

***Adapter clipping***

ClipAndMerge uses an overlap alignment of the respective forward or reverse adapter with the 3’ end of each read in order to remove sequencing adapter sequences. Regions at the 3’ end of each read that were contained in the alignment were clipped. Reads that were shorter than 25 nucleotides after adapter clipping or contained only adapter sequences (adapter dimmers) were removed. All remaining reads were then used in the merging step.

***Merging of overlapping paired reads***

Merging was performed for all remaining paired reads with a minimum overlap of 10 nucleotides and at most 5% mismatches in the overlap region. The algorithm selected the maximal overlap fulfilling these criteria. The consensus sequence was generated using the nucleotides in the overlap regions from the read with the higher PHRED quality score, maximizing the quality of the resulting read.

***Quality trimming***

In a final step, ClipAndMerge performed quality trimming of the reads and all nucleotides with PHRED scores smaller than 20 were trimmed from the 3’ end of each read. Finally, all reads with fewer than 25 nucleotides after quality trimming were removed. The resulting high-quality reads were used for the alignment.

***Virus screening***

Screening of the datasets was carried out with the software MALT using the ncbi-viral database as a reference. A sequence identity threshold of 85% was set and the alignment mode was changed to SemiGlobal. The analysis was carried out using the following command:

malt-run --mode BlastN -e 0.001 -id 85 --alignmentType SemiGlobal --index $index --inFile $FASTQCM --output $OUT

where $index is the index file, $FASTQCM is the clipped and merged file and $OUT is the output the file.

The resulting alignments were visually inspected using MEGAN 6. Reads mapping to the hepatitis B reference in the database (NC\_003977.2) were extracted and verified using a discontiguous megablast against the virus taxa (taxid: 10239) with default parameters.

***HBV alignment***

For identification of the genotypes, samples were aligned against one reference for each of the eight hepatitis B genotypes available in the NCBI hepatitis B genotyping project (https://www.ncbi.nlm.nih.gov/projects/genotyping/view.cgi?db=2) (table S1). Additional eight human monkey strains were used. All references were combined in one FASTA file and a competitive mapping was performed using BWA. The mapping was carried out using the following command:

bwa aln -n 0.01 -l 300 $INDEX $FASTQCM $OUT

where $INDEX is the reference, $FASTQCM is the input file and $OUT is the output file. Minimum mapping quality was set to 0.

***Duplicate removal***

We used DeDup version 0.11.3, part of the EAGER pipeline, to identify and remove all duplicate reads in the sample specific BAM files (table S2) with the default options and the following command:

java -jar DeDup.jar -i $IN -o $OUT

where $IN is the input BAM file and $OUT is the output BAM file.

***Extracting mapped reads***

After duplicate removal the resulting BAM files were converted to SAM files using SAMtools version 0.1.19-96b5f2294a with deafult parameters and the following command:

samtools view -h -o $OUT $IN

where $OUT is the SAM output file and $IN is the BAM input file. Reads from the SAM file where converted to FASTQ using the following awk script:

awk '/^[FMR]/{ print “@”$1”\n”$10”\n+\n”$11 }' $IN > $OUT

where $IN is a SAM file and $OUT is the resulting FASTQ file containing all the mapped reads.

***De novo assembly***

The the *de novo* assembly was performed using the SPAdes genome assembler vesion v3.9.0 with the following settings:

spades.py -t 20 -m 500 -k 11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41,43,45,47,49,51,53,55,57,59,61,63,65,67,69,71,73,75,77,79,81,83,85,87,89,91,93,95,97,99,101,103,105,107,109,111,113,115,117,119,121,123,125,127 -s $IN -o $OUT

where $IN is a FASTQ file containing the mapped reads and $OUT is the output folder for SPAdes.

Resulting contigs for each K-value were checked and the one which spawned the longest contig was selected for further processing (table S3).

***Mapping of contigs***

Contigs were mapped against the multi FASTA file containing all 16 references. The following command was used:

bwa mem $INDEX $IN $OUT

where $INDEX is the reference, $IN the file containing the contig/contigs and $OUT is the resulting BAM file.

***Consensus generation***

For genomic reconstruction of the ancient HBV strains, the results of the alignments were inspected visually with IGV version 2.3.92. Information about contig order and direction were used for the construction of a consensus sequence. Bases that were soft clipped in the alignment were cut off using SeqKit software version 0.7.0 and realigned to the 16 references as described above. This was done because of the circular genome structure of HBV. Big contigs needed to be split to preserve genomic order with respect to the reference sequences (table S4).

***Remapping raw reads against the consensus sequence***

Raw reads of each sample were mapped to their corresponding consensus sequence using the software CircularMapper version 1.93.4 and the following command line:

java -jar CircularGenerator.jar -e $E -i $IN -s "$N"

where $E is the length of elongation, $IN is the input file and $N is the name of the target sequence.

bwa aln -t 8 $IN $R -n 0.01 -l 300 -f $OUT

where $IN is the elongated consensus sequence, $R is the file containing the clipped and merged readsand $OUT is the output file.

bwa samse $RE $IN $R -f $OUT

where $RE is the elongated reference, $IN is the bwa aln output, $ R is the file containing the clipped and merged readsand $OUT is the output file.

java -jar realign-1.93.4.jar -e $E -i $IN -r $OR

where $E is the length of elongation, $IN is the output of bwa samse and $OR is the unmodified consensus.

***Phylogenetic analysis***

Hepatitis B reference strains for monkeys were collected using edirect with the following command:

esearch -db pubmed -query "hepatitis B AND Orangutan OR hepatitis B AND Gibbon OR hepatitis B AND Gorilla OR hepatitis B AND Chimpanzee OR hepatitis B AND Orang-utan" | elink -target nuccore | efetch -format fasta > $OUT

where $OUT is the output file in fasta format containing all sequences from the papers containing the search keys.

To control the received sequences a multiple sequence alignment using the linsi algorithm contained in MAFFT version 7.310 was carried out. The following command was used:

linsi $IN > $OUT

where $IN is the input file containing the retrieved sequences and $OUT is the multiple sequence alignment.

The alignment was visually inspected in AliView (v. 1.18.1) and sequences that differed from the majority were removed. This step was necessary due to the unrestricted research command which, by chance, could also return non-primate sequences. After filtering the set contained 74 ape infecting HBV strains.

Using the 74 ape strains and 5497 non-recombinant genomes available at hpvdb (https://hbvdb.ibcp.fr/HBVdb/HBVdbDataset?seqtype=0) clustering was carried out using UClust v 1.1.579 (Edgar et al., 2010). The clustering with an identity threshold of 97% percent yielded 493 representative HBV genomes. Combining them with the five ancient strains a multiple sequence alignment was carried out using Geneious version 10.1.2 (Kearse et al., 2012) with a 65% similarity cost matrix, a gap open penalty of 12 and a gap extension penalty of 3. The multiple sequence alignment was stripped of any sites (columns) that had gaps in more than 95% of sequences. The alignment was used to construct a network with the software SplitsTree v4 (Huson and Bryant, 2006), creating a NeighborNet with uncorrected P distances. Furthermore, information from the was used to finalize our consensus sequences with respect to gaps ant their position in the sequence.

