bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Cold Spring Harbor Laboratory

Article

Reconstructing prehistoric viral genomes from Neanderthal sequencing data

Renata C. Ferreira[1,2], Gustavo V. Alves[1], Marcello Ramon[3], Fernando Antoneli[1] and Marcelo R. S. Briones[1]*

[1]Center for Medical Bioinformatics, Escola Paulista de Medicina, Federal University of São Paulo, Brazil
[2]Epigene LLC, São Paulo, Brazil
[3]Computique LLC, São Paulo, Brazil
*Corresponding author. E-mail: marcelo.briones@unifesp.br.

Abstract: DNA viruses that produce persistent infections have been proposed as potential causes for the extinction of Neanderthals and therefore, the identification of viral genome remnants in Neanderthal sequencing reads is an initial step to address this hypothesis. Here, as proof of concept, we searched for viral remnants in sequencing reads of Neanderthal genome data by mapping to adenovirus, herpesvirus and papillomavirus, which are double stranded DNA viruses that may establish lifelong latency and can, produce persistent infections. The reconstructed ancient viral genomes of adenovirus, herpesvirus and papillomavirus revealed conserved segments, with nucleotide similarity to extant viral genomes, and variable regions in coding regions with substantial divergence to extant close relatives. Sequencing reads mapped to extant viral genomes showed deamination patterns of ancient DNA and that these ancient viral genomes showed divergence consistent with the age of these samples (≈50,000 years) and viral evolutionary rates (10^{-5} to 10^{-8} substitutions/site/year). Analysis of random effects shows that the Neanderthal mapping to genomes of extant persistent viruses is above the expected by random similarities of short reads. Also, negative control with a nonpersistent DNA virus does not yield statistically significant assemblies. This work demonstrates the feasibility of identifying viral genome remnants in archaeological samples with signal-to-noise assessment.

Keywords: Neanderthal genome; Ancient viruses; Sequence data; Genome assembly; Adenovirus; Herpesvirus; Papillomavirus

1. Introduction

Several hypotheses were proposed to address the problem of Neanderthals' extinction. These include: (I) Competition with modern humans, which suggests that modern humans outcompeted the Neanderthals for resources, such as food and shelter [1]. This could have resulted in the Neanderthals extinction as they were less efficient to survive in the same environments as anatomically modern humans (AMH). (II) Interbreeding with modern humans. This hypothesis proposes that the Neanderthals were absorbed into the modern human population through interbreeding [2]. Over time, this hybridization could have led to the eventual disappearance of the Neanderthal lineage. (III) Environmental factors. Changes in the climate or other environmental factors, such as the onset of the last Ice Age, could have contributed to the Neanderthals' decline and eventual extinction [3]. The Neanderthals may have been less adaptable to changing environmental conditions than AMH. (IV) Disease. It is possible that the Neanderthals were more susceptible to diseases carried by AMH [4]. This could have had a significant impact on their ability to survive and reproduce.

It is likely that a combination of these factors, along with other unknown factors, contributed to the eventual disappearance of Neanderthals. Among the proposed hypotheses, disease, especially infectious diseases, could have played a relevant role in Neanderthal extinction [5]. Viral diseases caused by DNA viruses that produce persistent infections might leave genetic traces that can be detected in DNA from Neanderthal bone remains [6]. The hypothesis of Wolff and Greenwood [6] propose that viruses that produce lifelong infections could have been introduced by modern humans among the Neanderthals, there is no evidence that these viruses were not already present in their common ancestor, or that it was Neanderthal populations that introduced these viruses into modern humans.

DNA viruses that are known to cause persistent infections in humans include: (I) Herpesviruses [7,8]. This family of viruses includes several members, such as herpes simplex virus (HSV), varicella-zoster virus (VZV), and Epstein-Barr virus (EBV). These viruses can establish lifelong infections in their host and remain dormant for prolonged periods of time, periodically reactivating to cause recurrent disease. (II) Papillomaviruses [9]. These viruses are responsible for causing human papillomavirus (HPV) infections, which can lead to the development of several types of cancers, including cervical cancer. The virus can persist in the body for many years, even decades, without causing any symptoms. (III) Polyomaviruses [10]. This family of viruses includes Human polyomavirus 2, commonly referred to as the JC virus or John Cunningham virus and the BK virus, also known as human polyomavirus 1, which can cause persistent infections in the kidneys and other organs. These viruses are often reactivated in people with weakened immune systems, such as transplant recipients. (IV) Adenoviruses [11]. Adenoviruses can cause persistent infections in the tonsils, adenoids, and other mucosal tissues. These viruses can also cause respiratory and gastrointestinal infections. (V) Hepatitis B virus (HBV) [12]. This virus can cause chronic hepatitis, which can lead to cirrhosis and liver cancer. It is estimated that around 250 million people worldwide are chronically infected with HBV.

The persistence of these viruses is dependent on their ability to evade the immune system and/or establish latency, where the virus becomes dormant within the host cells [13]. This makes it difficult for the host to eliminate the virus and can lead to chronic infections.

Paleogenomics provide insights on the evolution of transmissible and nontransmissible diseases [8,14]. DNA samples from bones of human ancestors are a source of genomic information of the hosts as well as pathogens [15]. Ancient viral genomes more likely to be identified in Neanderthal raw sequence data might be DNA viruses causing persistent infections [16]. Among DNA viruses causing persistent, or long-term, infections are herpesviruses, papillomaviruses, and adenoviruses. Herpesviruses, in particular, might have been a major cause for Neanderthal extinction as proposed in previous studies [4,6,17]. Recovery of herpesvirus DNA and microRNAs from nonprimary infections have been reported from teeth and subgingival plaque of living individuals [18,19]. HSV-1 DNA has been recovered from the trigeminal ganglia (TG) of cadavers, indicating that it can reactivate peri- or postmortem [20]. Although herpesvirus (HSV-1) DNA is detectable in blood during primary infection it is not detected in reactivation from latency [21]. Herpesvirus (European HSV-1) has been successfully recovered from samples dating from the 3rd to 17th century CE, with up to 9.5x coverage with paired human genomes up to 10.16x coverage [8]. Therefore, herpesvirus must be common in the archaeological record.

In this study, we addressed the hypothesis of Neanderthal infection with adenovirus, herpesvirus, and papillomavirus by testing if it is possible to identify remnants of viral genomes in Neanderthal sequencing data. As proof of concept, we searched for viral genomes in raw sequencing data of *Homo sapiens neanderthalensis* from Chagyrskaya cave in Russia (samples 1 and 7) [22]. Analysis of taxonomic output of the Sequence Taxonomic Analysis Tool (STAT) on the NCBI Sequence Read Archive (SRA) [23], revealed that reads corresponding to several microorganisms, such as bacteria, fungi and viruses were detected in these samples. Inspection of results from this analysis allowed the selection of genomes of human pathogenic DNA viruses to be used as references (templates) for genome assemblies.

2. Materials and Methods

2.1. Samples, representative sequences, and reference sequences

Neanderthal reads were downloaded from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) under accession numbers PRJEB55327, Runs ERR10073004 to ERR10073014 for Neanderthal Chagyrskaya 1 (from lower left deciduous canine) and Runs ERR10073054 to ERR10073181 for Neanderthal Chagyrskaya 7 (from thoracal vertebral process fragment). Representative sequences, used as assembly references, were: Human adenovirus 7 (KX897164), human alphaherpesvirus 1 (MN136523) and human papillomavirus type 12 (X74466). The human genome reference sequence used was from BioProject PRJNA31257, NCBI RefSeq assembly GCF_000001405.40 and GenBank assembly GCA_000001405.29 (https://www.ncbi.nlm.nih.gov/grc). Reference sequences (NCBI RefSeq https://www.ncbi.nlm.nih.gov/refseq/) used for assemblies were for human adenovirus: AC_000007.1, AC_000008.1, NC_012959.1, NC_001460.1, NC_011203.1, NC_011202.1, NC_001405.1, AC_000006.1, NC_010956.1, NC_003266.2, NC_001454.1, AC_000017.1, AC_000018.1, AC_000019.1; for human herpesvirus: NC_001806.2, NC_001798.2, NC_001348.1, NC_009334.1, NC_001716.2, NC_009333.1; for human papillomavirus: NC_001676.1, NC_013035.1, NC_001356.1, NC_001352.1, NC_001357.1, NC_001694.1, NC_027779.1, NC_026946.1, NC_001457.1, NC_001595.1, NC_001596.1, NC_001576.1, NC_001526.4, NC_001583.1, NC_001586.1, NC_001587.1, NC_001354.1, NC_001690.1, NC_001591.1, NC_001691.1, NC_001593.1, NC_001693.1, NC_001458.1, NC_034616.1, NC_010329.1, NC_004104.1, NC_004500.1, NC_005134.2, NC_008189.1, NC_008188.1, NC_012213.1, NC_012485.1, NC_012486.1, NC_014185.1, NC_014955.1, NC_017993.1, NC_017994.1, NC_017995.1, NC_017996.1, NC_017997.1, NC_028125.1, NC_022892.1, NC_023891.1, NC_027528.1, NC_001531.1. For DNA virus negative control, the human parvovirus B19 reference sequence used was NC_000883.2. Primate (non-human) and murid NCBI RefSeq viral sequences used were Chimpanzee adenovirus Y25 NC_017825.1, Murine adenovirus 3 NC_012584.1, Murine adenovirus A NC_000942.1, Chimpanzee alpha-1 herpesvirus strain 105640 NC_023677.1, Macacine herpesvirus 1 NC_004812.1, Murid herpesvirus 1 NC_004065.1, *Mus musculus* papillomavirus type 1 isolate MusPV NC_014326.1, *Rattus norvegicus* papillomavirus 3 isolate Rat_60S NC_028492.1, rhesus macaque papillomavirus NC_001678.1.

2.2. Genome assembly by reference mapping

For genome assembly by mapping, duplicate reads were removed using Dedupe Duplicate Read Remover 38.84 and trimmed for illumina TrueSeq adapter/primers using BBDuk. The resulting reads were mapped to viral assembly reference sequences (templates) and NCBI RefSeq sequences using BBMap with k-mer length = 13, random map of multiple best matches and maximal indel size = 16,000 and normal sensivity [24,25] (https://sourceforge.net/projects/bbmap/). Consensus sequences were obtained by strict 50% majority rule and high quality 60%. BBMap, a splice-aware global aligner for DNA and RNA sequencing reads, outperforms Bowtie 2 and BWA for short read sequences because it is fast and extremely accurate, particularly with highly mutated genomes or mapping that produce long indels (up to 100kb long) [25]. Images were processed using Geneious Prime 2023 (https://www.geneious.com).

2.3. Tracking and quantification of DNA damage patterns in DNA sequencing reads

Ancient DNA origin of Neanderthal viral reads was tested in BAM files by a Bayesian method as implemented in mapDamage with multiple time runs with convergence tests [26]. In each of the three assemblies, 50,000 iterations were run, and the acceptance ratio was in the admissible range of (0.1-0.3) for all estimated parameters. MapDamage run as a docker via "docker container run -it quay.io/biocontainers/mapdamage2:2.2.0--pyr36_1 /bin/bash" to create the environment. To run "mapDamage -i mymap.bam -r myreference.fasta" was used and "mapDamage -i mymap.bam -r myreference.fasta -t title" to run the program and change the plot title. Adenovirus reads used for the negative control of mapDamage analysis were from NCBI SRA SRX22331168: Whole genome sequencing of Human adenovirus sp., accession PRJNA930027 (run SRR26630751) containing 11,833,726 reads, 1.4G bases (illumina NextSeq 2000). The positive control for mapDamage analysis consisted of Neanderthal reads of Chagyrskaya 7 mapped to the human mitochondrial genome reference rCRS (NC_012920.1).

2.4. Viral Genealogies

The consensus sequences obtained from genome assembly by reference mapping were exported and used for a GenBank wide blast search (MegaBlast) to identify their closest relatives among extant viruses [27]. Alignments of consensus sequences (queries) with closest relatives were obtained by MegaBlast and genealogies by Neighbor-Joining as implemented in Blast with default parameters [28]. Additionally, viral genealogies were inferred using maximum likelihood as implemented in PAUP 4* (https://paup.phylosolutions.com/get-paup/[29] and the inference of Bayesian trees as implemented in MrBayes 3.2.7a (https://nbisweden.github.io/MrBayes/index.html) [30]. In maximum likelihood inference, the initial model parameters were obtained via ModelTest-NG 0.1.7 using the Bayesian Information Criterion (BIC) [31] as implemented in CIPRES (https://www.phylo.org/) [32]. For Bayesian trees inference, the model parameters (base frequencies, substitution rates, proportion of invariants and gamma distribution shape) were estimated during hot and cold chain tree search by MCMC with a burn-in length of 110,000, subsample frequency of 200 and total chain length of 1,100,000. All phylogenies were inferred from alignments produced with MAFFT program (version 7.490) with automatic selection for algorithms L-INS-i, FFT-NS-i, and FFT-NS-2 according to data size, scoring matrix 200PAM/k=2, gap penalty = 1.53 and offset value = 0.123 [33]. All alignments and trees were inferred from whole genome sequences. Phylogenies of N-adenovirus, N-herpesvirus and N-papillomavirus compared to NCBI RefSeq non-primate and murid sequences were inferred using a Maximum Likelihood algorithm as implemented in FastTree 2.1.11 with GTR model, 4 categories of substitution rates and branch support by Shimodaira-Hasegawa test [34]. Trees with closest extant relatives of Neanderthal viruses, RefSeq sequences and non-human primate sequences were inferred using the Bayesian method as implemented in MrBayes 3.2.7 with total chain length = 1,100,000, subsample frequency = 200, burn-in length = 110,000.

2.5. Assemblies with random reference sequences

Random mock reference sequences of adenovirus, herpesvirus and papillomavirus were obtained using the random sequence generator of UGENE program version 44 [35]. Random mock sequences were of the same size and base frequencies as actual assembly reference sequences. Random reference sequences were used for genome assembly by mapping using the same programs and parameters as used for actual reference sequences as described above.. Comparison of assemblies by mapping of real and random reference sequences was assessed by the Welch's *t*-test , a modification of the Student's *t*-test, that does not assume that the two samples have equal variances, as is the case with data here considered [36].

3. Results

3.1. Taxonomic assessment of Neanderthal raw genomic data

An initial taxonomic analysis of sequencing reads, based on MinHash-based k-mers, (STAT, https://www.ncbi.nlm.nih.gov/sra) [23], revealed that reads of human pathogenic DNA viruses such as herpesviruses, papillomaviruses, and adenoviruses, are present in Neanderthal sequencing runs in ≈0.01% reads (Chagyrskaya 1 and 7) (Table 1) [22]. The examination of the analysis, available in the SRA site, for Chagyrskaya 1, comprises 11 runs with 2.96×10^9 bases (3.25 Gb of data). To test if the reads indicated as viral in the SRA analysis of Neanderthal reads, we analyzed in more detail these reads by mapping to extant viral genomes.

Table 1. Viral segments identified by Sequence Taxonomy Analysis Tool (STAT) in Neanderthal Chagyrskaya 1 (SRA BioSample SAMEA110640442) and 7 (SRA BioSample SAMEA110640445) in SRA BioProject PRJEB55327. Includes only human pathogenic DNA viruses producing persistent infections. Excludes bacteriophages, environmental viruses, retroviruses, human endogenous retroviruses, and RNA viruses. Removal of duplicate reads, primers and adapters resulted in the exclusion of 50% reads approximately. Ten representative samples of Chagyrskaya 7 are in the table. Viral reads comprise approximately 0.01% of each run.