The same multiple sequence alignment was used for the generation of Maximum-Likelihood (ML) and Neighbour-Joining (NJ) Trees. MEGA7 version 7170509-x86\_64 with the following command line was used:

Megacc -a $MAO -d $IN -o $OUT

Where $MAO is the megacc configuration file, $IN is the multiple alignment and $OUT is the output directory. For both trees 1408 informative sites and Jukes-Cantor substitution model were used. Bootstrap replicates are 2000 for ML and 10000 for NJ. The trees are provided in source data 3.

***Molecular clock analysis***

The evolution of Hepatitis B virus over time is unclear in regard to its evolutionary rate and the role of recombination. Previous studies have attempted to detect a molecular clock like signature without success. We investigate if the ancient genomes presented here allow a molecular clock analysis using TempEst v1.5.1 (Rambaut et al. 2016). The data set shows little positive correlation between genetic divergence and sampling time (correlation coefficient 0.075) and there is very little temporal signal (TempEst R^2=0.006, see Figure S10). Therefore, we refrain from further dating analysis.

***LC-MS/MS analysis and database searches***

In-filter digested *in vitro* samples were analyzed on a Dionex Ultimate 3000 nano-HPLC coupled to a Q Exactive mass spectrometer (Thermo Scientific, Bremen). The samples were washed on a trap column (Acclaim Pepmap 100 C18, 10 mm × 300 μm, 3 μm, 100 Å, Dionex) for 5 min with 3% acetonitrile (ACN)/0.1% TFA at a flow rate of 30 μL/min prior to peptide separation on an Acclaim PepMap 100 C18 analytical column (15 cm × 75 μm, 3 μm, 100 Å, Dionex). A flow rate of 300 nL/min using eluent A (0.05% formic acid (FA)) and eluent B (80% ACN/0.04% FA) was used for gradient separation as follows: linear gradient 5% ± 50% B in 60 min, 50% ± 95% B in 5 min, 95% B for 10 min, 95% ± 5% B in 1 min, and equilibration at 5% B for 12 min. Spray voltage applied on a metal-coated PicoTip emitter (30 μm tip size, New Objective, Woburn, Massachusetts, US) was 1.25 kV, with a source temperature of 250°C. Full scan MS spectra were acquired from 5 to 145 min between 300 and 2,000 m/z at a resolution of 60,000 at m/z 400 (automatic gain control [AGC] target of 1E6; maximum ion injection time [IIT] of 500 ms). The five most intense precursors with charge states 2+ used were selected with an isolation window of 1.6 m/z and fragmented by HCD with normalized collision energies of 25. The precursor mass tolerance was set to 10 ppm, and dynamic exclusion (30 s) was enabled.

Acquired spectra were analyzed by database searches using Proteome Discoverer (PD) 2.2.0.388 with the search engines SequestHT (Thermo Scientific). Searches were performed against a combined database built by the combination of the full Swiss protein database (468,716 entries, downloaded from Uniprot, December, 21st, 2017), hepatitis B data base (7 entries, downloaded from Uniprot, December, 7th, 2017) and common laboratory contaminants (115 entries, downloaded from Uniprot, August, 15th, 2014). The following settings were used for the search: semi-tryptic specificity; two missed cleavage sites; mass tolerances of 10 ppm for precursors and for fragment masses 0.02 Da (HCD) and 0.5 Da (CID); static modifications: carbamidomethylation on Cys; dynamic modifications: oxidation of Met, Lys and Pro. An additional search was performed using 12 FASTA files from *in silico* translated DNA sequences. The DNA sequences were obtained from previous DNA sequencing of the samples.

A nearly complete y-ion series and two b-ion fragments allow for an assignment of the full peptide sequence. The peptide was identified in the biological sample from Petersberg with 4 peptide spectral matches, showing that the detection of this peptide is not a random event. Moreover, the same peptide could also be identified in the second biological sample from Karsdorf (not shown); blank runs between the LC-MS/MS runs of the two samples rule out potential artifacts due to sample carryover.

Note that the MS/MS method applied here does not allow us to distinguish leucine (L) or isoleucine (I) residues. Manual permutation of the leucine residues in the above stated sequence followed by a BLAST search (default search parameters) led to the identification of the HBV-protein external core antigen in all cases with the exception of the combinations D***II***DTASALYR and DLLDTASA***I***YR; these two variants were reported by BLAST search as the proteins hypothetical protein CR988\_04570 [*Treponema sp.*] and anti-GFP antibody [synthetic construct] with the HBV-protein external core antigen listed at rank 3. However, these proteins were not found in the genomic data. Hence, despite the uncertainty of the I/L assignment, the MS/MS data support the genomic finding of an HBV infection.

***Human genetics***

The mapping against hg19 was carried out using the following command for all three samples:

bwa aln -n 0.01 -l 300 $INDEX $FASTQCM $OUT

where $INDEX is the reference, $FASTQCM is the input file and $OUT is the output file. The duplicate removal after mapping was executed as described above.

(table S5)

Table S1. Accession numbers for the reference genomes used in the first alignment step to catch HBV diversity in the sample. Since monkey HBV strains are not classified into genotypes the column is left blank.

|  |  |  |
| --- | --- | --- |
| Accession | Genotype | Host |
| X51970 | A | Human |
| AB073846.1 | B | Human |
| M12906 | C | Human |
| M32138 | D | Human |
| AB032431 | E | Human |
| AB036910 | F | Human |
| AB064310 | G | Human |
| AY090454 | H | Human |
| AB032433 |  | Chimpanzee |
| AF222323 |  | Chimpanzee |
| AJ131567 |  | Gorilla |
| AY330911.1 |  | Chimpanzee |
| AJ131571.1 |  | Gibbon |
| U46935.1 |  | Gibbon |
| FM209516.1 |  | Gibbon |
| AF193863.1 |  | Orangutan |

Table S2. Number of reads mapping against the references shown in Table S1 before and after duplicate removal.

|  |  |  |
| --- | --- | --- |
| Sample | Number of reads mapping | After duplicate removal |
| Karsdorf | 12182 | 10718 |
| Sorsum | 5949 | 4299 |
| Petersberg | 2729 | 2125 |

Table S3. Number of contigs and combined contig length of the *de novo* assembly for choosen K-values.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Choosen K-value | Number of contigs | Combined contig length |
| Karsdorf | 17 | 3 | 3207 |
| Sorsum | 41 | 1 | 3276 |
| Petersberg | 17 | 2 | 2692 |

Table S4. Final consensus length after retrieving gap information from the multiple sequence alignment with Geneious.

|  |  |
| --- | --- |
| Sample | Consensus length |
| Karsdorf | 3183 |
| Sorsum | 3182 |
| Petersberg | 3161 |

Table S5.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Number of mapped reads | Number of mapped reads (after DeDup) | Coverage >=1x |
| Karsdorf | 131860289 | 122568310 | 70,2% |
| Sorsum | 25300485 | 9856001 | 18.8% |
| Petersberg | 143954577 | 105476677 | 77,4% |