Run	Chagyrskaya	# Reads	DNA Viruses
ERR10073004	01	5,187,065	Human gammaherpesvirus 4, Human papillomavirus 12 & 36
ERR10073005	01	1,106,287	Human papillomavirus 20
ERR10073006	01	12,197,881	Human gammaherpesvirus 4, Hepatitis B virus, Human papillomavirus 10, 12, 19 & 209,
ERR10073007	01	5,384,224	Gammapapillomavirus 12
ERR10073008	01	581,197	Not detected
ERR10073009	01	5,434,074	Human alphaherpesvirus 1
ERR10073010	01	11,535,641	Human gammaherpesvirus 4, Human papillomavirus 36, Human adenovirus 7
ERR10073011	01	2,145,096	Not detected
ERR10073012	01	813,154	Not detected
ERR10073013	01	5,836,042	Papillomaviridae
ERR10073014	01	10,283,890	Human papillomavirus 12
ERR10073060	07	31,299,145	Human papillomavirus type 18, Human respirovirus 1, Human mastadenovirus C
ERR10073061	07	30,544,456	Human mastadenovirus C, Human papillomavirus type 16
ERR10073062	07	31,517,112	Human mastadenovirus C
ERR10073063	07	34,004,864	Human alphaherpesvirus 2, Betaherpesvirinae Stealth virus 1, Human mastadenovirus C
ERR10073064	07	34,089,383	Human mastadenovirus C, Human papillomavirus type 16

ERR10073065	07	34,548,780	Human betaherpesvirus 7, Cytomegalovirus 1, Mastadenovirus
ERR10073066	07	34,530,209	Human mastadenovirus C, Human papillomavirus type 16
ERR10073067	07	34,336,757	Human alphaherpesvirus 2, Mastadenovirus
ERR10073068	07	34,800,717	Mastadenovirus
ERR10073069	07	35,065,248	Human mastadenovirus C

3.2. Genome assembly by mapping

Because of higher coverage of Neanderthal genome Chagyrskaya 7 (4.9 fold coverage) [22] it was used for reference mapping using BBMap, an assembler optimized for short reads. Before processing, the data comprised 128 runs with 152.71 x 10^9 bases (170.21 Gb of data). After removal of duplicate reads, trimming and size selection (exclusion of reads < 10 bases), a total of 1,221,449,544 reads of Neanderthal Chagyrskaya 7 were mapped simultaneously to representative genomes of adenovirus (180,420 reads), herpesvirus (1,224,414 reads) and papillomavirus (23,999 reads) (Figure 1). A total of 1,220,020,414 reads did not map to the references. In simultaneous mapping with different references each reference works as a decoy for the other assemblies.

In Figure 1 A, the assembly of Neanderthal reads to adenovirus has a mean read length of 17 bp with 3,090 (8.0%) identical sites and pairwise nucleotide identity of 86.5%. The coverage of 38,528 bases is 102.2 (mean) with standard deviation of 249.4. The maximum coverage is 3,802 with confidence mean of 40.8, Expected Errors at 2,374, Error Free Odds < 0.0001%, Q20 = 99.6%, Q30 = 99.4% and Q40 = 98.7%.

In Figure 1 B, the assembly of Neanderthal reads to herpesvirus has a mean read length of 18 bp with 5,383 (2.9%) identical sites and pairwise nucleotide identity of 81.6%. The coverage of 185,422 bases is 171.2 (mean) with standard deviation of 247.4. The maximum coverage is 6,771 with confidence mean of 40.8, Expected Errors at 15,416, Error Free Odds < 0.0001%, Q20 = 99.6%, Q30 = 99.4% and Q40 = 98.7%.

In Figure 1 C, the assembly of Neanderthal reads to papillomavirus has a mean read length of 20 bp with 1,069 (13.0%) identical sites and pairwise nucleotide identity of 85.5%. The coverage of 8,223 bases is 115.4 with standard deviation of 609.1. The maximum coverage is 6,401 with confidence mean of 40.8, Expected Errors at 378.97, Error Free Odds < 0.0001%, Q20 = 99.6%, Q30 = 99.4% and Q40 = 98.8%.

The average read lengths that mapped to viral reference genomes is between 17 and 20 bases and the mapping is consistent in the three viral assemblies (Figures 2, 3 and 4). In the adenovirus assembly (Figure 1 A) the distribution of read lengths ranges from 13 to 139 bases (Figure 5 A). In the herpesvirus assembly (Figure 1 B) the read length distribution ranges from 13 to 93 bases (Figure 5 B). In the papillomavirus assembly (Figure 1 C) the read length distribution ranges from 13 to 61 bases (Figure 5 C).

3.3. Analysis of deamination patterns in Neanderthal genomic reads

Deamination is a process that can affect ancient DNA (aDNA) data. Ancient DNA is often degraded and can be chemically altered over time. One of the common types of chemical damage in aDNA is cytosine deamination. Cytosine deamination involves the conversion of cytosine (C) to uracil (U) over time. This process can occur more frequently in ancient samples due to environmental factors, such as heat and moisture. The deamination of cytosine to uracil introduces changes in the DNA sequence. Therefore, to show that the reads mapped to viral genomes are of aDNA origin and not contamination with modern DNA, the reads mapped to viral genomes were tested by analysis of deamination as implemented in program mapDamage (Figure 6) that uses as inputs the BAM files of assemblies shown

in Figure 1. In Figure 6 A and C, the assembly of adenovirus and papillomavirus shows that C-to-T deamination is significantly more abundant in the 5′ overhangs as compared to the other changes away from termini. In the herpesvirus assembly, although C-to-T deamination is more prevalent in the termini, C-to-T is also more abundant in central regions than other changes (Figure 6 B). The likelihood provides statistical support for C-to-T deamination in all three assemblies, which is a main feature of reads derived from aDNA as opposed to recent DNA [26]. The G-to-A as observed in Figure 6 could be caused by deamination of guanine residues to xanthine (X) residues. This causes basecalling adenine residues (A) by DNA sequencing methods, and therefore leads the observed G-to-A pattern in terminal residues in aDNA reads (Figure 6). For the mapDamage analysis a positive and a negative control were included. The positive control consists of Neanderthal reads aligned to the modern human mitochondrial genome (mtDNA). This is to show the deamination pattern with *bona fide* human aDNA reads mapped to the reference. The negative control consists of reads from a present-day human adenovirus 7 genome project (PRJNA930027, run SRR26630751) aligned to the human adenovirus 7 reference sequence as this is supposed not to have the aDNA deamination pattern. These controls are depicted in Figure 6 D and E.

3.4. Genome mapping with taxonomic reference sequences

To verify that assemblies in Figure 1 were supported by assemblies with viral reference sequences Neanderthal reads were also mapped to reference viral genomes of adenovirus (14 NCBI RefSeq sequences), herpesvirus (6 NCBI RefSeq sequences) and papillomavirus (45 NCBI RefSeq sequences) to test the preferential mapping of Neanderthal reads to reference sequences (Table 2). The analysis with taxonomy reference sequences shows that the two adenovirus reference sequences with most reads mapped are AC_000018.1 (human adenovirus type 7 with 78,255 reads) and NC_011203.1 (human adenovirus B1 with 54,835 reads), for herpesvirus the two best matches are NC_001806.2 (human herpesvirus 1 strain 17 with 937,258 reads) and NC_001798.2 (human herpesvirus 2 strain HG52 with 167,080 reads) and for papillomavirus the two best matches are NC_001531.1 (human papillomavirus type 5 with 1,889 reads) and NC_001591.1 (human papillomavirus type 49 with 494 reads) (Table 2).

Table 2. Assembly of Neanderthal reads to NCBI viral reference sequences (NCBI RefSeq). A total of 14 adenovirus, 6 herpesvirus and 45 papillomavirus NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) sequences were used for assembly by mapping with Neanderthal reads using BBMap as described in Material and Methods. Three assemblies were generated, one for each viral famility and in each assembly all sequences of each family were mapped simultaneously to all references in that family. The 6 top results, ranked by the number of mapped reads, are shown.

RefSeq accession	# mapped reads	GenBank ID	Virus species
Adenovirus			
AC_000018.1	78,255	GI:56160876	*Human adenovirus type 7*
NC_011203.1	54,835	GI:197944726	*Human adenovirus B1*
NC_011202.1	8,854	GI:197944766	*Human adenovirus B2*
AC_000019.1	8,573	GI:56160914	*Human adenovirus type 35*
NC_003266.2	6,449	GI:51527264	*Human adenovirus E*
NC_010956.1	2,716	GI:190340974	*Human adenovirus D*
Herpesvirus			
NC_001806.2	937,258	GI:820945227	*Human herpesvirus 1 strain 17*
NC_001798.2	167,080	GI:820945149	*Human herpesvirus 2 strain HG52*
NC_009334.1	13,928	GI:139424470	*Human herpesvirus 4*

NC_009333.1	10,443	GI:139472801	*Human herpesvirus 8 strain GK18*
NC_001348.1	4,056	GI:9625875	*Human herpesvirus 3*
NC_001716.2	1,857	GI:51874225	*Human herpesvirus 7*
Papillomavirus			
NC_001531.1	1,889	GI:9627145	*Human papillomavirus type 5*
NC_001591.1	494	GI:9627363	*Human papillomavirus type 49*
NC_005134.2	167	GI:50253426	*Human papillomavirus type 96*
NC_023891.1	82	GI:607064610	*Human papillomavirus type 178*
NC_001596.1	77	GI:9627396	*Human papillomavirus type 9*
NC_014955.1	76	GI:319962668	*Human papillomavirus type 132*

3.5. Genetic relatedness of Neanderthal viruses and extant viruses

To verify the relatedness of the inferred Neanderthal viral sequences to its closest extant relatives in a wide database, the consensus from mapping assemblies were exported and used in a MegaBlast search of the GenBank. Each consensus sequence from the three assemblies in Figure 1 (Supplementary Data S1, S2 and S3) was used as a query. The closest extant relatives identified in the search are depicted as Neighbor Joining trees (Supplementary Figures S1, S2 and S3). In Supplementary Figure S1, the Neanderthal adenovirus is distant from any extant adenovirus genome by at least 0.06 substitutions/site which corresponds to 2,312 substitutions (100 closest relatives) (alignment in Supplementary Data S4). In Supplementary Figure S2, the Neanderthal herpesvirus is 0.07 substitutions/site which corresponds to 12,780 substitutions along ≈50,000 years. The Neanderthal herpesvirus is very distant from any present-day herpesvirus genome (709 closest relatives) (alignment in Supplementary Data S5). In Supplementary Figure S3, the Neanderthal papillomavirus clusters with sequences of the type 12 although its branch length indicates its distant relationship to other sequences in the same cluster at 0.06 substitutions/site, or 494 substitutions along the branch (72 closest relatives) (alignment in Supplementary Data S6).

From the MegaBlast alignments the ten closest relatives of each Neanderthal viruses were extracted for phylogenetic inference by maximum likelihood (Supplementary Figures S4, S5 and S6). In papillomavirus alignment only 9 closest relatives are included due to the exclusion of a partial sequence (MH777358.2). These phylogenies are consistent with relatedness inferred by MegaBlast (Supplementary Figures S1, S2 and S3) and were inferred from alignments in Supplementary Data S7, S8 and S9. The starting parameters of maximum likelihood trees were estimated by ModelTest-NG using the Bayesian Information Criterion (BIC) (Supplementary Tables 1, 2 and 3). Bootstrap analysis shows that the Neanderthal sequences are separated from their closest extant relatives with 100% frequency (Supplementary Figures S1, S2 and S3).

The consensus sequences of N-adenovirus, N-herpesvirus and N-papillomavirus here reconstructed are included in the Supplementary Data S1, S2 and S3 and the differences between the reconstructed Neanderthal sequences and the respective assembly reference sequences are in Supplementary Figures S7, S8 and S9 (VCF files are in Supplementary Tables 4, 5 and 6).

The closest relatives in the Phylogenies shown in Supplementary Figures S4, S5 and S6 were used along with the corresponding NCBI RefSeq sequences shown in Table 2 for each corresponding virus family plus a non-human primate sequence (chimpanzee for adenovirus and herpesvirus) and rhesus macaque for papillomavirus. These three alignments are in Supplementary Data S10, S11 and S12. The Bayesian trees (Figure 7) shows that the closest relative to N-adenovirus (excluding the assembly reference sequence) is KP670856.2 (human adenovirus 7 strain human/CHN/GZ6965/2011/7[P7H7F7]) with 94.9% nucleotide identity and the closest relative to N-herpesvirus

(excluding the assembly reference sequence) is KX265031.1 (human alphaherpesvirus 1 strain 914-H2) with 93.8%. The closest relative of N-papillomavirus (excluding the assembly reference sequence) is MH777286.2 (Metagenome Assembled Genome: human papillomavirus isolate HPV-mSK_144) with 82.3% (Supplementary Tables 7, 8 and 9).

Relationships of N-adenovirus, N-herpesvirus, and N-papillomavirus with RefSeq non-human primate viral sequences were assessed by phylogenies (Supplementary Figure S11). The branching patterns follow the expected distances according to the divergence of the hosts. Also, as expected, the Neanderthal viral sequences cluster together with modern human viral sequences and are distant to chimpanzee sequences (there are no full-length papillomavirus chimpanzee genomes in RefSeq and rhesus macaque papillomavirus genome NC_001678.1 was used instead). Nucleotide identity matrices (Supplementary Tables 10 and 11) show that N-adenovirus and N-herpesviruses are ≈95% identical with human sequences and ≈70% identical with chimpanzee sequences, while N-papillomavirus is 60-75% identical with human viral counterparts while ≈40% identical with rhesus macaque counterpart (Supplementary Table 12).

Regarding the possible degeneration of codons due to ambiguities generated by deamination biases, it must be noted that the exclusion sets in phylogenetic analysis remove characters (positions) with IUPAC ambiguity codes. This degeneration would affect only C-to-T and G-to-A transitions. Even if the C-to-T deamination is so high as to generate a position with an IUPAC ambiguity code it will be removed in the phylogenetic analysis. In the N-adenovirus alignment (34,662 positions) the excluded characters were: 16 gapped, 1,526 missing/ambiguous and 32,235 non-gapped invariants, leaving a total of 885 positions from which trees were inferred. In the N-herpesvirus alignment (150,608 positions) the excluded characters were: 9,880 gapped, 7,417 missing/ambiguous and 128,255 non-gapped invariants, leaving a total of 5,056 positions from which trees were inferred. In the N-herpesvirus alignment (7,538 positions) the excluded characters were: 934 gapped, 254 missing/ambiguous and 3,869 non-gapped invariants, leaving a total of 2,481 positions from which trees were inferred. The C-to-T and G-to-A deamination does not occur in all Cs and all Gs at termini and therefore will not be confounded with a bona fide C-to-T or G-to-A substitution, which will be called in the consensus when above 50% and with high quality (phred score ≥ 30) position (Supplementary Figure S10).

3.6. Putative changes in proteins of Neanderthal viruses

Alignments of the inferred N-adenovirus, N-herpesvirus and N-papillomavirus to the extant closest relatives revealed the sequence changes distributed along most coding regions including those encoding relevant infection proteins (Figure 8). In adenovirus changes are observed in genes encoding Fiber (L5), Hexon (L3) and Penton base and core proteins (L2) (Supplementary Figure S12 and Supplementary Table 13) [37]. In herpesvirus changes are observed in glycoproteins involved in interaction with host cell surface such as gB (UL27), gD (US6), gH (UL22) and gL (UL1) (Supplementary Figures S13 and S14 and Supplementary Table 14) [38]. In papillomavirus changes were observed in capsid protein L1 as well as oncoproteins E6 and E7 (Supplementary Figure S15 and Supplementary Table 15) [39].

The non-synonymous SNPs detected in Neanderthal viral genes show that the effect of deamination is minimal because most C-to-T and G-to-A deamination changes are below the threshold for consensus sequences generation. In other words, the C-to-T and G-to-A changes are located preferentially in the termini of individual reads in quantity and pattern compatible with aDNA, as show above, but not all reads have these changes. Therefore, although the deamination pattern is present, it is below 50% frequency as shown in Supplementary Figure S10. For this reason, the deamination bias does not have to be compensated for phylogenetic analysis and does not compromise the inference of amino acid changes from viral DNA sequences. In the case of C-to-T changes that lead to amino acid changes, the

frequency of the T in a "C" position has to be higher than the frequency expected by mere deamination alone and therefore the mere deamination can be differentiated from a bona fide C-to-T transition.

3.7. Assessment of random effects in genome mapping

To assess whether mapping of Neanderthal reads to viral references could be obtained by chance alone, due to short reads, random viral references sequences were generated and mapped (Figure 9). Random references have the same size and base compositions as the actual references, namely, human adenovirus 7 (KX897164), human alphaherpesvirus 1 (MN136523) and human papillomavirus type 12 (X74466). The idea is to observe the statistics of Neanderthal reads mapped to actual viral sequences versus their random versions (Table 3). In Figure 9 A, the assembly of Neanderthal reads to the random adenovirus reference has coverage of 62.2 (mean) with standard deviation of 180.2 with 126,613 reads mapped. In Figure 9 B, the assembly of Neanderthal reads to random herpesvirus reference has a mean coverage of 154.8 with standard deviation of 704.3 with 1,166,326 reads mapped. In Figure 9 C, the assembly of Neanderthal reads to random papillomavirus reference has a mean coverage of 51.0 with standard deviation of 116.0 with 22,682 reads mapped (Table 3). The Welch's *t*-test was used to assess if mapping of Neanderthal reads with real templates (assembly references) produced assemblies with higher coverage as compared to random reference sequences (Table 3). Comparison of the *t* statistics real references of human adenovirus 7 (KX897164), human alphaherpesvirus 1 (MN136523) and human papillomavirus type 12 (X74466) and random references shows that real template assembly coverages are always above random reference assemblies (the *t* value is always above the critical value for infinte degrees of freedom, because the number of high number of reads) with *p*-values=0 (probability zero of rejecting the null hypothesis that the means of the two samples are equal).

Table 3. Comparison of mapping with real assembly references and random references and Neanderthal sequencing reads. N-adenovirus, N-herpesvirus and N-papillomavirus real template used are the same used in assemblies of Figure 1: Human adenovirus 7 (KX897164), human alphaherpesvirus 1 (MN136523) and human papillomavirus type 12 (X74466). SD = standard deviation and # Reads = mapped reads. Random reference sequences here used are in Supplementary Data. Cov. Indicate "Coverage", column "*t*" indicates the'Welch's *t*-test *t* statistics, *c.v.*=critical value for infinite degrees of freedom and *p*-value is the probability of rejecting the correct null hypothesis, that is, that both samples are statistically equivalent.

	Real			Random					
	Cov. Mean	Cov. SD	# Reads	Cov. Mean	Cov. SD	# Reads	*t*	*c.v.*	*p*-value
N-adenovirus	102.2	249.4	180,419	62.2	180.2	126,613	51.58	1.64	0
N-herpesvirus	171.2	247.4	1,224,713	154.8	704.3	1,166,326	23.78	1.64	0
N-papillomavirus	115.4	609.1	23,998	51.0	116.0	22,682	16.07	1.64	0

To further test for artifactual mapping of Neanderthal reads on viral references, the mapped reads in Figure 1 were extracted ("unassembled") from the bam files and used for mapping to the human genome scaffolds (hg38). No reads mapped to human nuclear chromosomes and only a limited number of reads mapped to human mitochondrial DNA (Figure 10) which shows that reads that map to viral genomes do not map to human genome and therefore random similarities due to short reads do not explain *per se* the results observed in Figure 1. The 180,419 N-adenovirus reads were assembled to hg38 and (140 reads mapped to mtDNA 180,279 reads were not assembled to any human chromosome. The 1,224,713 N-herpesvirus reads were assembled to hg38 and (465 reads mapped to mtDNA 1,224,248 reads were not assembled to any human chromosome. Finally, the 23,998 N-papillomavirus reads were assembled to hg38 and (16 reads mapped to mtDNA 23,982 reads were not assembled to any human chromosome.

The same set of unassembled reads were used for assembly by mapping to the corresponding random reference sequences to test if simple random similarities could account for the mapping coverage, in other words, if random noise could explain the results (Figure 11). The analysis revealed that 232 of 180,419 N-adenovirus reads were assembled to adenovirus random reference while 180,187 reads were not assembled (Figure 11 A) which is the expected result if mapping is not due to random similarities due to short size of reads. A total of 84,062 of 1,224,713 N-herpesvirus reads were assembled to herpesvirus random reference while 1,140,651 reads were not assembled (Figure 11 B) and 1 of 23,998 reads were assembled to papillomavirus random reference while 23,997 reads were not assembled (Figure 11 C).

Taken together the results of controls in Figures 9, 10 and 11 show that pure randomness does not explain the results observed in Figure 1. In Figure 9 we tested the level of coverage when Neanderthal reads are mapped to random mock sequences and compare the coverage level with real assembly references. In Figure 10 we took the reads mapped to viral genomes and mapped them to the human genome (hg38). In Figure 11 the reads that mapped to random references were used to map to the human genome (hg38). The analysis in Figure 9 shows that the coverage of mapping to viral genomes is higher than with random references. The test in Figure 10 shows that none of the reads that mapped to real viral genomes mapped to human nuclear sequences, only few reads mapped to mitochondrial DNA. In Figure 11 the analysis revealed that the reads that mapped to random viral references did not map to human nuclear genome and only very few sequences mapped to portions of the mitochondrial DNA. Therefore, random effects due to artifactual similarities due to the read lengths here used did not explain *per se* the mapping of Neanderthal reads to viral genomes used for mapping (Figure 1).

3.8. Control with a nonpersistent DNA virus

To test whether a nonpersistent DNA virus would show a positive result with Neanderthal reads we mapped the Neanderthal reads against the human parvovirus B19 reference sequence (NC_000883.2) [40]. Parvovirus B19 has a DNA genome, only infects humans and is known for causing disease in the pediatric population and rarely in adults. It is the cause of the childhood rash called fifth disease or erythema *infectiosum*, or "slapped cheek syndrome" [41]. Individuals with anti-parvovirus B19 IgG antibodies are generally considered immune to recurrent infection [42]. The assembly was done with the same set of Neanderthal reads used for mapping of adenovirus, herpesvirus and papillomavirus shown in Figure 1. To estimate the signal-to-noise level, mapping against parvovirus also included 3 random mock sequences whose length and percentage of bases A, C, G and T were identical with the parvovirus reference sequence. By signal we mean the mapping coverage and distribution against the parvovirus reference while by noise we mean the coverage and distribution against random references. A positive result is when the mapping statistics are above the observed by using random references while a negative result is when the mapping statistics are equivalent to observed for random references. The assembly by mapping with parvovirus (Figure 12) shows that 714 Neanderthal reads map to parvovirus reference with a mean coverage of 2x, standard deviation of 9.2, and 100% single

strand (Figure 12 A). Mapping of Neanderthal reads against random references 1, 2 and 3 (Figure B, C and D) shows mean coverages of 3.5x (1,271 reads, SD=16.9), 7.2x (2,595 reads, SD=41.6) and 2.7x (975 reads, SD=10). The distribution of reads along the references also show that the pattern observed for the parvovirus is equivalent to the pattern for random references and therefore the result is negative, in other words, parvovirus DNA cannot be detected in Neanderthal reads above the level expected by random noise. The Welch's *t*-test was used to determine if mapping of Neanderthal reads with human parvovirus B19 produced assemblies with higher coverage as compared to random reference sequences (Table 4). Comparison of the *t* statistics of parvovirus B19 and random references shows that parvovirus B19 assemblies were always below (random references 1 and 2) or equal to random references (random reference 3) with proper *p*-values. Therefore, the qualitative pattern observed in Figure 12 is statistically supported (Table 4). When parvovirus B19 is tested against random Ref. 1 and Ref. 2 the coverage with random references is statistically higher than with the real reference while testing against Ref. 3 shows that mapping coverage with random reference is statistically identical with the real reference. In all three cases the coverage with the real reference is never statistically higher than with the random references and therefore the result is negative, in other words the null hypothesis of randomness can't be refuted.

Table 4. Comparison of mapping Neanderthal sequencing reads with real human parvovirus B19 () references and random references. SD = standard deviation and # Rreads = mapped reads. Random reference sequences here used are in Supplementary Data. Column "Cov." indicates "Coverage", column "*t*" indicates the'Welch's *t*-test *t* statistics, *c.v.*=critical value for infinite degrees of freedom and *p*-value (one-tailed) is the probability of rejecting the correct null hypothesis that both samples are statistically equivalent.

Reference	**Cov. Mean**	**Cov. SD**	**# reads**	***t***	***c.v.***	***p*-value**
Parvovirus B19	2.0	9.2	714	-	-	-
Random Ref. 1	3.5	16.9	1,271	-2.56	1.64	0.005267
Random Ref. 2	7.2	41.6	2,595	2.25	1.64	0.012046
Random Ref. 3	2.7	10.0	975	-1.48	1.64	0.068387

3.9. Herpesvirus type 1 (HSV-1) and type 2 (HSV-2)

Regarding the detection of herpesvirus types in Neanderthal genome data, the initial STAT binning analysis suggested that HSV-2 was the type detected in the Chagyrskaya 7 individual (Table 1) while our genome mapping used an assembly reference sequence of the type 1 (HSV-1) that was detected in Chagyrskaya 1 (Table 1). To test whether HSV-1 and HSV-2 can be discriminated with assembly by mapping, a total of 1,221,442,735 Neanderthal Chagyrskaya 7 reads were assembled by mapping using reference sequences of HSV-1 (NC_001806.2) and HSV-2 (NC_001798.2) (Figure 13). A total of 1,221,263,284 reads were not assembled to either viral reference. The HSV-1 mapping produced a contig with mean coverage of 8.1x (SD=19, Q40=98.6%) with 74,405 reads, while HSV-2 produced a contig with mean coverage of 11.4x (SD=29.2, Q40=98.6%) with 105,046 reads. The alignment of reference genomes for HSV-1 and HSV-2 shows that the global sequence nucleotide similarity is 73% with 80% percent in the most conserved domains and 50% in the most divergent regions (Supplementary Figure S16). Given the read lengths and algorithms used it is possible

that reads mapping to the conserved regions might be responsible for identifying either HSV-1 and HSV-2 in Neanderthal samples. Another possibility is that mixed infections with HSV-1 and HSV-2 explain these results. Mixed infections with HSV-1 and HSV-2 occur in 6-11% HSV infections, according to PCR testing with type specific primers [43,44].

4. Discussion

In the present study we present an initial assessment on the possible presence of viral DNA remnants in Neanderthal sequencing data. For this we investigated the STAT SRA analysis of Neandertal reads of the Chagyrskaya Cave. This binning method indicated that approximately 0.01% of Neanderthal reads, more specifically, Chagyrskaya 1 and 7, contained potential viral sequences. To investigate in more detail this possibility and determine the exact number of viral reads in these data we mapped Neanderthal reads to viral reference sequences. The initial SRA analysis indicated that double stranded DNA viruses that produce lifelong infections could be present in these Neanderthal samples as previously hypothesized [6]. Mapping of Neanderthal reads to adenovirus, herpesvirus and papillomavirus reference genomes produced consistent assemblies by mapping and the inferred consensus sequences were analyzed and compared to extant closest relatives. We show that changes of Neanderthal viruses are distributed along the sequences and nonsynonymous changes are observed in coding regions of surface proteins such as glycoproteins gB (gene UL27), gD (gene US6), gH (gene UL22) and gL (gene UL1) involved in cell adhesion and infectivity of herpesviruses (Supplementary Figures S13 and S14) [45]. Changes in major surface viral proteins of adenovirus (genes L2, L3 and L5) and oncoproteins of papillomavirus (genes E6 and E7) have been revealed as well (Figure 8, Supplementary Figures S12 and S15). A detailed analysis of these amino acid changes and their putative functional effects on these proteins is a matter for future research. Most positions of Neanderthal viruses here inferred are conserved and therefore several segments, surrounding variable regions and regions with low coverage in mapping, can be used as targets for PCR primer design. The amplicons obtained by PCR from Neanderthal purified DNA can be sequenced to fill low coverage regions, which will allow the full reconstruction of these paleoviruses and their Neanderthal epidemiology. Further detailed characterization of variable regions of these ancient viral genomes might reveal Neanderthal adaptive signatures.

In the case of Chagyrskaya 7, because DNA was isolated from a thoracal vertebral process it is likely that DNA from a latent viral infection, especially Herpesvirus, was coisolated with Neanderthal DNA. Herpes simplex virus 1 and 2 (HSV-1 and HSV-2) are part of the alpha subfamily of human herpesviruses. The lytic phase of infection occurs within mucoepithelial cells, and the latent infection occurs in neurons. HSV-1 establishes latency preferentially in sensory neurons of peripheral ganglia and is associated with orofacial infection. While HSV-1 is isolated from the upper half of the body, innervated by the trigeminal ganglia, HSV-2 is isolated from the lower half of the body innervated by the sacral ganglia, which explains the clinical pattern more associated with genital herpes infection [43]. However, recent studies indicate this viral tropism is not always applicable and mixed infections occur both in ocular and genital herpes patients [43]. Human adenoviruses persistence and reactivation occurs in the gastrointestinal tract and the best-documented sites of latency include tonsillar and adenoidal T lymphocytes [46]. Latency of papillomavirus is associated with basal epithelial stem cell pool and periodically induced to reactivate when the stem cell divides, and one daughter cell is committed to terminal differentiation and induction of the viral life cycle. Tissue resident memory T-cells are

hypothesized to control these periodic reactivation episodes and thus limit their duration [47]. Taken together these data suggest that once infected by these viruses the Neanderthal host might have remained with viral particles in body tissues for its whole life. Depending on viral loads it is likely that sufficient viral DNA remained *postmortem* and was coisolated and sequenced. Although it is relevant to determine if Neanderthal individuals are infected with HSV-1 and HSV-2 it seems that with the methodology used in this study it is not possible to give a definitive answer because Neanderthal reads could map with approximate probability, depending on the binning algorithm and parameters, to conserved regions (Figure 14) and therefore, a deeper specific analysis should be carried in future studies. Also, mixed infections with HSV-1 and HSV-2, are not uncommon (approximately 10% patients) and these are detected in samples from patients by PCR amplification, using HSV-1/HSV-2 specific primers [43,44]. Our result in Figure 13 is consistent with the result in Table 1 which suggests that HSV-2 is probably the type present in Chagyrskaya 7. Our result in Table 2, with more reads mapped to HSV-1 reference sequence, is explained by the fact that the reads used in this analysis were from the assembly in Figure 1 and therefore enriched with HSV-1 reads. It must be noted however, that the objective of the analysis presented in Table 2, was to show that reads that mapped to the assembly reference human alphaherpesvirus 1 (MN136523), used in Figure 1, also mapped consistently to the taxonomic references (NCBI RefSeq) of HSV-1 and HSV-2.