Table S6. Basic statistics for the mapping against the references shown in table S1. Shown are mean coverage, mean coverage for the covered region, genome length, number of missing bases and covered bases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Karsdorf** |  |  |  |  |  |
| Reference | Mean coverage | Mean coverage cov. | Genome length | Bases missing | Covered Bases |
| A | 4,12139 | 10,0492 | 3221 | 1900 | 1321 |
| gibbon\_orang\_1\_restored | 5,96826 | 12,4941 | 3182 | 1662 | 1520 |
| B | 3,46345 | 9,69104 | 3215 | 2066 | 1149 |
| C | 20,6267 | 44,3876 | 3215 | 1721 | 1494 |
| AY330911.1\_chimp\_gorilla | 16,0003 | 22,3205 | 3182 | 901 | 2281 |
| D | 24,7942 | 42,0773 | 3182 | 1307 | 1875 |
| U46935.1\_gibbon\_orang | 12,8231 | 24,1581 | 3182 | 1493 | 1689 |
| E | 16,0311 | 24,3922 | 3212 | 1101 | 2111 |
| F | 1,06065 | 4,2203 | 3215 | 2407 | 808 |
| AF193863.1\_gibbon\_orang | 12,6266 | 26,4677 | 3182 | 1664 | 1518 |
| G | 24,1961 | 41,7139 | 3248 | 1364 | 1884 |
| chimp\_gorilla\_3\_resotred | 13,1691 | 17,8239 | 3182 | 831 | 2351 |
| H | 2,37574 | 11,0856 | 3215 | 2526 | 689 |
| AB032433\_chimp\_gorilla | 18,5654 | 26,5386 | 3182 | 956 | 2226 |
| FM209516.1\_gibbon\_orang | 5,4428 | 13,436 | 3182 | 1893 | 1289 |
| AF222323\_chimp\_gorilla | 16,456 | 23,7474 | 3182 | 977 | 2205 |

Tables S6. Continues

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sorsum** |  |  |  |  |  |
| Reference | Mean coverage | Mean coverage cov. | Genome length | Bases missing | Covered Bases |
| A | 2,23812 | 5,54965 | 3221 | 1922 | 1299 |
| gibbon\_orang\_1\_restored | 3,08674 | 6,65898 | 3182 | 1707 | 1475 |
| B | 1,87278 | 4,49664 | 3215 | 1876 | 1339 |
| C | 4,70638 | 12,1048 | 3215 | 1965 | 1250 |
| AY330911.1\_chimp\_gorilla | 8,39189 | 13,4321 | 3182 | 1194 | 1988 |
| D | 6,54808 | 12,9095 | 3182 | 1568 | 1614 |
| U46935.1\_gibbon\_orang | 4,60245 | 8,99018 | 3182 | 1553 | 1629 |
| E | 5,82067 | 10,5152 | 3212 | 1434 | 1778 |
| F | 1,59285 | 6,02471 | 3215 | 2365 | 850 |
| AF193863.1\_gibbon\_orang | 2,47046 | 6,00535 | 3182 | 1873 | 1309 |
| G | 4,47691 | 7,56162 | 3248 | 1325 | 1923 |
| chimp\_gorilla\_3\_resotred | 6,19485 | 10,3638 | 3182 | 1280 | 1902 |
| H | 1,67683 | 8,25574 | 3215 | 2562 | 653 |
| AB032433\_chimp\_gorilla | 8,78536 | 14,4695 | 3182 | 1250 | 1932 |
| FM209516.1\_gibbon\_orang | 5,39126 | 12,7168 | 3182 | 1833 | 1349 |
| AF222323\_chimp\_gorilla | 13,336 | 20,0354 | 3182 | 1064 | 2118 |

Tables S6. Continues

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Petersberg** |  |  |  |  |  |
| Reference | Mean coverage | Mean coverage cov. | Genome length | Bases missing | Covered Bases |
| A | 1,52313 | 4,92076 | 3221 | 2224 | 997 |
| gibbon\_orang\_1\_restored | 0,434632 | 3,34867 | 3182 | 2769 | 413 |
| B | 0,970451 | 3,97959 | 3215 | 2431 | 784 |
| C | 1,90513 | 6,30144 | 3215 | 2243 | 972 |
| AY330911.1\_chimp\_gorilla | 0,69296 | 2,54032 | 3182 | 2314 | 868 |
| D | 22,9139 | 25,3255 | 3182 | 303 | 2879 |
| U46935.1\_gibbon\_orang | 0,2467 | 2,13315 | 3182 | 2814 | 368 |
| E | 8,81071 | 18,2228 | 3212 | 1659 | 1553 |
| F | 0,167341 | 1,92143 | 3215 | 2935 | 280 |
| AF193863.1\_gibbon\_orang | 0,335324 | 2,02467 | 3182 | 2655 | 527 |
| G | 1,12192 | 6,11409 | 3248 | 2652 | 596 |
| chimp\_gorilla\_3\_resotred | 0,226587 | 1,79353 | 3182 | 2780 | 402 |
| H | 0,298289 | 2,42172 | 3215 | 2819 | 396 |
| AB032433\_chimp\_gorilla | 0,461031 | 2,10776 | 3182 | 2486 | 696 |
| FM209516.1\_gibbon\_orang | 0,268699 | 1,36581 | 3182 | 2556 | 626 |
| AF222323\_chimp\_gorilla | 0,804525 | 3,44549 | 3182 | 2439 | 743 |



Figure S1 Skull of the investigated Karsdorf individual 537 is from a male with an age at death of around 25-30 years.



Figure S2. Mandible fragment of the Sorsum individual XLVII 11 analyzed in this study is from a male.



Figure S3. Skull of the analyzed Petersberg individual from grave 820 is from a male with an age at death of around 65-70 years.