Analysis of deamination in the terminal ends of reads assembled to reference sequences show that the expected pattern of aDNA is observed (Figure 6). This is central to rule out contamination with modern extant viral DNA. Furthermore, the sequence differences observed between inferred N-adenovirus, N-herpesvirus and N-papillomavirus suggest that they are distant from any extant viral genomes of the respective groups (Figure 7). Neighbor-Joining trees obtained as results of blast analysis (Supplementary Figures S1, S2 and S3) were compared with topologies inferred by maximum likelihood (Supplementary Figures S4, S5 and S6) and Bayesian methods (Figure 7) detailing the relationships of the Neanderthal viruses with their respective ten closest extant relatives. The inferred Neanderthal sequences were separated in 100% bootstrap pseudoreplicates in maximum likelihood trees (Supplementary Figures S4, S5 and S6). Bayesian trees also supported the topologies and distances of inferred Neanderthal sequences and their extant relatives (Figure 7).

Although the assembly of adenovirus and papillomavirus showed that C-to-T deamination was significantly more abundant in the 5′ overhangs as compared to the other changes away from termini, it can be argued that the C-to-T substitution rate of Neanderthal reads was around 30 % higher than the results described by Skov et al. [22]. The estimates of aDNA damage here calculated (Figure 6), using MapDamage 2.0 [26], which implements a fast approximation of the DNA damage by a Bayesian framework and extends the Briggs-Johnson model of DNA damage [48]. The MapDamage model employs a maximum likelihood to estimate DNA damage parameters. In the paper describing MapDamage 2.0 [26] the authors compare the two methods (in Supplementary Information, Figures S2 and S3) which shows that the estimates of cytosine deamination rates obtained by the MapDamage 2.0 program are consistently higher than those obtained by the Briggs-Johnson model. Since Skov et al. [22] estimates are based on Briggs et al. [48], it is expected that a difference in the estimates of the parameters based on these two different methods will be observed.

As discussed by Briggs et al. [48] When strand-equivalent reciprocal nucleotide substitutions are analyzed in DNA sequences from Pleistocene organisms, C-to-T changes are more frequent than G-to-A changes. Also, the rates of G-to-A and C-to T changes are higher than the rates of the other two transitions in contrast to DNA sequences determined from modern DNA. Deamination of cytosine (C) residues to uracil (U) residues in ancient DNA is responsible for the excess of C-to-T substitutions. The G-to-A could be caused by deamination of guanine residues to xanthine (X) residues,

which are read by methods of DNA sequencing as adenine residues (A), and therefore causing the observed G-to-A misincorporations in ancient DNA.

It must also be noted that in our analysis we show that C-to-T changes are higher than G-to-A changes and both are much higher in terminal positions than central positions (Figure 6). Also, in adenovirus and papilloma viruses the C-to-T and G-to-A are less frequent in the internal sections as compared to terminal positions which are fully consistent with ancient DNA patterns (Figure 6 A and 6 C. In the herpesvirus the C-to-T and G-to-A are slightly higher than other changes in internal sections and significantly higher in terminal positions (Figure 6 B).

The comparison of Neanderthal viral C-to-T and G-to-A in 5′-termini and 3′-termini with negative control (Figure 6 A, B, C and D) shows that aDNA patterns are significantly different, as the C-to-T and G-to-A transitions in termini are 10x higher. The negative control consists of 6,000,000 reads of a present-day Adenovirus infected cell culture and sequenced. The comparison od Neanderthal viruses C-toT and G-to-A patterns with the positive control shows that the N-adenovirus, N-herpesvirus and N-papillomavirus reads have about half of C-to-T at termini (5′ and 3′) but about 5 times more G-to-A. The positive control consists of Neanderthal reads (from Chagyrskaya 7, the same pool used for viruses) mapped to the human reference mtDNA (Figure 7 E). Although all viral patterns and mtDNA patterns are consistent with aDNA material, the base frequencies in the different references might explain the difference since in mtDNA G+C=44.4 (C=31.3%, G=13.1%), and therefore more C-to-T than G-to-A is expected in mtDNA reads termini because the sheer number of Cs as compared to Gs is more than double. In adenovirus G+C=51.2% (C=25.9%, G=25.3%) in herpesvirus (HSV-1) G+C=68.3% (C=33.8%, G=25.3%), HSV-2 G+C=70.4% (C=35%, G35.3%) and in papillomavirus G+C=42.4 (C=20%, G=22.4%).

The possibility that simple random similarities between Neanderthal reads and viral reference sequences could explain the mapping results shown in Figure 1, was tested with random mock reference sequences of adenovirus, herpesvirus, and papillomavirus. Size and base composition of random references were identical with actual counterparts. Mapping random references would indicate the random noise in the analysis (Table 3). As shown, the assemblies obtained with real sequences have more reads mapped and higher coverage than the random sequences for all three viruses as assessed by the Welch's *t*-test (Table 3). For adenovirus, the real reference has coverage 102.2 (180,419 reads mapped) as opposed to coverage of 62.2 (126,613 reads mapped) for random reference. For herpesvirus, the real reference has coverage 171.2 (1,224,613) as compared to coverage 154.8 (1,166,326 reads mapped). For papillomavirus, the real reference has coverage 115.4 (23,998 reads mapped) as opposed to random reference coverage 51.0 (22,682 reads mapped). All p-values of these comparisons above reject the null hypothesis (that the samples are equal) with *p*-value = 0 (Table 3). Also, in assemblies of random references the number of regions with single strand coverage is higher than with real references (Figure 1 and Figure 9). This indicates that random similarities between Neanderthal reads, and viral sequences are substantial and that the *bona fide* signal obtained by mapping of Neanderthal reads and actual viral genomes is above the level expected by random noise.

A negative control for nonpersistent DNA virus was included (Figure 12). This analysis revealed the relevance of using random reference sequences for assessing the level of random noise in this type of analysis. Mapping of Neanderthal reads to human parvovirus B19 reference is within the expected for random similarities of short reads as shown by the Welch's *t*-test (Table 4). Also, it indicates that not all DNA viruses are picked by our approach, but it is likely that our results reveal bona fide DNA viruses that produced persistent infection in the Neanderthal individual tested. The impact of this finding on populations, as suggested by Wolff and Greenwood [6] remains to be tested in future studies.

Additional tests for random similarities and short read lengths were performed, such as mapping to the human genome (hg38) (Figure 10) and reassembly to random reference sequences (Figure 11). Mapping of Neanderthal viral

reads of assemblies in Figure 9 do not map to any human nuclear chromosomes and only a small number of reads map to mtDNA (Figure 10). Regarding mapping to random reference sequences, we show that when just the subset of reads that map to viral references are considered in the analysis, very few reads map to random references which strongly suggests a particularly good signal-to-noise ratio in our analysis. Taken together, these tests (Figure 10 and 11) further suggest that our results are not caused by random mapping and short read lengths. The difference between mapping to random reference sequences observed in Figure 9 and Figure 11 is explained by the high abundance of non-human DNA in the Neanderthal SRA raw data (Supplementary Figures S10 and S11). The taxonomic analysis of sequencing reads, based on MinHash-based k-mers implemented in SRA STAT, reveals that more than 90% of reads in runs of Chagyrskaya 7 are not human DNA. In several run files only between 0.5% to 1% are human DNA. Viruses are no more than 0.01% and more than 90% are unidentified sequences, comprising the dark matter (unidentified spots) as shown in SRA Krona view (Supplementary Data S16) [49]. The analysis in Figure 9 includes "dark matter" reads while the analysis in Figure 11 these reads are filtered-out.

The assessment of phylogenetic relationships of Neanderthal viral sequences to modern human counterparts and viral sequences of non-human primates shows that the inferred Neanderthal sequences follow the expected phylogenetic pattern, and no further analysis is necessary within the scope of this study. Therefore, as expected, and regarding RefSeq non-human viral sequences, when modern humans and Neanderthal viral sequences are analyzed, it is evident that chimpanzee's viral sequences, and other primates, function as outgroup to the Neanderthal-modern human ingroup (Supplementary Figure S11, Supplementary Tables 10, 11 and 12). Murid sequences are a more distant outgroup relative to the primate ingroup. This is the expected pattern. We propose that a deeper, more detailed phylogenetic analysis in a future study, would require the sequencing of viral genomes directly amplified from Neanderthal DNA to close the low coverage gaps and confirm synonymous and non-synonymous SNPs to determine the intraspecific and interspecific relationships more precisely between these viral sequences.

Regarding the weight of deamination patterns in non-synonymous SNPs detected in Neanderthal viral genes shows that the effect of deamination is minimal. Most C-to-T and G-to-A changes are below the threshold for consensus sequences generation. In other words, the C-to-T and G-to-A changes are located preferentially at the termini of individual reads in a quantity and pattern compatible with aDNA, as show above, but not all reads have these changes (Supplementary Figure S10). Although the deamination pattern is present, it is below 50% frequency. Therefore, the deamination bias does not have to be compensated in phylogenetic analysis and does not compromise the inference of amino acid changes in viral DNA sequences. Deamination in ancient DNA is consistently around 15-20% of C-to-T changes in the termini of reads [8]. In the case of HSV-1 retrieved from human remains dating back to 500 years old and 1,500 years old the levels of C-to-T deamination in 5' termini were 20% and 12.5%, respectively. Although C-to-T deamination in 5' termini is higher than in the middle of the reads it is not 100% and therefore the C-to-T deamination can be differentiated from a bona fide C-to-T transition and, therefore, does not affect the inference of synonymous and non-synonymous changes in Neanderthal viral sequences.

A tentative classification of the Neanderthal viruses, inferred from our data, suggests that N-adenovirus can be classified within family Adenoviridae, genus *Mastedenovirus* and in species human Adv-B or Adv-C. The guidelines of the International Committee of Taxonomy of Viruses (ICTV, 9th Report, 2011) define that "if virus neutralization data are available, lack of cross-neutralization combined with a phylogenetic distance of more than 15% separates two types into different species. If the phylogenetic distance is less than 5%, any additional common grouping criteria from the list above may classify separate types into the same species even if they were isolated from different hosts" [50]. Therefore, based on these guidelines there is no justification to propose that N-adenovirus constitutes a prototype new species. Neutralization data are not available, and the phylogenetic distance is not superior to 15% either at nucleotide

or aminoacid levels in any regions of the genome. The distances between N-adenovirus and any of its extant relatives is 5.1% on average (Supplementary Table 7). In CDS E2B (DNA Polymerase) the identity between N-adenovirus and Human adenovirus 7 (*Human mastadenovirus B* or HAdV-7) is 95.7% (4.3% distance) and the ICTV species demarcation criteria in this genus depends on phylogenetic distance (>5–15%, based primarily on distance matrix analysis of the DNA polymerase amino acid sequence) and host range. Therefore, based on both criteria N-adenovirus should be maintained in the same species of *Human mastadenovirus* since distances are within the 5% limit and infect the same host. Although modern humans and Neanderthals constitute two different subspecies, they belong in the same species, namely, *Homo sapiens sapiens* and *Homo sapiens neanderthalensis*.

Nucleotide identity between N-herpesvirus and HSV-1 and HSV-2 doesn't suggest that N-herpesvirus is a prototype new species as shown in Figure 13. According to ICTV guidelines the species demarcation, at present, uses sequence data flexibly to support taxonomic proposals based on phylogenetic grouping of viruses, does not require the availability of a complete genome sequence, does not depend on any specified gene or group of genes, and does not specify genetic distance thresholds for taxa differentiation.

In the case of N-papillomavirus the nucleotide distance to human papillomavirus strains RTRX7 and mSK144 is 18% at the whole genome level. The N-papillomavirus seems a bona fide member of Betapapilomavirus 1 and not a new prototype species. According to ICTV the species demarcation criteria is that for a putative novel papillomavirus genome with complete genome sequence data available that is <70% related to papillomaviruses within the genus Betapapillomavirus, which is not observed in our data where whole genome nucleotide identities range from 74 to 82% within the group (Figure 7).

The refinement of the classification of these three Neanderthal recovered viruses ultimately depends on the possibility to amplify their genome segments by PCR (using primers based on highly conserved regions) directly from DNA extracted from Neanderthal bone material, sequencing of these amplicons and incorporation of these reads into our assemblies to resolve nucleotide ambiguities and improve the sequencing accuracy of regions with low coverage.

On the probability of finding all three viruses in the same sample of a single individual, it has been shown that single individuals of present-day humans had been exposed to about 10 viral species over their lifetime [51]. Particularly in the case of lifelong infections, it is not surprising to verify that an individual had been exposed to these three viruses because they are highly contagious and associated with social behavior [6]. As a matter of fact, proviral remnants of ancient retroviral infection of germ cells now make up about 8% of the human genome in the form of human endogenous retroviruses (HERVs) which indicates that exposition of humans to viral agents is extensive [52]. A single sample of one individual might contain remnants of lifelong infections and this probabilitydependensd on many external factors associated with the distribution of these viruses in tissues as well as environmental factors affecting the preservation of this material. The assessment of individual coinfection probability with these three different viruses, based on the data at hand, is highly speculative and depends on the specific epidemiological patterns of the Chagyrskaya population, a conjecture that is beyond the scope of this study.

Variable regions identified in these reconstructed viral genomes could reveal Neanderthal adaptive signatures, such as nonsynonymous substitutions in viral surface proteins that interact with host cell receptors. Conserved regions can be exploited as potential PCR targets for amplification of these viral genomes from Neanderthal isolated DNA.

5. Conclusions

In this study we present proof of concept on the possibility to detect vestigial viral DNA from sequencing reads of Neanderthal genome projects. This would be the first step to address the hypothesis that DNA viruses that produce persistent infections might have had an impact on the extinction of Neanderthals. We show that remnants of adenovirus, herpesvirus and papillomavirus could be detected in Neanderthal genome data, above the level of pure random similarities, given the read lengths and assembly coverages here considered. Under the same conditions a nonpersistent DNA virus was not detected above the level of pure random nucleotide similarity. These results could be taken in the context of the hypothesis by Wolff and Greenwood [6] that adenovirus, herpesvirus and papillomavirus infected Neanderthals which, in turn, could have contributed to processes associated with Neanderthal extinction.

Supplementary Materials

The supporting figures and tables can be downloaded from the journal website.

Author Contributions

Conceptualization, M.R.S.B. and R.C.F.; methodology, G.V A., M.R and F.A.; performed the experiments, R.C.F., M.R.S.B and G.V A.; analyzed the data contributed to fruitful discussions and key ideas, R.C.F, M.R.S.B., G.V.A., M.R. and F.A.; writing-original draft preparation, M.R.S.B.; writing-review and editing, M.R.S.B., R.C.F and F.A.. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Brazil) grant 20/08943-5 to M.R.S.B. and F.A. and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) grant 311154/2021-2 to M.R.S.B.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data produced by analyses in this study are available upon request after acceptance of the manuscript for publication.

Acknowledgements

The authors thank Prof. Fulvio A. Scorza, director of *Escola Paulista de Medicina* at *UNIFESP*, for continual support to our group.

References

1. Banks, W.E.; d'Errico, F.; Peterson, A.T.; Kageyama, M.; Sima, A.; Sánchez-Goñi, M.-F. Neanderthal Extinction by Competitive Exclusion. *PLoS One* **2008**, *3*, e3972, doi:10.1371/journal.pone.0003972.
2. Jones, D. The Neanderthal Within. *New Scientist* **2007**, *193*, 28–32, doi:10.1016/S0262-4079(07)60550-8.
3. Dalén, L.; Orlando, L.; Shapiro, B.; Brandström-Durling, M.; Quam, R.; Gilbert, M.T.P.; Díez Fernández-Lomana, J.C.; Willerslev, E.; Arsuaga, J.L.; Götherström, A. Partial Genetic Turnover in Neandertals: Continuity in the East and Population Replacement in the West. *Molecular Biology and Evolution* **2012**, *29*, 1893–1897, doi:10.1093/molbev/mss074.
4. Greenbaum, G.; Getz, W.M.; Rosenberg, N.A.; Feldman, M.W.; Hovers, E.; Kolodny, O. Disease Transmission and Introgression Can Explain the Long-Lasting Contact Zone of Modern Humans and Neanderthals. *Nat Commun* **2019**, *10*, 5003, doi:10.1038/s41467-019-12862-7.
5. Enard, D.; Petrov, D.A. Evidence That RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **2018**, *175*, 360-371.e13, doi:10.1016/j.cell.2018.08.034.
6. Wolff, H.; Greenwood, A.D. Did Viral Disease of Humans Wipe out the Neandertals? *Medical Hypotheses* **2010**, *75*, 99–105, doi:10.1016/j.mehy.2010.01.048.
7. Chayavichitsilp, P.; Buckwalter, J.V.; Krakowski, A.C.; Friedlander, S.F. Herpes Simplex. *Pediatrics In Review* **2009**, *30*, 119–130, doi:10.1542/pir.30-4-119.
8. Guellil, M.; van Dorp, L.; Inskip, S.A.; Dittmar, J.M.; Saag, L.; Tambets, K.; Hui, R.; Rose, A.; D'Atanasio, E.; Kriiska, A.; et al. Ancient Herpes Simplex 1 Genomes Reveal Recent Viral Structure in Eurasia. *Science Advances* **2022**, *8*, eabo4435, doi:10.1126/sciadv.abo4435.
9. Van Doorslaer, K.; Chen, Z.; Bernard, H.-U.; Chan, P.K.S.; DeSalle, R.; Dillner, J.; Forslund, O.; Haga, T.; McBride, A.A.; Villa, L.L.; et al. ICTV Virus Taxonomy Profile: Papillomaviridae. *J Gen Virol* **2018**, *99*, 989–990, doi:10.1099/jgv.0.001105.
10. Moens, U.; Calvignac-Spencer, S.; Lauber, C.; Ramqvist, T.; Feltkamp, M.C.W.; Daugherty, M.D.; Verschoor, E.J.; Ehlers, B. ICTV Virus Taxonomy Profile: Polyomaviridae. *J Gen Virol* **2017**, *98*, 1159–1160, doi:10.1099/jgv.0.000839.
11. Dhingra, A.; Hage, E.; Ganzenmueller, T.; Böttcher, S.; Hofmann, J.; Hamprecht, K.; Obermeier, P.; Rath, B.; Hausmann, F.; Dobner, T.; et al. Molecular Evolution of Human Adenovirus (HAdV) Species C. *Sci Rep* **2019**, *9*, 1039, doi:10.1038/s41598-018-37249-4.
12. Hassan, M.M.; Li, D.; El-Deeb, A.S.; Wolff, R.A.; Bondy, M.L.; Davila, M.; Abbruzzese, J.L. Association Between Hepatitis B Virus and Pancreatic Cancer. *J Clin Oncol* **2008**, *26*, 4557–4562, doi:10.1200/JCO.2008.17.3526.
13. Speck, S.H.; Ganem, D. Viral Latency and Its Regulation: Lessons from the Gamma-Herpesviruses. *Cell Host Microbe* **2010**, *8*, 100–115, doi:10.1016/j.chom.2010.06.014.
14. Sankararaman, S.; Mallick, S.; Dannemann, M.; Prüfer, K.; Kelso, J.; Pääbo, S.; Patterson, N.; Reich, D. The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans. *Nature* **2014**, *507*, 354–357, doi:10.1038/nature12961.

15. Spyrou, M.A.; Bos, K.I.; Herbig, A.; Krause, J. Ancient Pathogen Genomics as an Emerging Tool for Infectious Disease Research. *Nat Rev Genet* **2019**, *20*, 323–340, doi:10.1038/s41576-019-0119-1.

16. Lion, T. Adenovirus Persistence, Reactivation, and Clinical Management. *FEBS Lett* **2019**, *593*, 3571–3582, doi:10.1002/1873-3468.13576.

17. Houldcroft, C.J.; Underdown, S.J. Neanderthal Genomics Suggests a Pleistocene Time Frame for the First Epidemiologic Transition. *American Journal of Physical Anthropology* **2016**, *160*, 379–388, doi:10.1002/ajpa.22985.

18. Zhong, S.; Naqvi, A.; Bair, E.; Nares, S.; Khan, A.A. Viral MicroRNAs Identified in Human Dental Pulp. *J Endod* **2017**, *43*, 84–89, doi:10.1016/j.joen.2016.10.006.

19. Kazi, M.M.A.G.; Bharadwaj, R.; Bhat, K.; Happy, D. Association of Herpes Viruses with Mild, Moderate and Severe Chronic Periodontitis. *J Clin Diagn Res* **2015**, *9*, DC05-08, doi:10.7860/JCDR/2015/13781.6187.

20. Baringer, J.R.; Swoveland, P. Recovery of Herpes-Simplex Virus from Human Trigeminal Ganglions. *N Engl J Med* **1973**, *288*, 648–650, doi:10.1056/NEJM197303292881303.

21. Juhl, D.; Mosel, C.; Nawroth, F.; Funke, A.-M.; Dadgar, S.M.; Hagenström, H.; Kirchner, H.; Hennig, H. Detection of Herpes Simplex Virus DNA in Plasma of Patients with Primary but Not with Recurrent Infection: Implications for Transfusion Medicine? *Transfus Med* **2010**, *20*, 38–47, doi:10.1111/j.1365-3148.2009.00951.x.

22. Skov, L.; Peyrégne, S.; Popli, D.; Iasi, L.N.M.; Devièse, T.; Slon, V.; Zavala, E.I.; Hajdinjak, M.; Sümer, A.P.; Grote, S.; et al. Genetic Insights into the Social Organization of Neanderthals. *Nature* **2022**, *610*, 519–525, doi:10.1038/s41586-022-05283-y.

23. Katz, K.S.; Shutov, O.; Lapoint, R.; Kimelman, M.; Brister, J.R.; O'Sullivan, C. STAT: A Fast, Scalable, MinHash-Based k-Mer Tool to Assess Sequence Read Archive next-Generation Sequence Submissions. *Genome Biology* **2021**, *22*, 270, doi:10.1186/s13059-021-02490-0.

24. Bushnell, B.; Rood, J.; Singer, E. BBMerge – Accurate Paired Shotgun Read Merging via Overlap. *PLOS ONE* **2017**, *12*, e0185056, doi:10.1371/journal.pone.0185056.

25. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*; Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 2014;

26. Jónsson, H.; Ginolhac, A.; Schubert, M.; Johnson, P.L.F.; Orlando, L. mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters. *Bioinformatics* **2013**, *29*, 1682–1684, doi:10.1093/bioinformatics/btt193.

27. Morgulis, A.; Coulouris, G.; Raytselis, Y.; Madden, T.L.; Agarwala, R.; Schäffer, A.A. Database Indexing for Production MegaBLAST Searches. *Bioinformatics* **2008**, *24*, 1757–1764, doi:10.1093/bioinformatics/btn322.

28. Saitou, N.; Nei, M. The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* **1987**, *4*, 406–425, doi:10.1093/oxfordjournals.molbev.a040454.

29. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) [All Versions, 1/4] Available online: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=H1jbCPkAAAAJ&citation_for_view=H1jbCPkAAAAJ:_FxGoFyzp5QC (accessed on 23 November 2023).

30. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst Biol* **2012**, *61*, 539–542, doi:10.1093/sysbio/sys029.

31. Wit, E.; Heuvel, E. van den; Romeijn, J.-W. 'All Models Are Wrong...': An Introduction to Model Uncertainty. *Statistica Neerlandica* **2012**, *66*, 217–236, doi:10.1111/j.1467-9574.2012.00530.x.

32. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway - Mark A. Miller, Terri Schwartz, Brett E. Pickett, Sherry He, Edward B. Klem, Richard H. Scheuermann, Maria Passarotti, Seth Kaufman, Maureen A. O'Leary, 2015 Available online: https://journals.sagepub.com/doi/10.4137/EBO.S21501 (accessed on 23 November 2023).

33. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.

34. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **2010**, *5*, e9490, doi:10.1371/journal.pone.0009490.

35. Okonechnikov, K.; Golosova, O.; Fursov, M.; the UGENE team Unipro UGENE: A Unified Bioinformatics Toolkit. *Bioinformatics* **2012**, *28*, 1166–1167, doi:10.1093/bioinformatics/bts091.

36. Welch, B.L. The Generalization of 'Student's' Problem When Several Different Population Varlances Are Involved. *Biometrika* **1947**, *34*, 28–35, doi:10.1093/biomet/34.1-2.28.

37. Kulanayake, S.; Tikoo, S.K. Adenovirus Core Proteins: Structure and Function. *Viruses* **2021**, *13*, 388, doi:10.3390/v13030388.

38. Connolly, S.A.; Jardetzky, T.S.; Longnecker, R. The Structural Basis of Herpesvirus Entry. *Nat Rev Microbiol* **2021**, *19*, 110–121, doi:10.1038/s41579-020-00448-w.

39. Ault, K.A. Epidemiology and Natural History of Human Papillomavirus Infections in the Female Genital Tract. *Infectious Diseases in Obstetrics and Gynecology* **2006**, *2006*, e40470, doi:10.1155/IDOG/2006/40470.

40. Heegaard, E.D.; Brown, K.E. Human Parvovirus B19. *Clinical Microbiology Reviews* **2002**, *15*, 485–505, doi:10.1128/cmr.15.3.485-505.2002.

41. Vafaie, J.; Schwartz, R.A. Parvovirus B19 Infections. *International Journal of Dermatology* **2004**, *43*, 747–749, doi:10.1111/j.1365-4632.2004.02413.x.

42. Lehmann, H.W.; von Landenberg, P.; Modrow, S. Parvovirus B19 Infection and Autoimmune Disease. *Autoimmunity Reviews* **2003**, *2*, 218–223, doi:10.1016/S1568-9972(03)00014-4.

43. Kaneko, H.; Kawana, T.; Ishioka, K.; Ohno, S.; Aoki, K.; Suzutani, T. Evaluation of Mixed Infection Cases with Both Herpes Simplex Virus Types 1 and 2. *J Med Virol* **2008**, *80*, 883–887, doi:10.1002/jmv.21154.

44. Susloparov, M.A.; Susloparov, I.M.; Zagoruĭko, T.I.; Noskova, N.V.; Makhova, N.M. [Herpes simplex virus type 1 and 2 (HSV-1,2) DNA detection by PCR during genital herpes]. *Mol Gen Mikrobiol Virusol* **2006**, 38–41.

45. Farnsworth, A.; Wisner, T.W.; Webb, M.; Roller, R.; Cohen, G.; Eisenberg, R.; Johnson, D.C. Herpes Simplex Virus Glycoproteins gB and gH Function in Fusion between the Virion Envelope and the Outer Nuclear Membrane. *Proceedings of the National Academy of Sciences* **2007**, *104*, 10187–10192, doi:10.1073/pnas.0703790104.

46. Kosulin, K.; Geiger, E.; Vécsei, A.; Huber, W.-D.; Rauch, M.; Brenner, E.; Wrba, F.; Hammer, K.; Innerhofer, A.; Pötschger, U.; et al. Persistence and Reactivation of Human Adenoviruses in the Gastrointestinal Tract. *Clinical Microbiology and Infection* **2016**, *22*, 381.e1-381.e8, doi:10.1016/j.cmi.2015.12.013.

47. Gravitt, P.E. Evidence and Impact of Human Papillomavirus Latency. *Open Virol J* **2012**, *6*, 198–203, doi:10.2174/1874357901206010198.

48. Briggs, A.W.; Stenzel, U.; Johnson, P.L.F.; Green, R.E.; Kelso, J.; Prüfer, K.; Meyer, M.; Krause, J.; Ronan, M.T.; Lachmann, M.; et al. Patterns of Damage in Genomic DNA Sequences from a Neandertal. *Proc Natl Acad Sci U S A* **2007**, *104*, 14616–14621, doi:10.1073/pnas.0704665104.

49. Wu, D.; Wu, M.; Halpern, A.; Rusch, D.B.; Yooseph, S.; Frazier, M.; Venter, J.C.; Eisen, J.A. Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees. *PLOS ONE* **2011**, *6*, e18011, doi:10.1371/journal.pone.0018011.

50. Virus Taxonomy - 1st Edition | Elsevier Shop Available online: https://shop.elsevier.com/books/virus-taxonomy/king/978-0-12-384684-6 (accessed on 2 March 2024).

51. Xu, G.J.; Kula, T.; Xu, Q.; Li, M.Z.; Vernon, S.D.; Ndung'u, T.; Ruxrungtham, K.; Sanchez, J.; Brander, C.; Chung, R.T.; et al. Comprehensive Serological Profiling of Human Populations Using a Synthetic Human Virome. *Science* **2015**, *348*, aaa0698, doi:10.1126/science.aaa0698.

52. Burn, A.; Roy, F.; Freeman, M.; Coffin, J.M. Widespread Expression of the Ancient HERV-K (HML-2) Provirus Group in Normal Human Tissues. *PLOS Biology* **2022**, *20*, e3001826, doi:10.1371/journal.pbio.3001826.