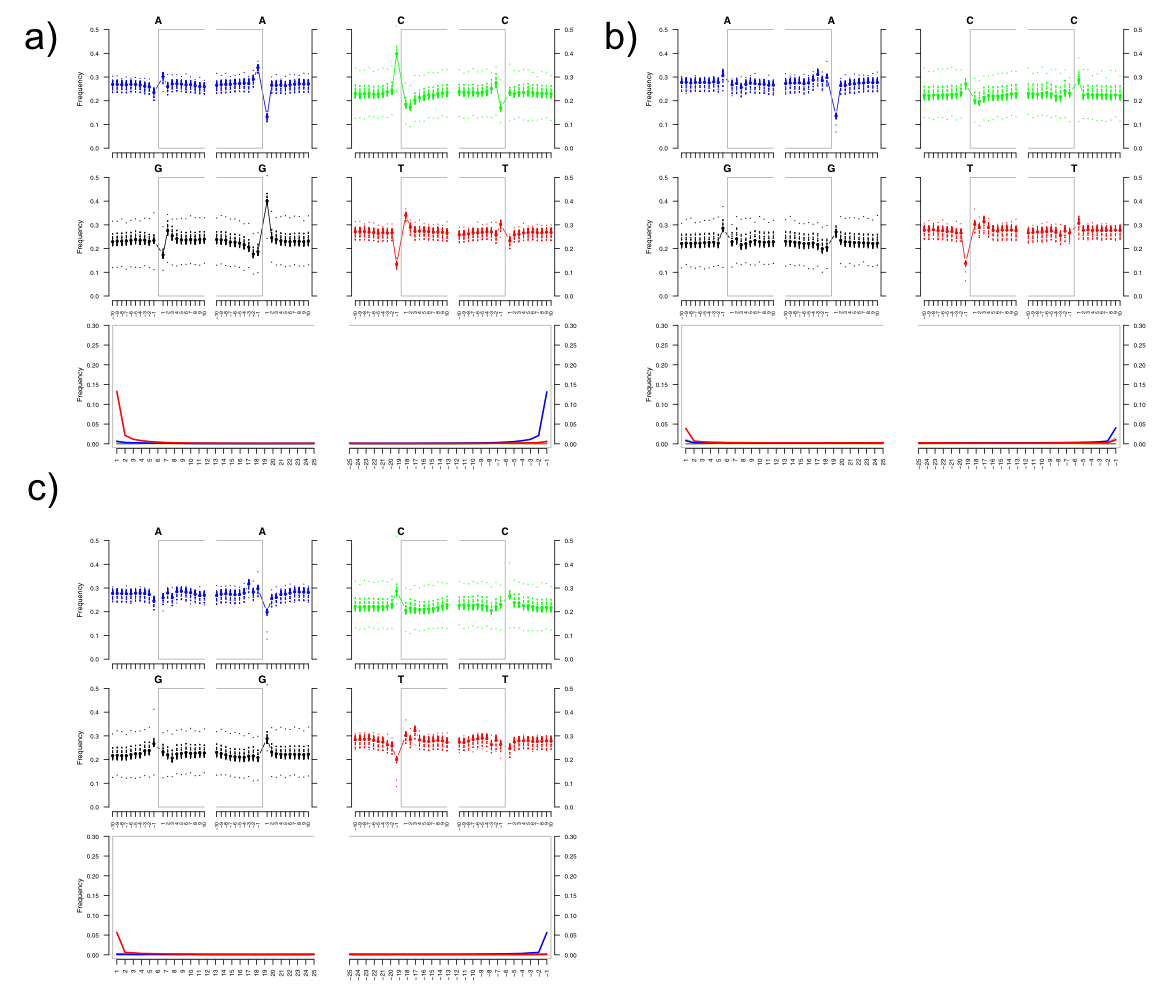


Figure S4) Damage plots showing deamination patterns of hg19 specific reads for the HalfUDG-treated libraries of a) Karsdorf, b) Sorsum, c) Petersberg

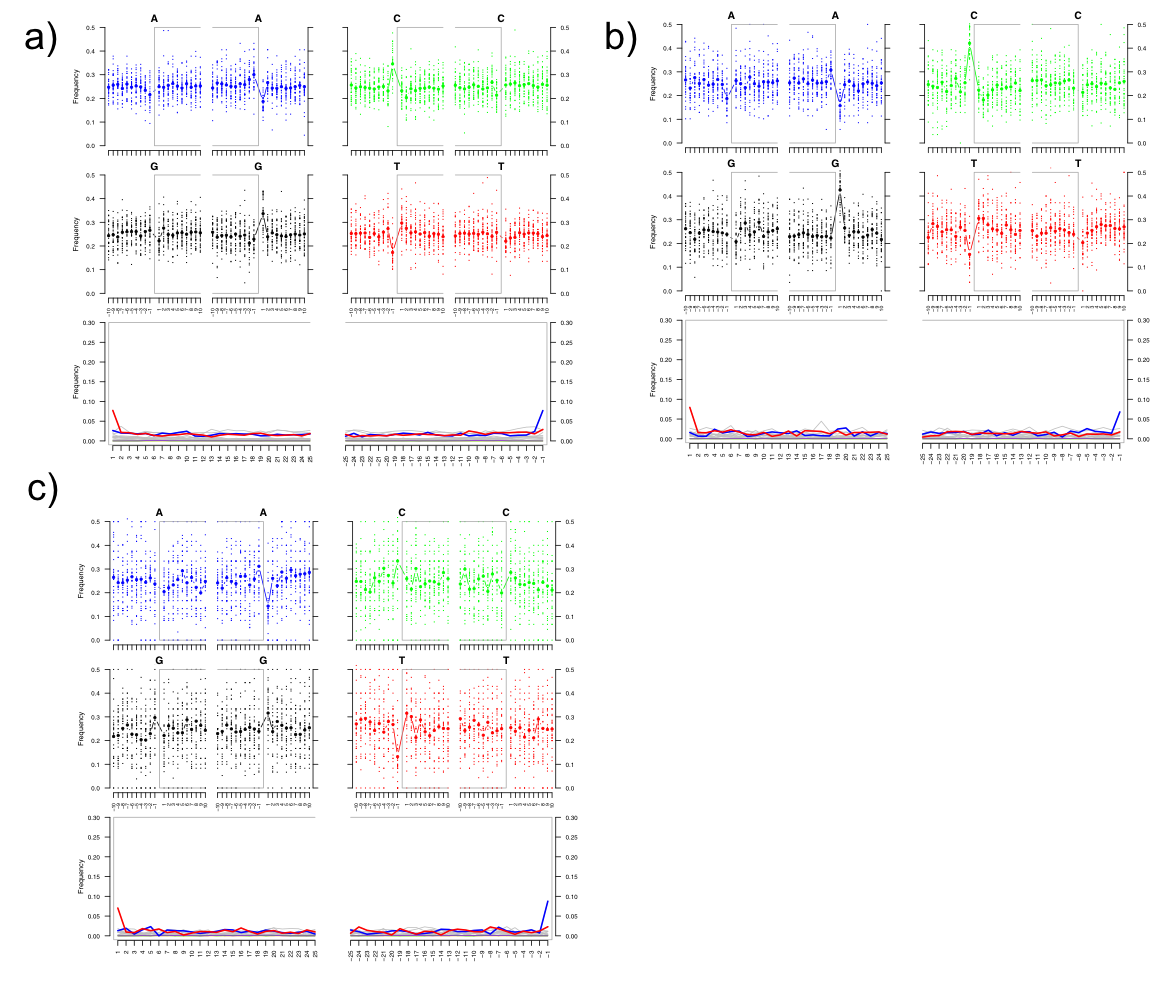


Figure S5) Damage plots showing deamination patterns of HBV specific reads for the HalfUDG-treated libraries of a