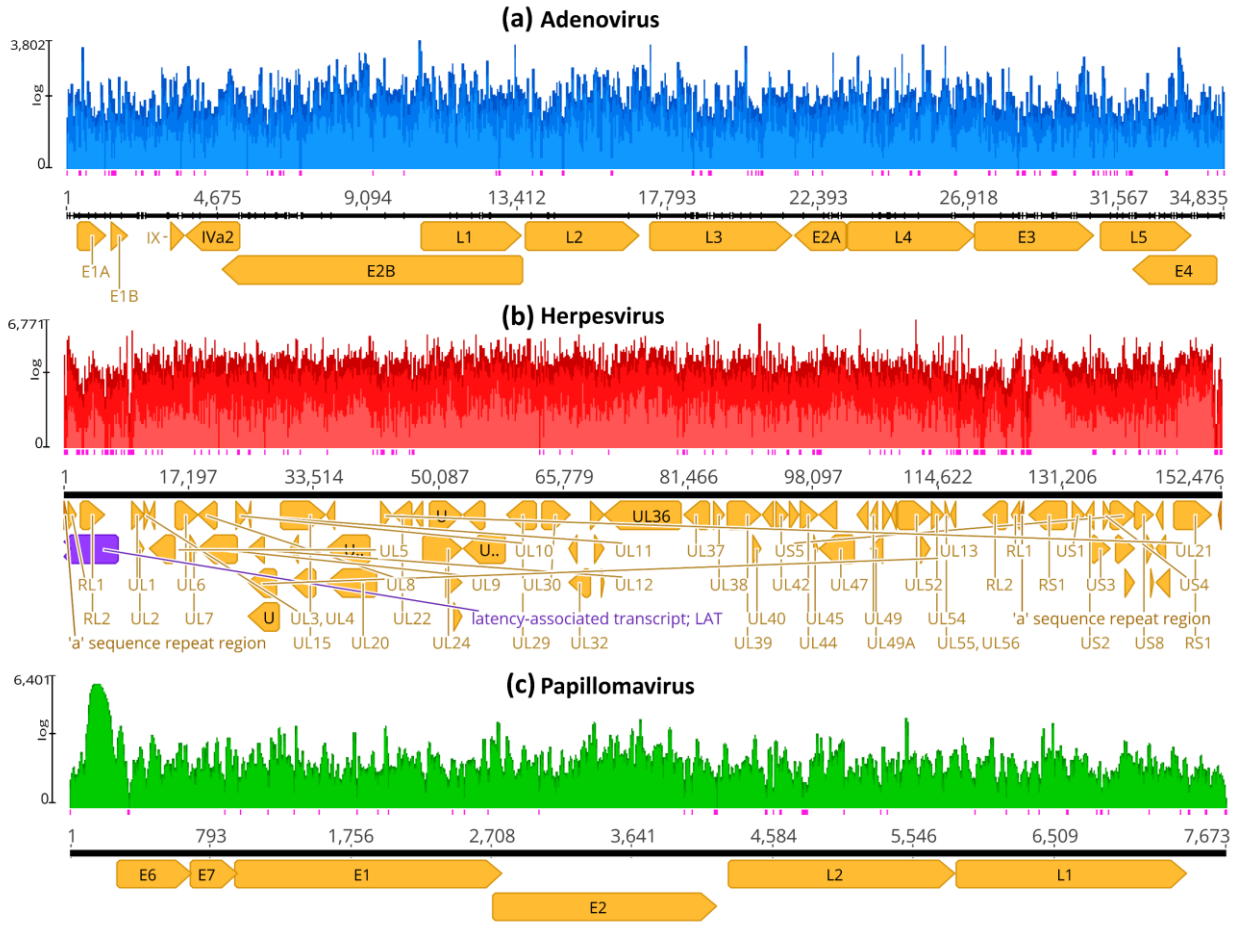

Figure 1. Mapping coverage of herpesvirus, papillomavirus and adenovirus using Neanderthal sample derived reads. Reference mapping was done using BBMap [24,25] (https://sourceforge.net/projects/bbmap/). In (a) is the mapping of 180,419 Neanderthal reads (mean coverage 102.2) to adenovirus assembly reference (or template) (KX897164). In (b), the mapping of 1,224,413 reads (mean coverage 171.2) to herpesvirus assembly reference (MN136523). In (c), the mapping of 23,998 Neanderthal reads (mean coverage 115.4) to papillomavirus assembly reference (X74466). Pink vertical bars below the assemblies indicate single strand regions and orange arrows indicate the protein coding genes. The mapping was performed with the same read dataset and these three reference (template) sequences simultaneously, as detailed in Material and Methods, but each depiction shows the corresponding assembly reference sequence of each bam file with the corresponding mapped reads.

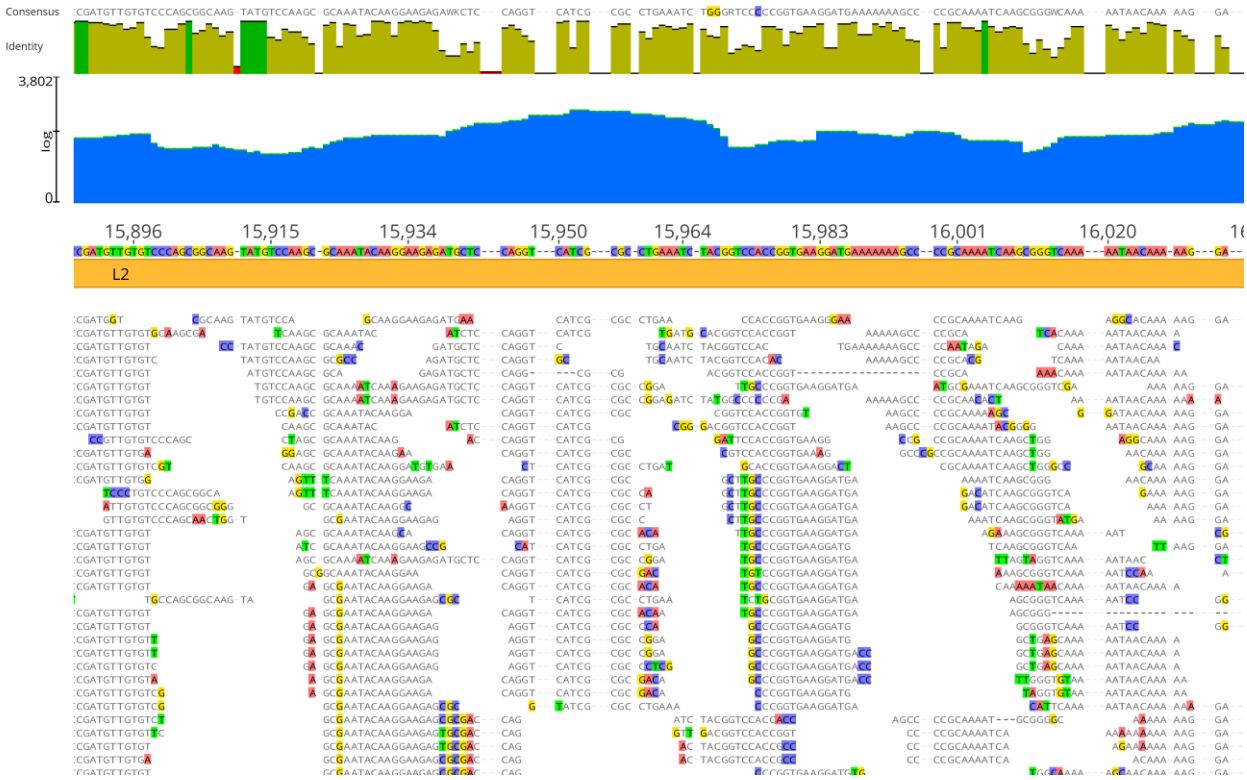

Figure 2. Detail of Neanderthal reads (Chagyrskaya 7) mapped to adenovirus assembly reference genome sequence (GenBank KX897164). Blue area indicates coverage and Dark green vertical bars indicate 100% similarity, olive green vertical bars indicate nucleotide identity <100% and ≥30%, red vertical bars indicate <30% identity. Colored residues are those that differ from the assembly reference. The consensus (50% strict rule) is above nucleotide identity bars. The mapping was performed with three reference sequences simultaneously, as detailed in Materials and Methods, but the depiction is with the adenovirus assembly reference sequence for clarity.

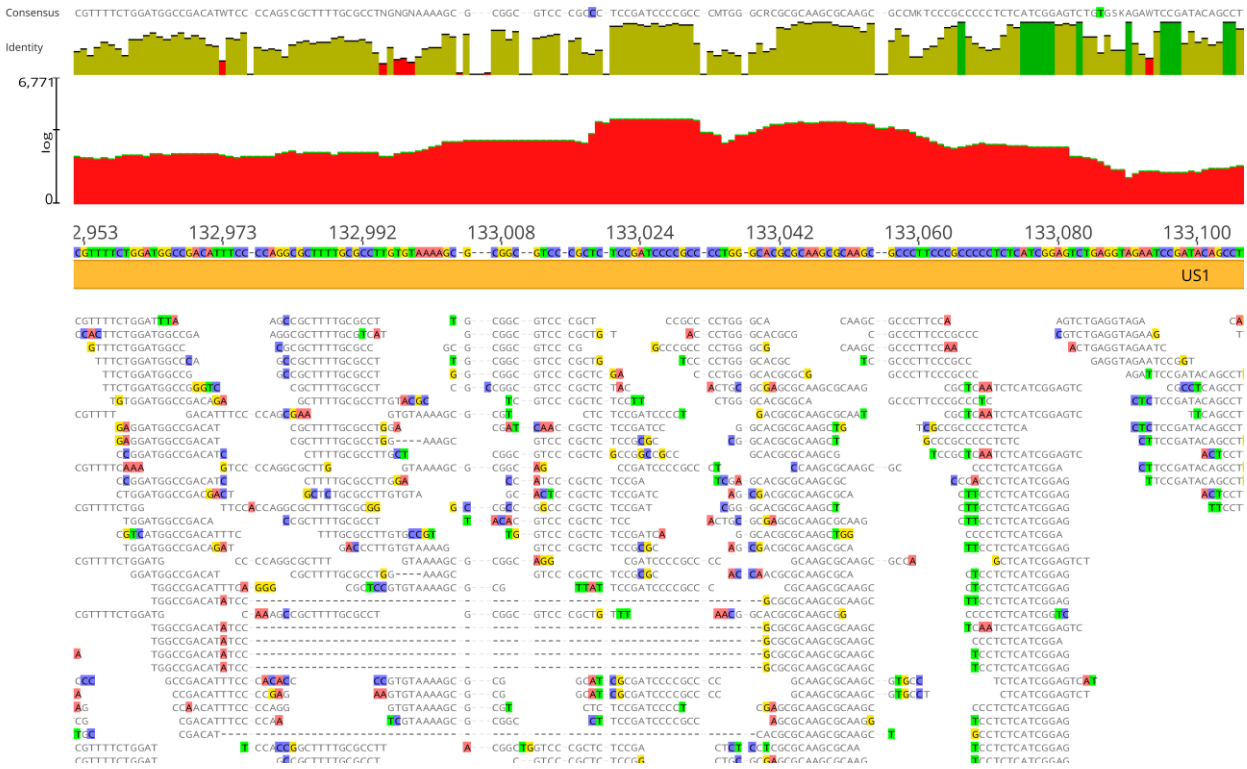

Figure 3. Detail of Neanderthal reads (Chagyrskaya 7) mapped to herpesvirus assembly reference genome sequence (GenBank MN136523). Red area indicates coverage Dark green vertical bars indicate 100% similarity, olive green vertical bars indicate nucleotide identity <100% and ≥30%, red vertical bars indicate <30% identity. Colored residues are those that differ from the assembly reference. The consensus (50% strict rule) is above nucleotide identity bars. The mapping was performed with three reference sequences simultaneously, as detailed in Material and Methods, but the depiction is with the herpesvirus assembly reference sequence for clarity.

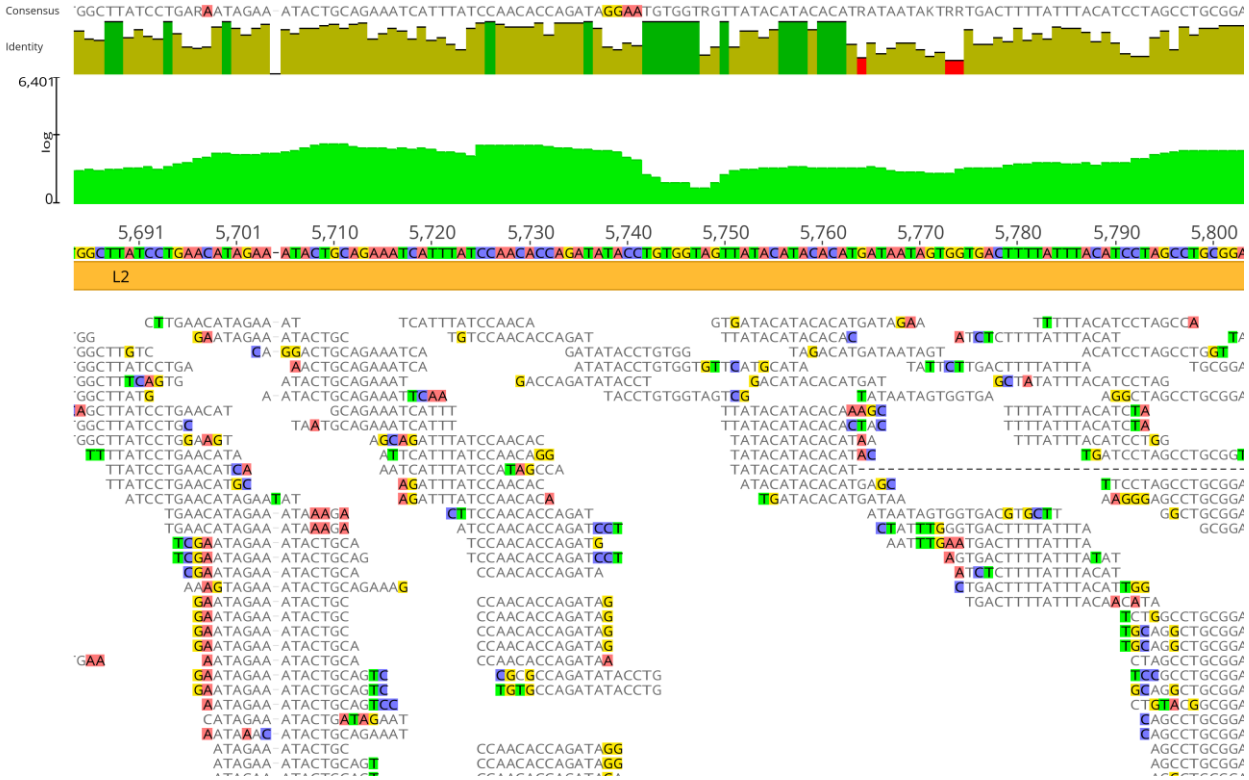

Figure 4. Detail of Neanderthal reads (Chagyrskaya 7) mapped to papillomavirus assembly reference genome sequence (GenBank X74466). Light green area indicates coverage and Dark green vertical bars indicate 100% similarity, olive green vertical bars indicate nucleotide identity <100% and ≥30%, red vertical bars indicate <30% identity. Colored residues are those that differ from assembly reference. The consensus (50% strict rule) is above nucleotide identity bars. The mapping was performed with three reference sequences simultaneously, as detailed in Material and Methods, but the depiction is with the papillomavirus assembly reference sequence for clarity.

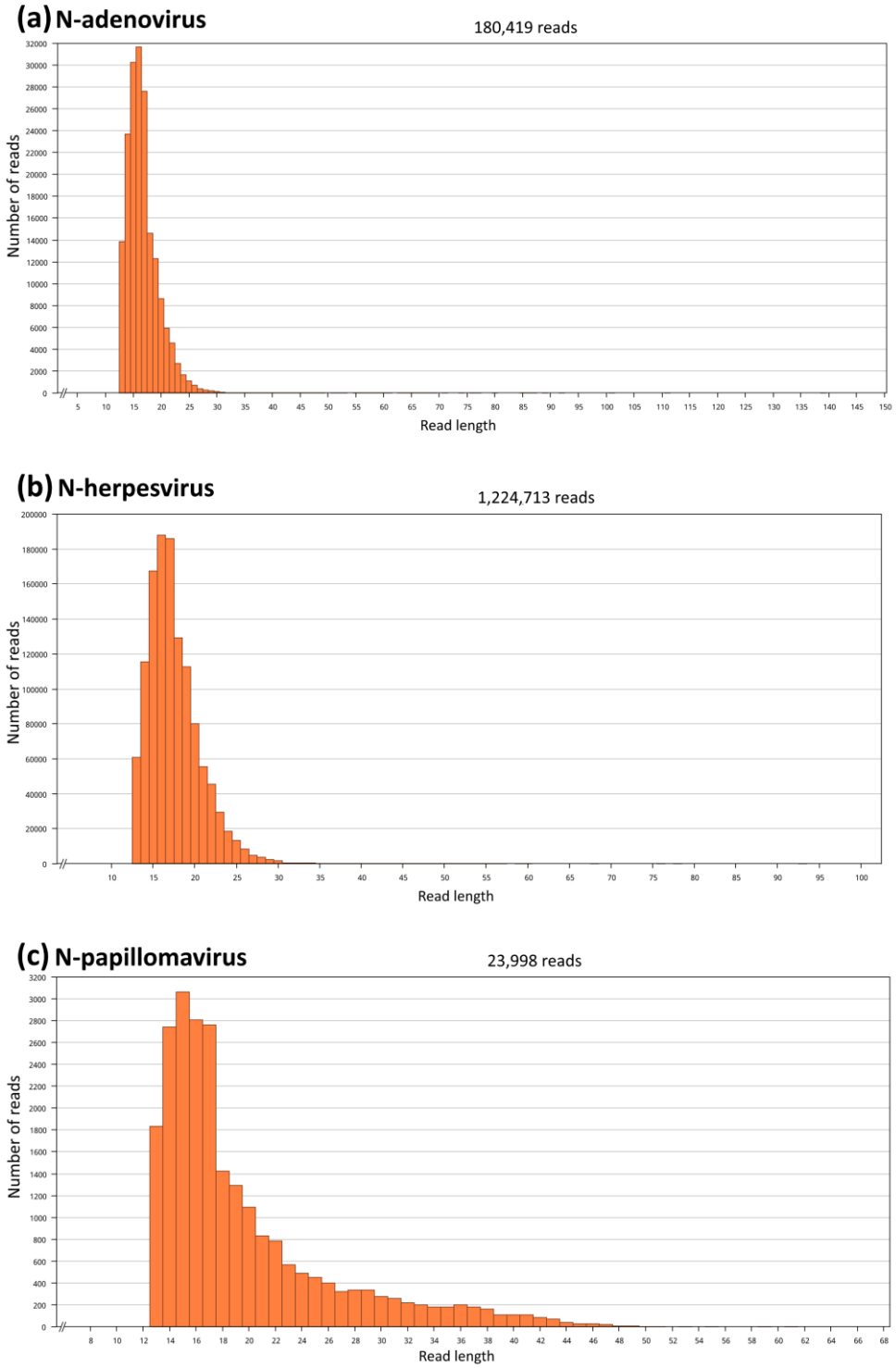

Figure 5. Neanderthal read length distributions. Read lengths used for mapping of (a) N-adenovirus (from 13 to 139 bases), (b) N-herpesvirus (from 13 to 93 bases) and (c) N-papillomavirus (from 13 to 61 bases).

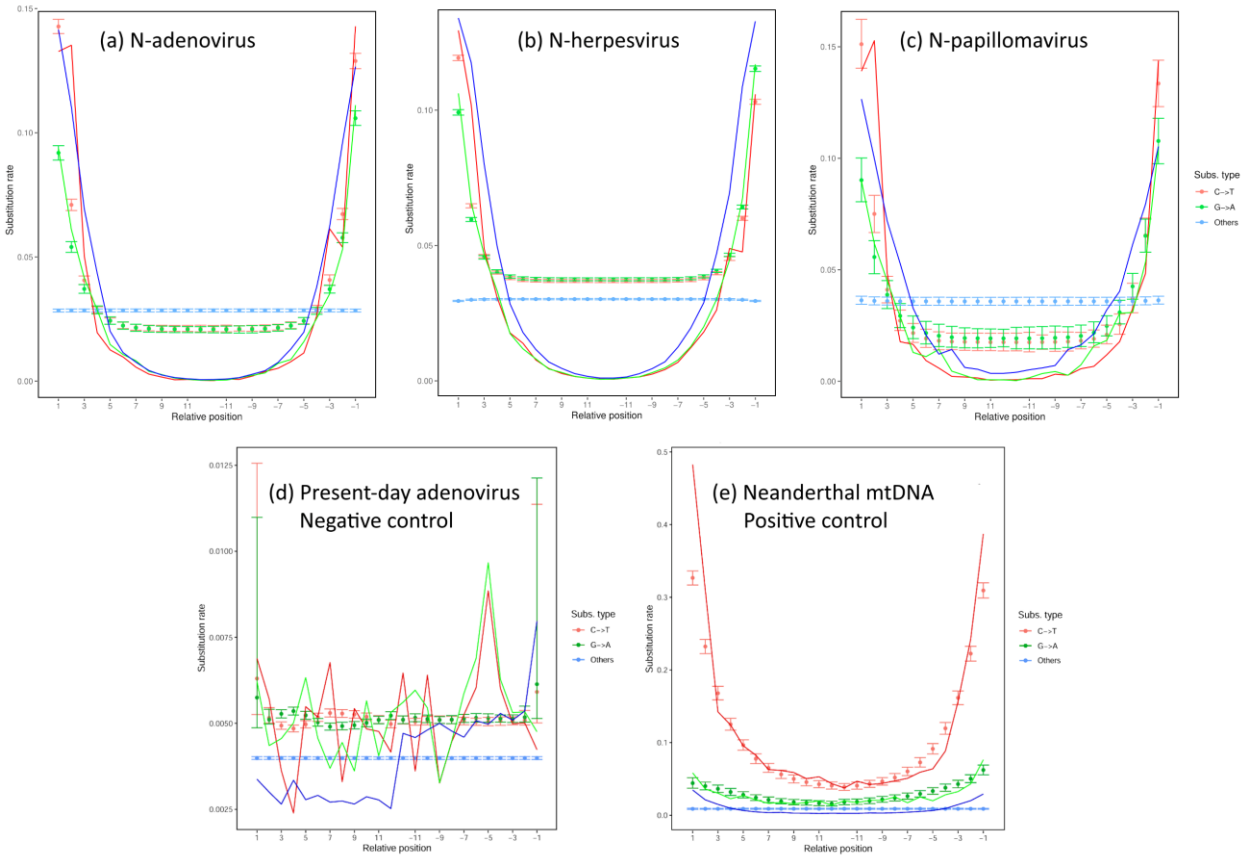

Figure 6. Substitution types and their relative positions in reads used for mapping of the N-adenovirus assembly (a), the N-herpesvirus assembly (b) and the N-papillomavirus assembly (c). The BAM files of theassemblies were analyzed by mapDamage program [26]. Discrete points per relative position indicate the substitution rate per read position of Chagyrskaya 7 and lines indicate the likelihood curves. In all three assemblies the C-to-T deamination occurs at the terminal overhangs as expected in ancient DNA reads. G-to-A changes are produced in ancient DNA by deamination of G to xanthine (X) which is read by sequencing methods as A. The number of reads analyzed for N-adenovirus is 180,419, for herpesvirus, 1,224,413 reads and for papillomavirus, 23,998 reads. The abscissa indicates the substitution type relative to the termini of reads, being 1 the 5′ termini and up to 11 positions inwards while the negative numbers provide the opposite end. A negative control (d) consists of sequencing reads from a present-day sample of adenovirus (SRA run SRR26630751) mapped to human adenovirus 7 reference (AC_000018.1) (6,205,335 mapped reads). The positive control (e) consists of Neanderthal reads of Chagyrskaya 7 mapped to the human mitochondrial genome reference sequence rCRS (NC_012920.1) (29,500 mapped reads). For negative and positive controls, the BBMap (mapping) and BBDuk (trimming) parameters used were identical with N-adenovirus, N-herpesvirus and N-papillomavirus.

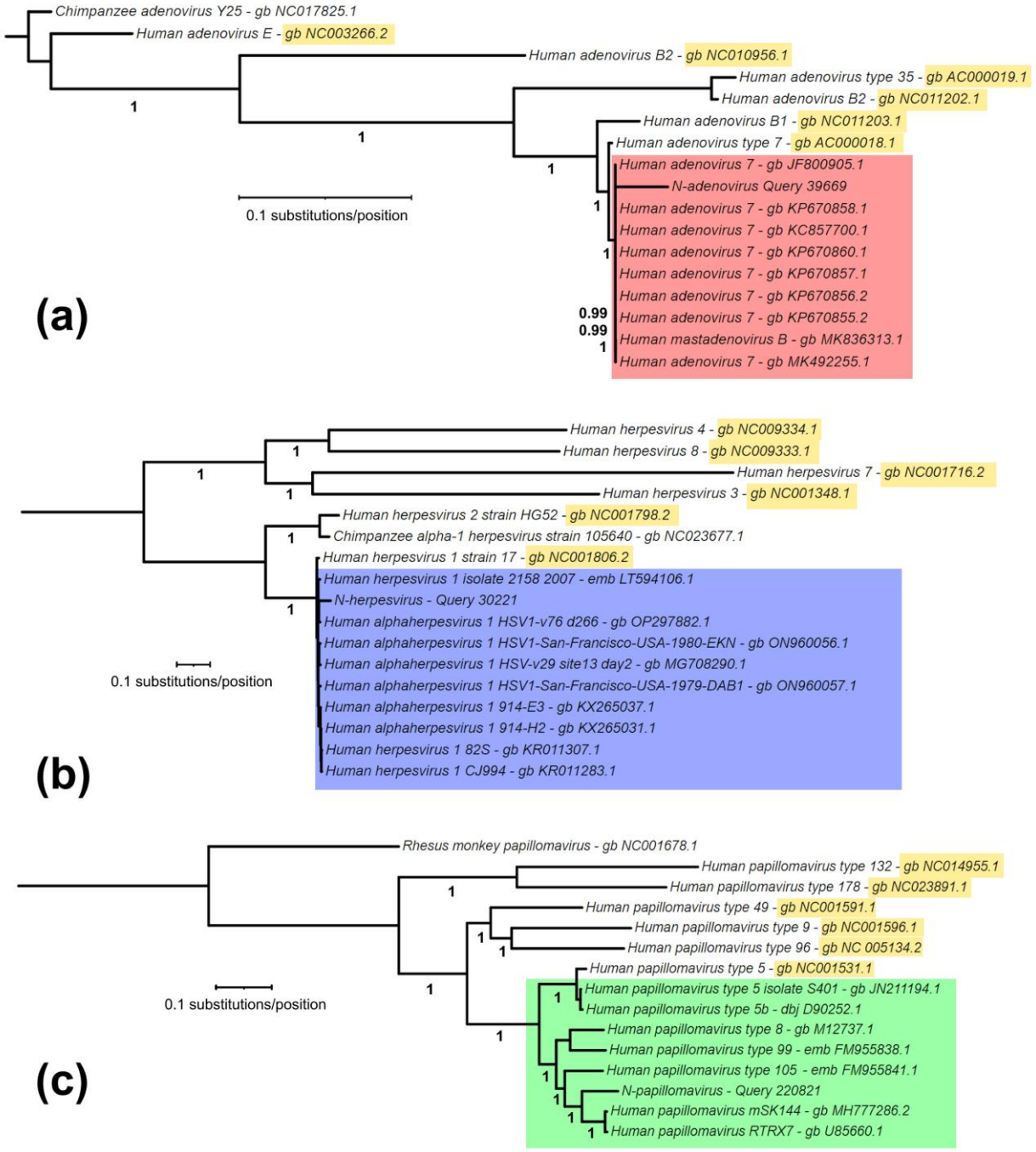

Figure 7. Bayesian trees of reconstructed ancient viral genomes and extant viral relatives. The genealogies of viral sequences were inferred from Chagyrskaya 7 reads of N-adenovirus, its closest relatives and RefSeq (a), N-herpesvirus its closest relatives and RefSeq(b) and N-papillomavirus its closest relatives and RefSeq (c). Numbers within boxes are the posterior probabilities. The scale bars indicate substitutions per position. The likelihoods of trees are (a) LnL= -149,860.813, (b) LnL= -1,187,183.249, and (c) LnL= -76,429.53635,496.56. Terminal taxa are identified with accession numbers. Whole genomes were aligned by MAFFT and used as input for Bayesian tree inference with MrBayes program version 3.2.7. Red, blue and green boxes indicate the subtrees with sequences of the closest extant relatives of N-adenovirus, N-herpesvirus and N-papillomavirus, respectively, as identified by MegaBlast analysis. and yellow boxes

indicate the NCBI RefSeq sequences for the corresponding viral taxa. The numbers below branches indicate the measure of branch support (posterior probabilities).

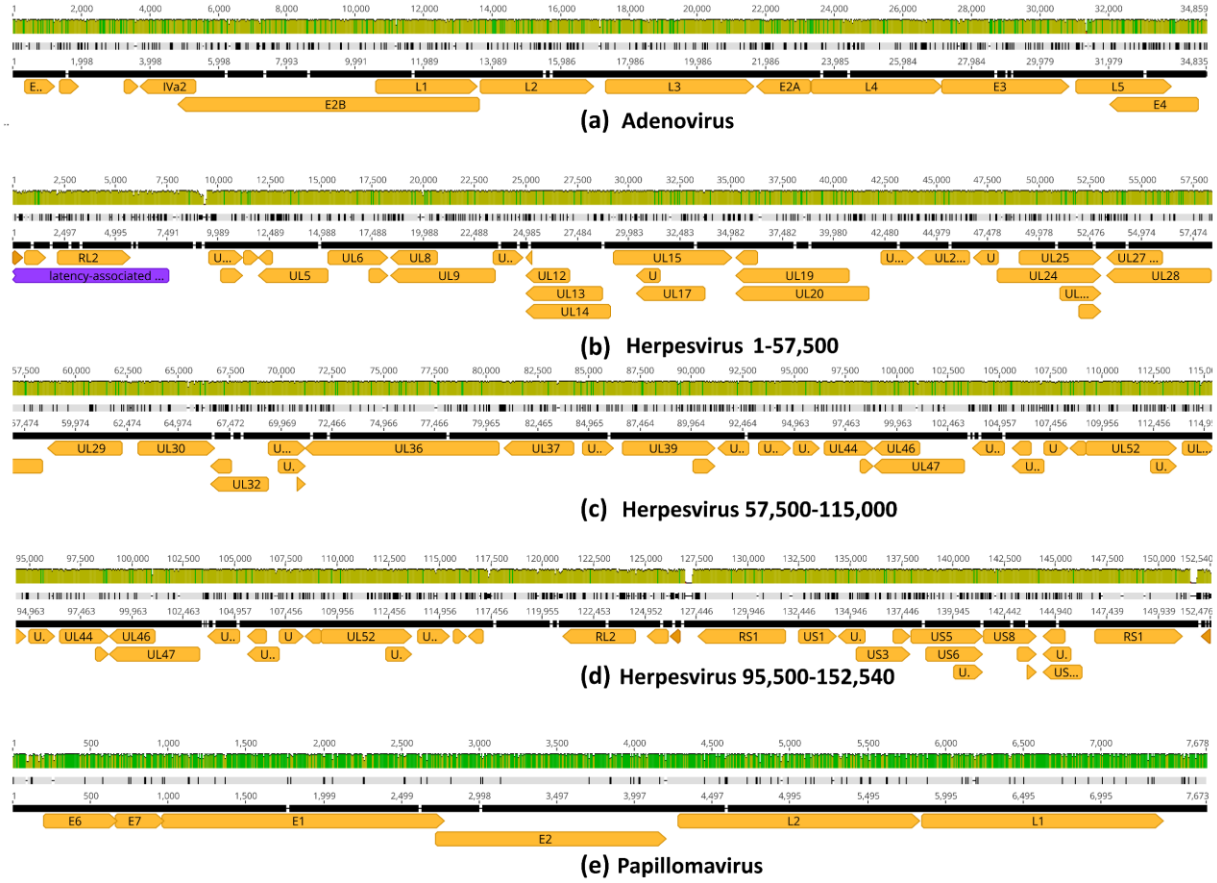

Figure 8. Alignment of reconstructed Neanderthal viral genomes to extant closest relatives. In (a) N-adenovirus vs adenovirus (94.4% global nucleotide identity). In (b-d) N-herpesvirus vs alphaherpesvirus 1 (94.1% nucleotide identity). In (e) N-papillomavirus vs papillomavirus (94.5% nucleotide identity). Pairwise alignments were performed using Geneious aligner. Olive green bars indicate partial nucleotide identity while dark green 100% nucleotide identity. Genes are indicated by orange arrows. Solid black lines are reference sequences of extant genomes. Black tick marks in grey lines indicate sequence differences of Neanderthal viruses relative to reference sequences.