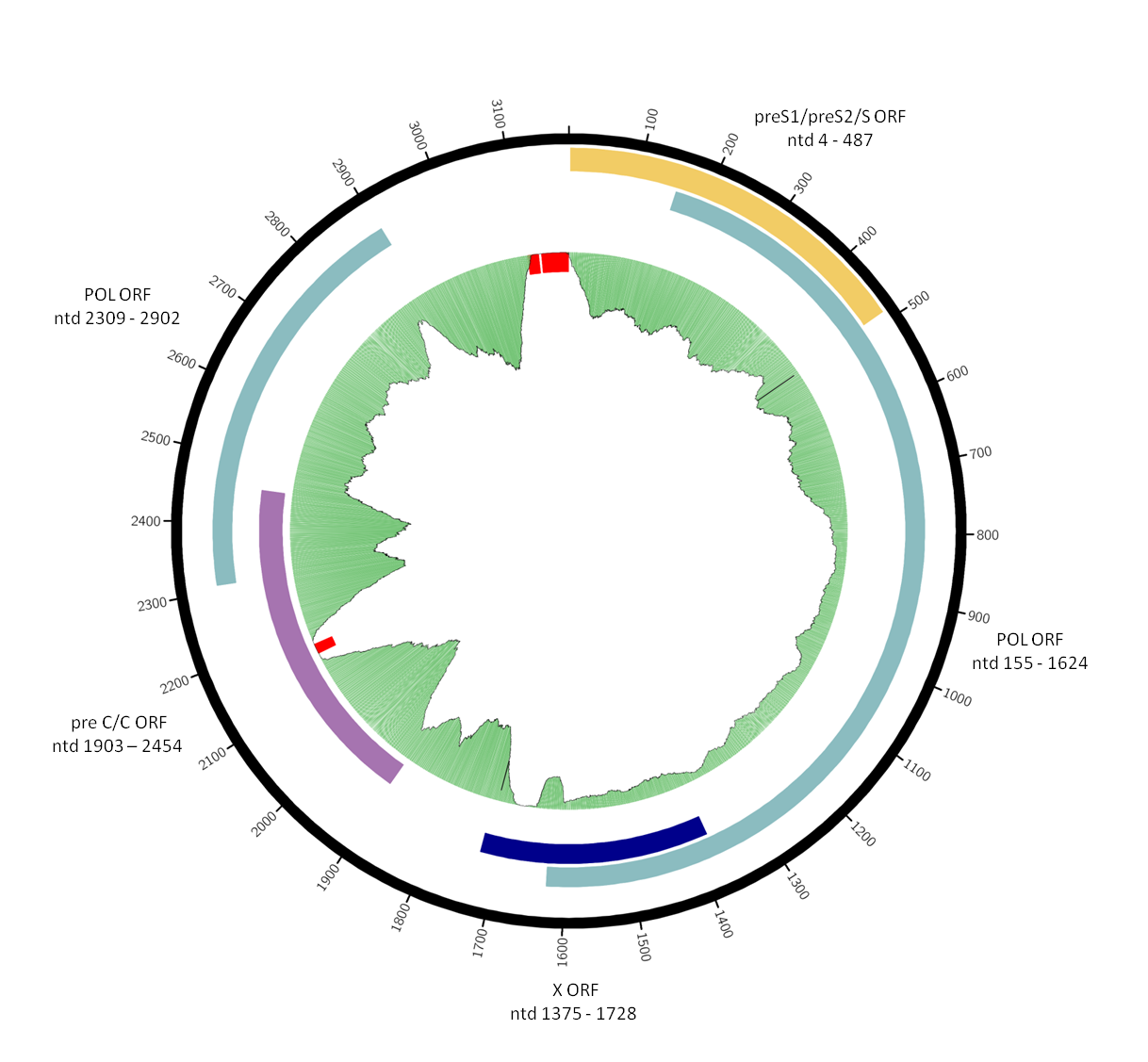
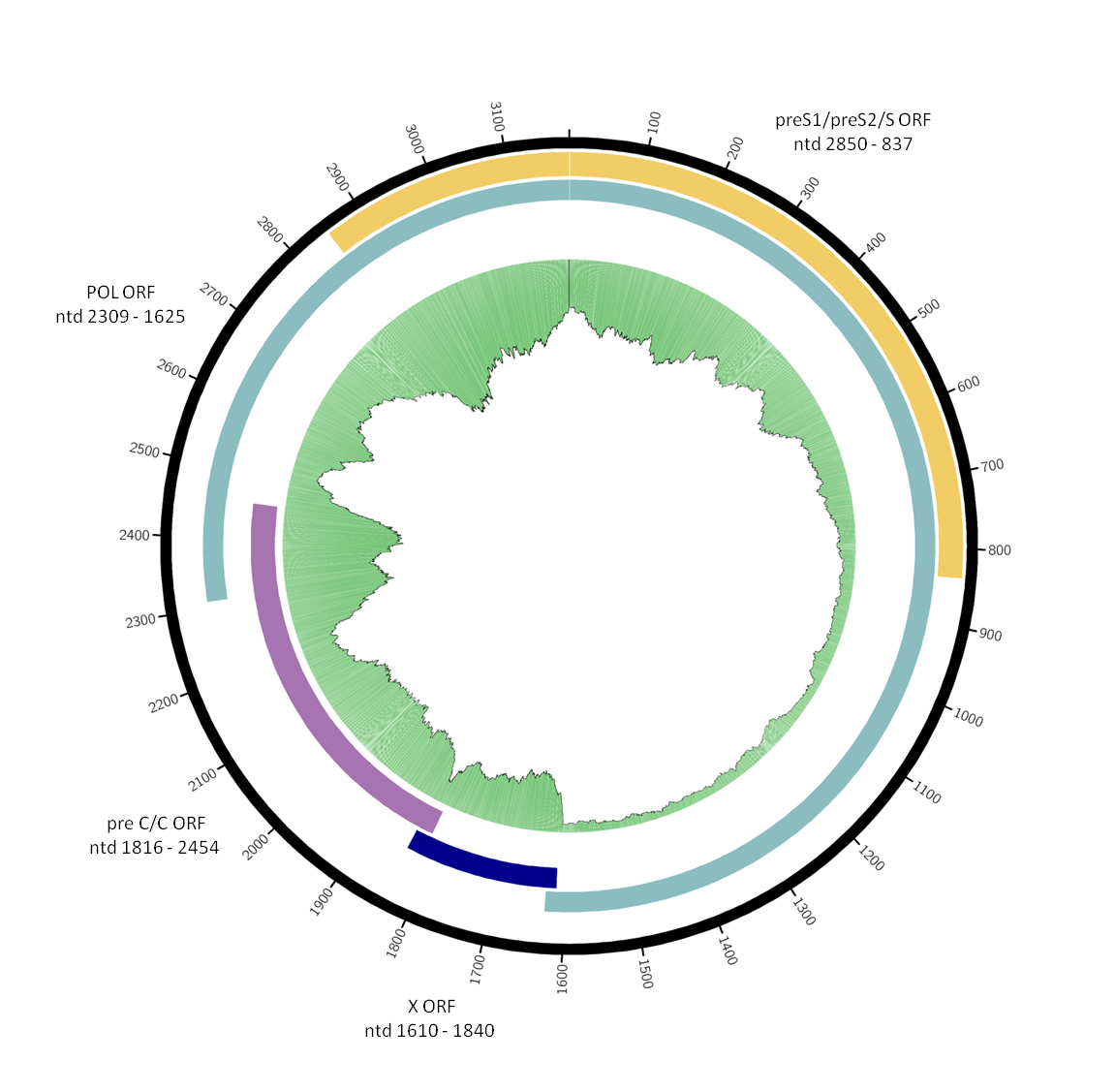


Figure S6. Consensus sequence of the Karsdorf HBV genome as constructed from mapping next-generation sequencing reads against the *de novo* reconstructed consensus sequence. Genomic organization of overlapping open reading frames and approximate location of single-stranded portion of plus strand are indicated, as well as the coverage (green) gaps in the alignment (red). Circular plots were generated using circos-0.69-6 and coverage information from the remapping. YMDD motiv, epsilon motiv, 33nt deletion and the a-determinant are present. No mutations regarding the HbeAg status were detected in the basal core promotor.