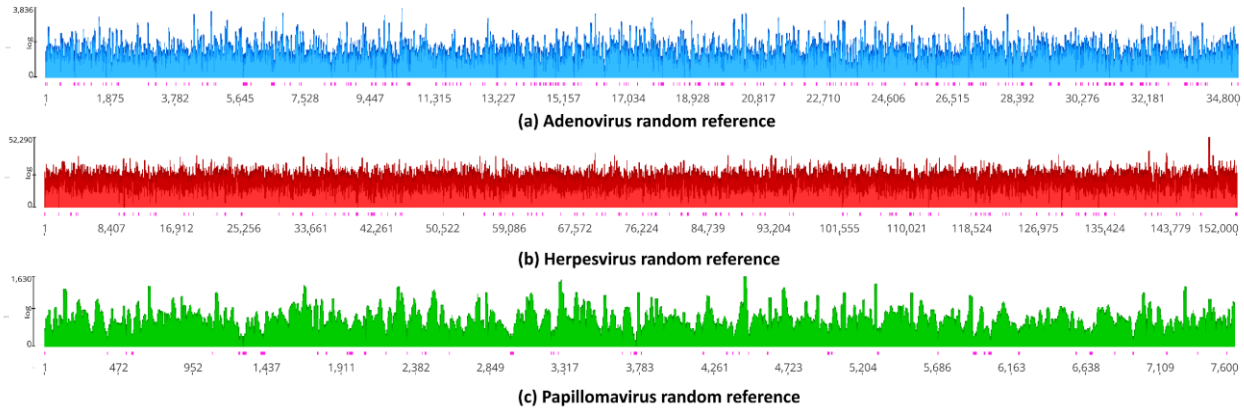

Figure 9. Assembly by mapping of Neanderthal reads against random reference sequences. Random reference sequences of adenovirus, herpesvirus and papillomavirus were obtained by UGENE version 44 [29]. Random sequences have the same size and base frequencies as actual reference sequences. Random reference sequences were used for genome assembly by mapping using the same programs and parameters as used for actual reference sequences. Pink vertical bars below coverage indicate single strand regions.

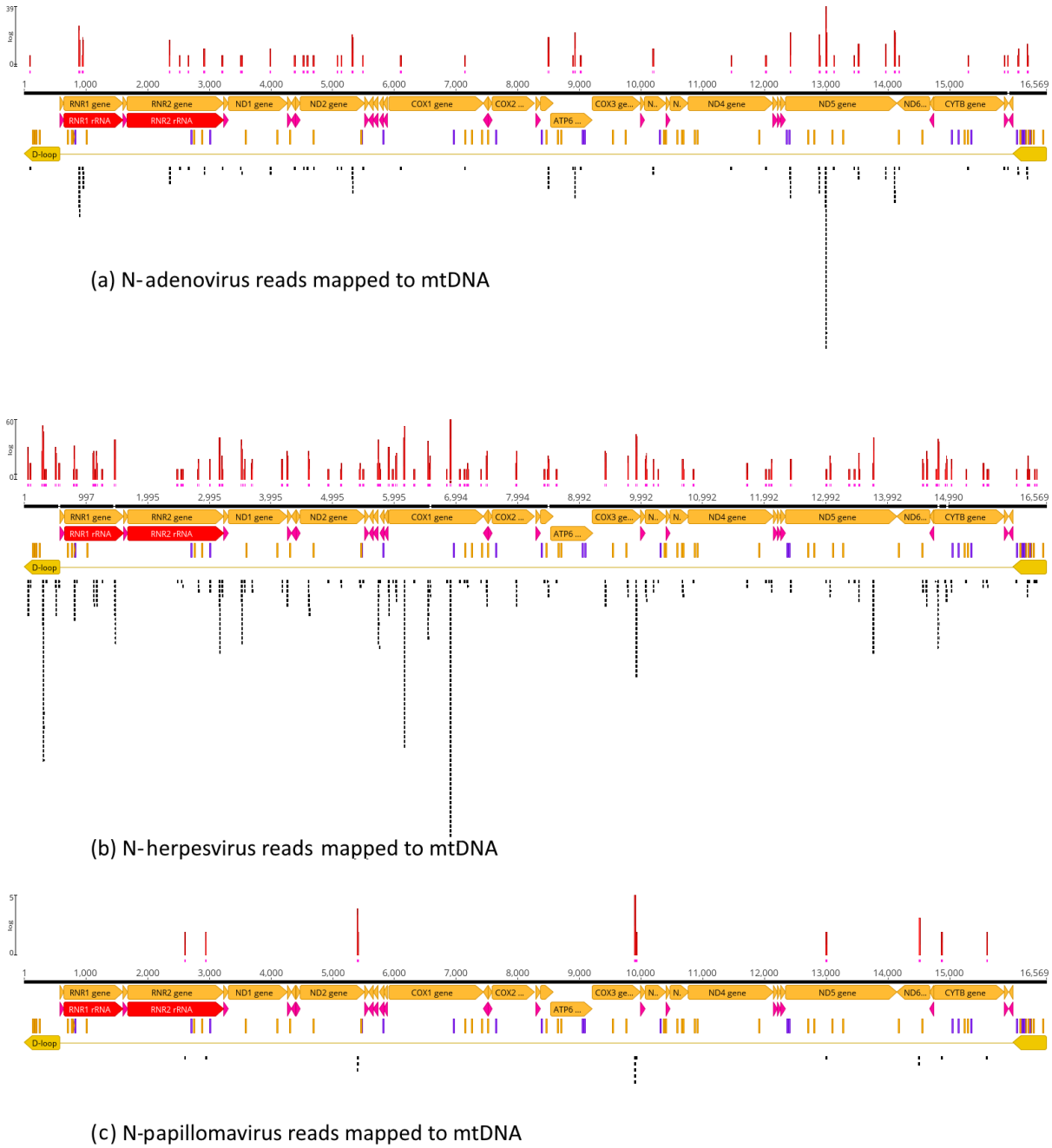

Figure 10. Mapping of N-adenovirus, N-herpesvirus e N-papillomavirus reads against the human genome (hg38) reference sequence. Reads that mapped to references in Figure 1 were "unmapped" and used for mapping the twenty-five chromosomes of the human genome (22 autosomes, X, Y and mtDNA). Genome assembly by mapping used the same programs and parameters as used for actual reference sequences. Pink vertical bars below coverage indicate single strand regions.

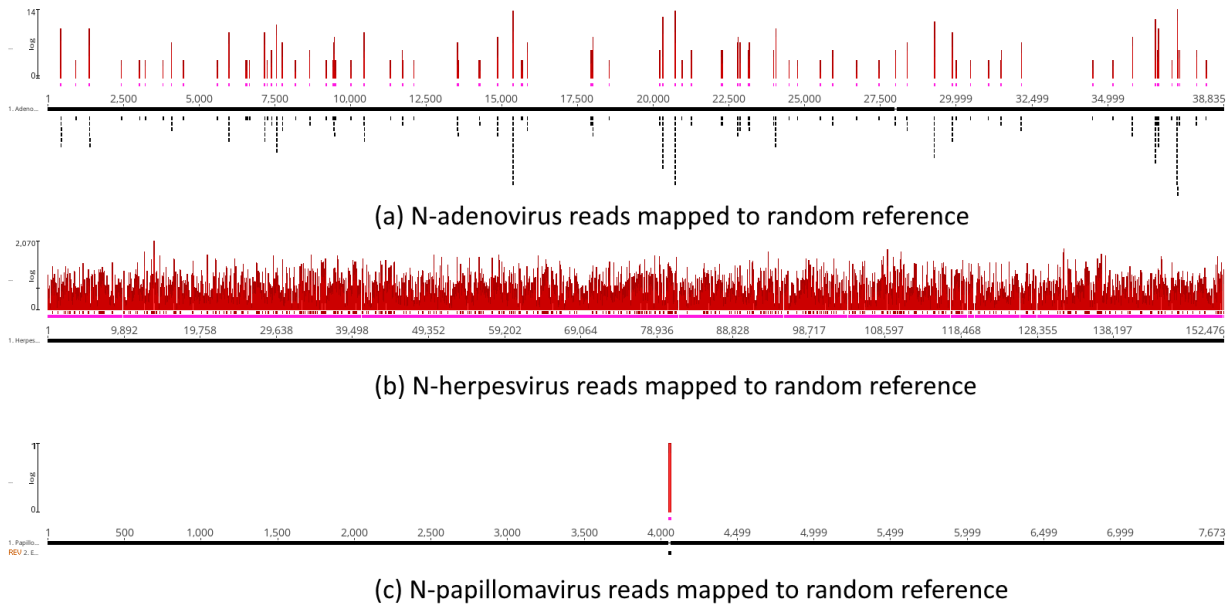

Figure 11. Mapping of N-adenovirus, N-herpesvirus e N-papillomavirus reads against equivalent random reference sequences. Reads that mapped to references in Figure 1 were "unmapped" and used for mapping random reference sequences. Random sequences have the same size and base frequencies as actual reference sequences. Random reference sequences were used for genome assembly by mapping using the same programs and parameters as used for actual reference sequences. Pink lines below coverage indicate single strand regions.

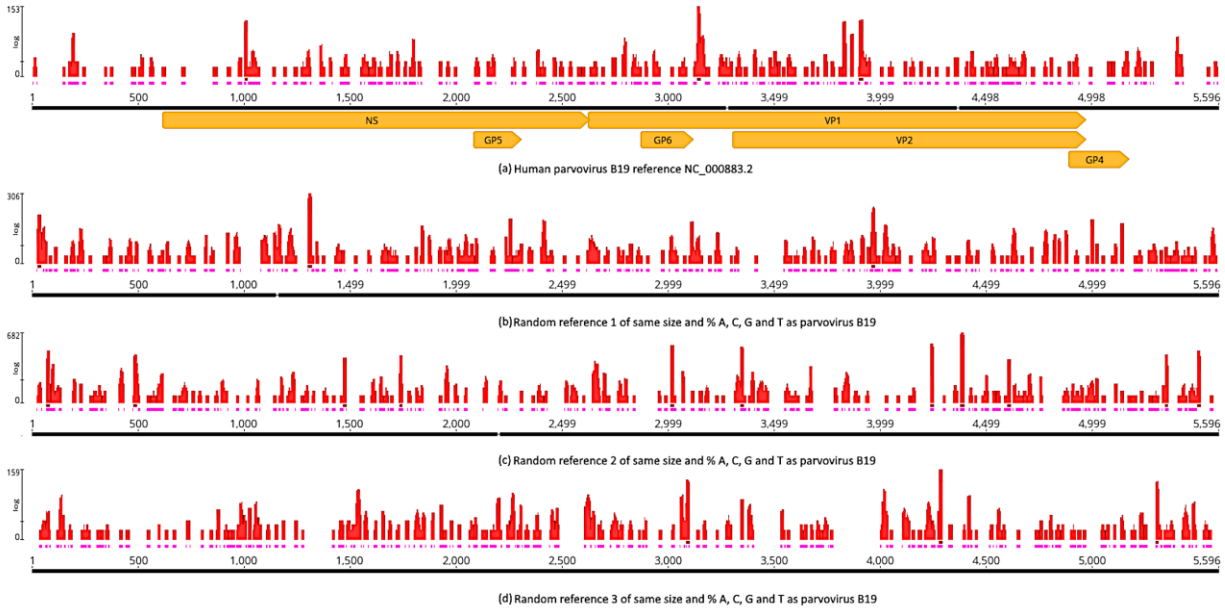

Figure 12. Mapping of Neanderthal reads to human parvovirus B19, a DNA virus that produces nonpersistent infection. In (a) coverage and reads distribution along the parvovirus reference sequence while in (b), (c) and (d) is the coverage and reads distribution along the random reference sequences of same size and base frequencies as the parvovirus B19 reference sequence NC_000883.2 (NCBI RefSeq). Red bars above reference sequences indicate coverage, pink bars just below red bars indicate single strand and orange arrows indicate the parvovirus genes. Mapping was obtained with BBMap using the same parameters used for adenovirus, herpesvirus and papillomavirus here analyzed.

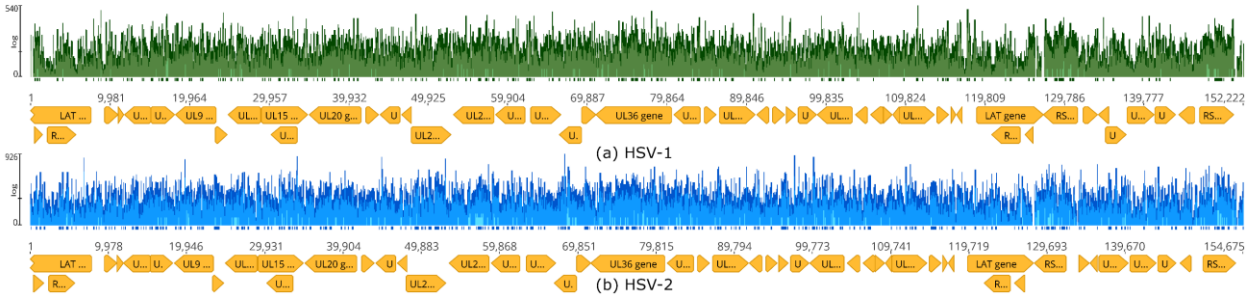

Figure 13. Mapping of Neanderthal reads to HSV-1 and HSV-2 reference sequences. The total set of 1.2 billion reads of Neanderthal Chagyrskaya 7 was assembled simultaneously to HSV-1 NC_001806.2 (a) and HSV-2 NC_001798.2 (b) reference sequences using BBMap. Green vertical bars indicate HSV-1 coverage, blue vertical bars indicate HSV-2 coverage. Orange arrows indicate genes. Tick marks below coverage bars indicate coverage above 50x. A total of 1,221,442,735 Neanderthal Chagyrskaya 7 reads were assembled to reference sequences to produce an HSV-1 contig with 74,405 reads and HSV-2 contig with 105,046 reads. A total of 1,221,263,284 reads were not assembled.