Figure S7. Consensus sequence of the Sorsum HBV genome as constructed from mapping next-generation sequencing reads against the *de novo* reconstructed consensus sequence. Genomic organization of overlapping open reading frames and approximate location of single-stranded portion of plus strand are indicated, as well as the coverage (green) gaps in the alignment (red). Circular plots were generated using circos-0.69-6 and coverage information from the remapping. YMDD motiv, epsilon motiv, 33nt deletion and the a-determinant are present. No mutations regarding the HbeAg status were detected in the basal core promotor.

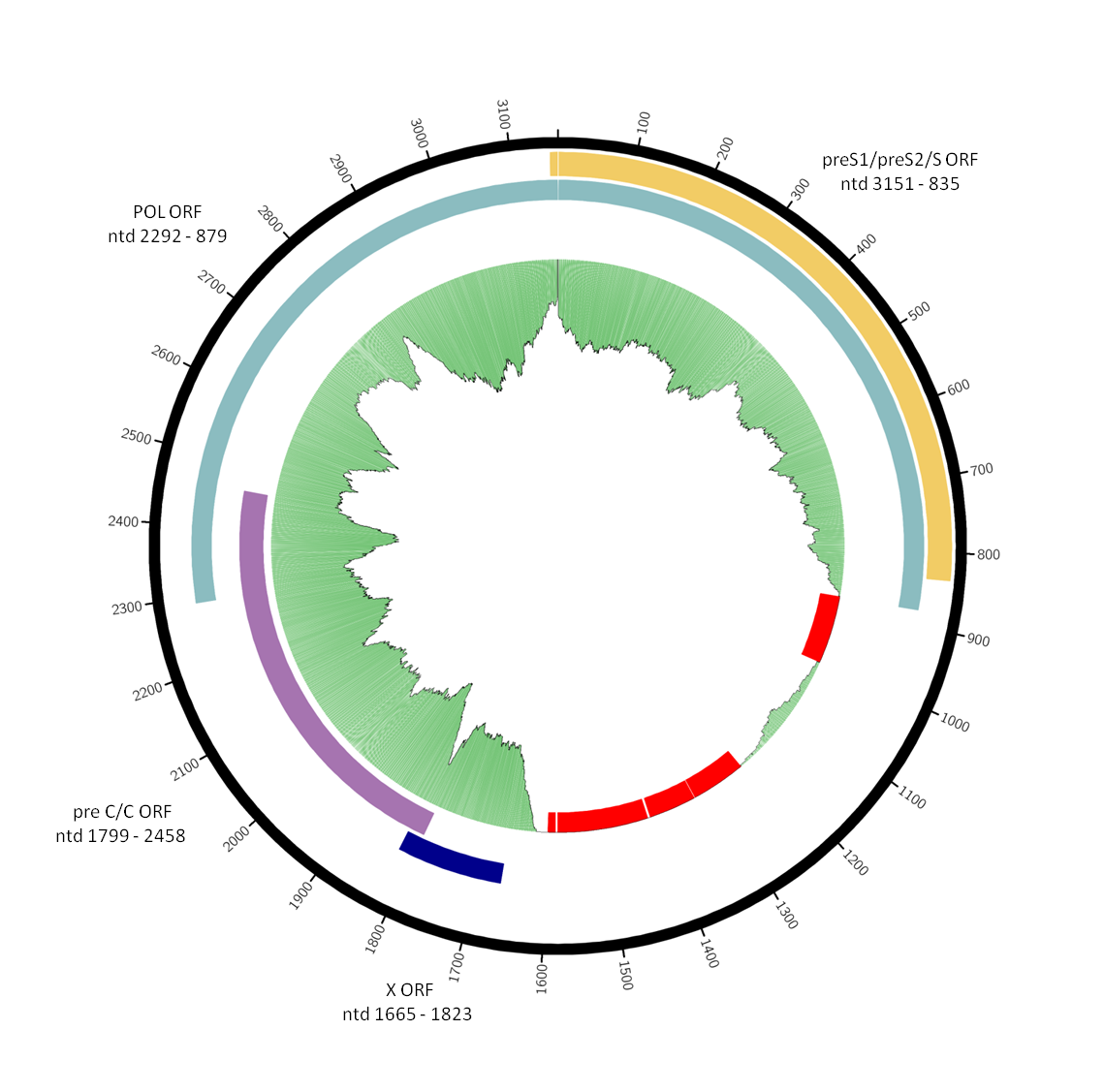


Figure S8. Consensus sequence of the Petersberg HBV genome as constructed from mapping next-generation sequencing reads against the *de novo* reconstructed consensus sequence. Genomic organization of overlapping open reading frames and approximate location of single-stranded portion of plus strand are indicated, as well as the coverage (green) gaps in the alignment (red). Circular plots were generated using circos-0.69-6 and coverage information from the remapping. YMDD motiv, epsilon motiv, 33nt deletion and the a-determinant are present. No mutations regarding the HbeAg status were detected in the basal core promotor.



Figure S9. Genetic (hamming) distance of our three ancient HBV genomes compared to all 493 reference genomes. Gaps or non-called sites (’N') were ignored.



Figure S10: Plot of phylogenetic root-to-tip distance relative to sampling time (TempEst). Each dot represents one sample.



Figure S11. Principal Component Analysis (PCA) of the human Karsdorf and Sorsum samples together with previously published ancient populations projected on 27 modern day West Eurasian populations (shown in gray) based on a set of 1.23 million SNPs (Matthieson et al., 2015).

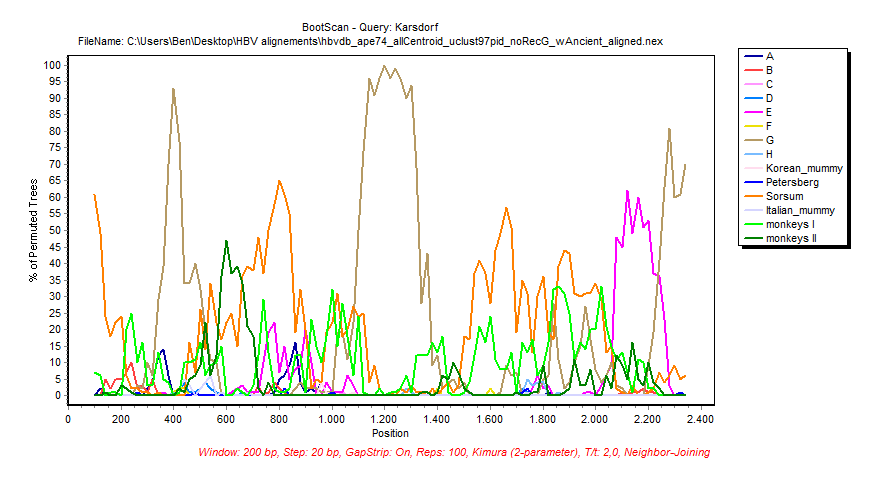


Figure S12. Principal Component Analysis (PCA) of the human Karsdorf and Sorsum samples together with previously published ancient populations projected on 27 modern day West Eurasian populations (not shown) based on a set of 1.23 million SNPs (Matthieson et al. ,2015).



Figure S13. Principal Component Analysis (PCA) of the human Petersberg sample projected on 27 modern day West Eurasian populations based on a set of 1.23 million SNPs (Matthieson et al., 2015).

a)



b)

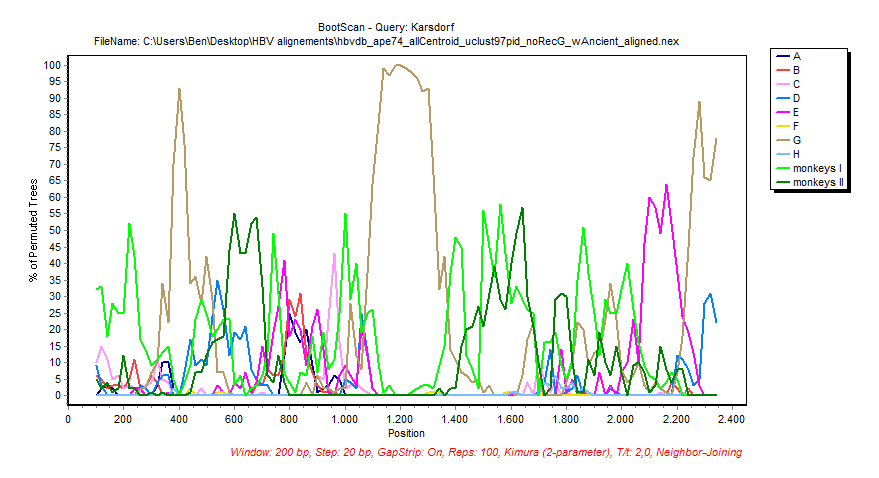
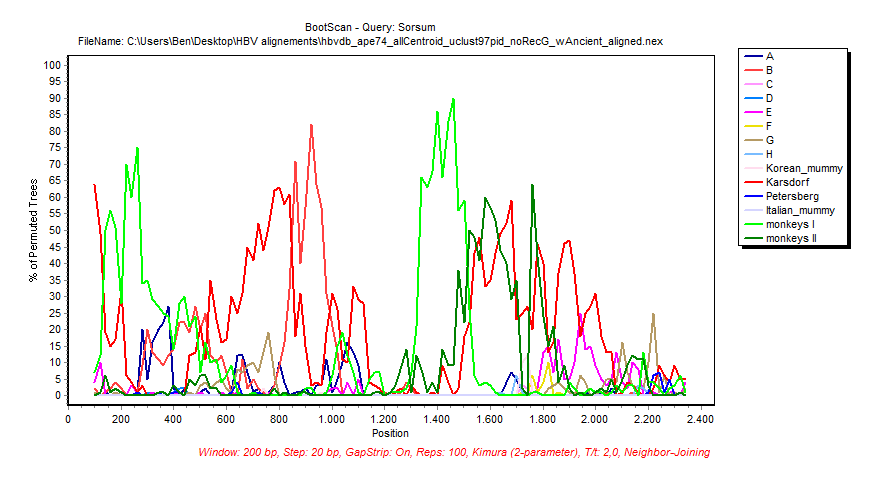


Figure S14. BootScan analysis of the sequence Karsdorf. In each case, sequence fragments of 200 bases incrementing by 20 bases, 100 bootstrap replicates, were compared with sequence groups of a) the eight human genotypes, two primate genotypes, and four ancient genomes and b) the eight human genotypes, two primate genotypes (color coded as described in the legend).

a)



b)

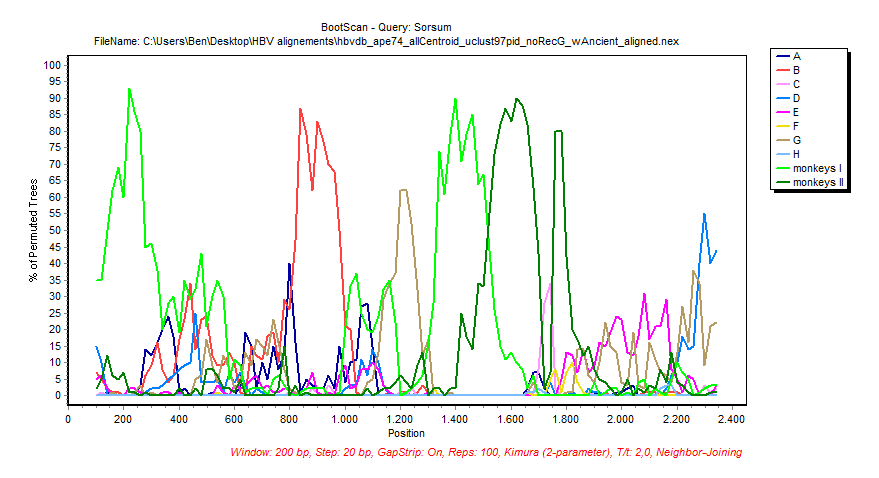
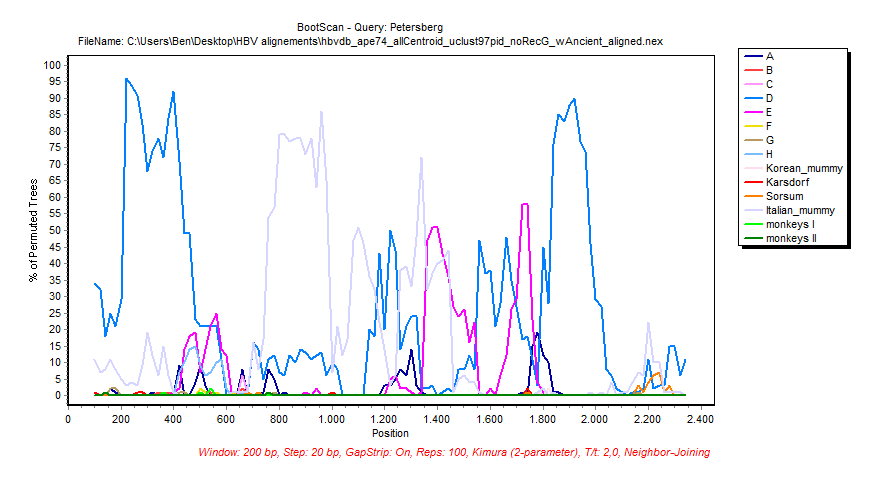


Figure S15. BootScan analysis of the sequence Sorsum. In each case, sequence fragments of 200 bases incrementing by 20 bases, 100 bootstrap replicates, were compared with sequence groups or 50% consensus sequences of a) the eight human genotypes, two primate genotypes, and four ancient genomes and b) the eight human genotypes, two primate genotypes (color coded as described in the legend).

a)



b)

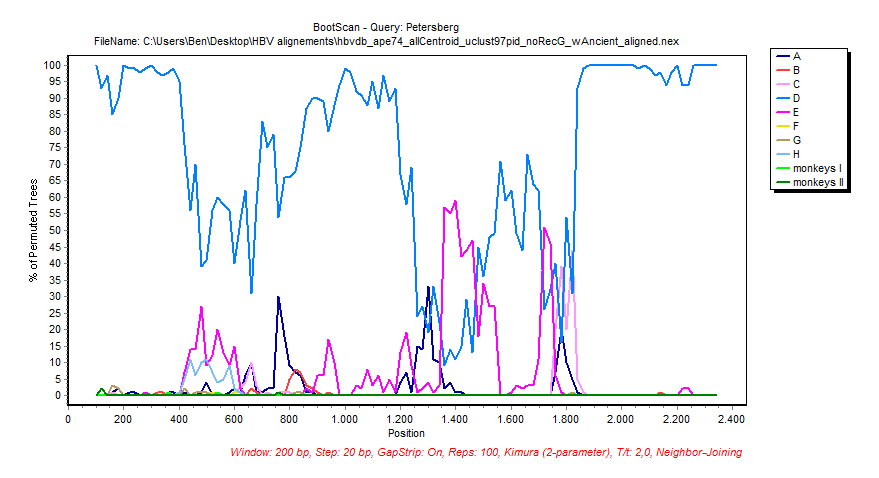
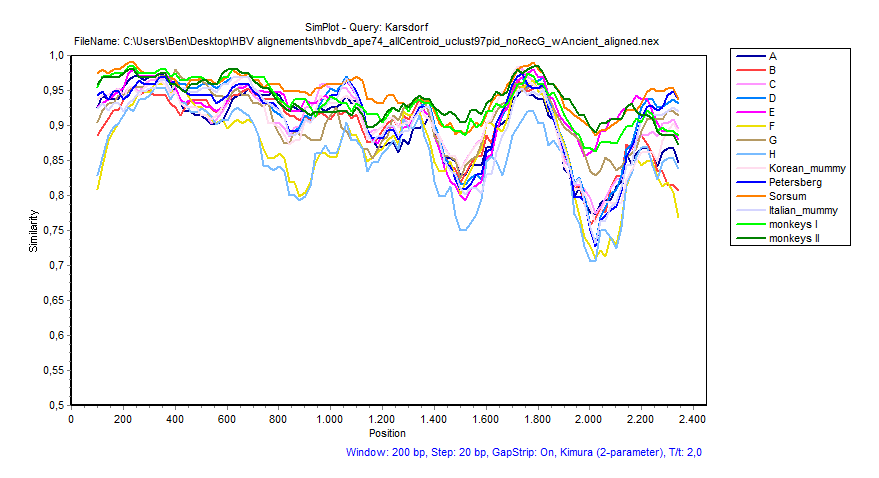
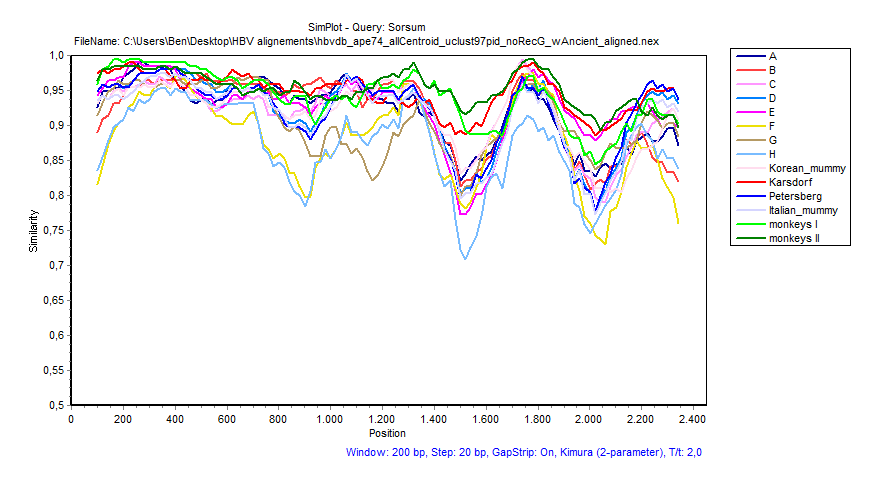


Figure S16. BootScan analysis of the sequence Petersberg. In each case, sequence fragments of 200 bases incrementing by 20 bases, 100 bootstrap replicates, were compared with sequence groups or 50% consensus sequences of a) the eight human genotypes, two primate genotypes, and four ancient genomes and b) the eight human genotypes, two primate genotypes (color coded as described in the legend).

a)



b)



c)

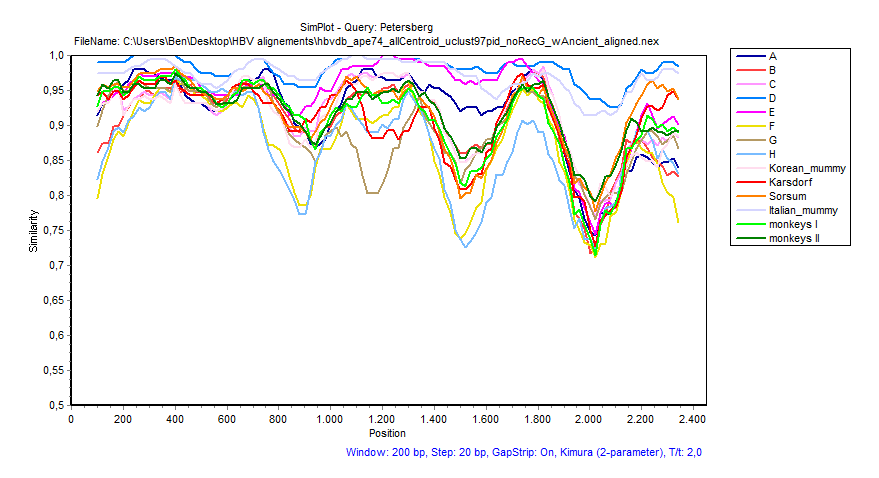


Figure S17. SimPlot analysis of a) Karsdorf, b) Sorsum and c) Petersberg. In each case, sequence fragments of 200 bases incrementing by 20 bases, 100 bootstrap replicates, were compared with sequence groups of the eight human genotypes, four primate genotypes and four ancient genomes (color coded as described in the legend).



Figure S18. MS/MS spectrum of the proteotypic HBV-peptide DLLDTASALYR from the HBV-protein external core antigen (residues 58-68). ([M+2H]2+): m/z = 1237.6429 Da. Mass accuracy of the precursor peptide = 0.56 ppm.