Cold Spring Harbor Laboratory

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Reconstructing prehistoric viral genomes from Neanderthal sequencing data

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

Affiliations:
[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

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

*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. Here we searched for DNA viral genomes in sequencing reads of Neanderthal genome projects by mapping to references of Adenovirus, Herpesvirus and Papillomavirus, which are double stranded DNA viruses that establish lifelong latency. Reconstructed ancient viral genomes exhibited conserved segments with significant similarity to extant viral genomes. Variable positions were identified in coding regions and revealed substantial divergence to any extant closest relatives. Sequencing reads mapped to references showed typical deamination patterns of ancient DNA, suggesting that contamination with modern viral DNA is unlikely. These ancient viral genomes showed divergence corresponding with the age of these samples (≈50,000 years) and the viral evolutionary rates (10^{-5} to 10^{-8} substitutions/site/year). Variable regions reveal Neanderthal adaptive signatures, such as nonsynonymous substitutions in viral surface proteins that interact with receptors on host cells. Conserved regions here identified can be used as PCR targets for amplification of viral genomes from Neanderthal isolated DNA, in future experiments, that might close low coverage gaps in the assemblies for accurate reconstruction of these viral paleogenomes. Mapping Neanderthal and signatures of anatomically modern humans (AMH) in these viruses is essential to establish whether Neanderthals might have acquired these infections from invading modern human populations, as proposed to explain the extinction of Neanderthal upon contact with AMH.

1. Introduction

Several hypotheses were proposed to explain the extinction of the Neanderthals. 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 is a major drive in human evolution [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].

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 long 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 various 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 JC virus and BK virus, 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 has been successfully recovered as ancient European HSV-1 genomes in samples dating from the 3rd to 17th century CE, with up to 9.5× with paired human genomes up to 10.16x [8]. Therefore, Herpesvirus must be common in the archaeological record.

In this study, we searched for viral genomes in raw sequencing data of *Homo sapiens neanderthalensis* from Chagyrskaya cave in Russia (samples 1 and 7) [22]. An initial taxonomic analysis using the Sequence Taxonomic Analysis Tool (STAT) on the NCBI Sequence Read Archive (SRA) [23], identified reads corresponding to several microorganisms, such as bacteria, fungi and viruses. Inspection of results from this analysis allowed the selection of genomes of human pathogenic DNA viruses to be used as references for genome assemblies.

2. Materials and Methods

2.1. Samples analyzed

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). Reference sequences used were: Human adenovirus 7 (KX897164), Human alphaherpesvirus 1 (MN136523) and Human papillomavirus type 12 (X74466).

2.2. Genome assembly by reference mapping

For genome assembly by mapping, duplicates 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 reference 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 reads with 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 was 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.

2.4. *Viral Genealogies*

The consensus sequences obtained from genome assembly by reference mapping were exported and used for blast (MegaBlast) analyses to identify 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].

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 [29]. Random mock sequences were of the same size as actual reference sequences and have the same base frequencies. 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.

3. Results

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) [30]. Chagyrskaya 1, comprises 11 runs with 2.96×10^9 bases (3.25 Gb of data).

Because of higher coverage of Neanderthal genome Chagyrskaya 7 (4.9 fold coverage) [30] it was used for reference mapping using BBMap, an assembler optimized for short reads. Before processing, the data comprised 128 runs with 152.71×10^9 bases (170.21 Gb of data). After removal of duplicates, 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 reference 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 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 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 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).

Ancient DNA origin of reads mapped to references was verified by analysis of deamination (Figure 6). In Figure 6 A and C, the assembly of Adenovirus and Papillomavirus shows that C-T deamination is significantly more abundant in the 5' overhangs as compared to the other changes away from termini (Fig. 1F, left and right panels). In Figure 6 B. 1F, in the Herpesvirus assembly, although C-T deamination is more prevalent in the termini the C-T is more abundant in central regions than other changes. The likelihood provides statistical support for C-T deamination in all three assemblies, which is a main feature of reads derived from ancient DNA as opposed to recent DNA [26].

The genetic relatedness of the Neanderthal N-Adenovirus, N-Herpesvirus and N-Papillomavirus with extant viruses was inferred by MegaBlast. Each consensus was used as query and their closest extant relatives aligned and used for the inference of genealogies by the Neighbor Joining method (Figure 7). In Figure 7 A, 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). In Figure 7 B, 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). In Figure 7 C, bottom, the Neanderthal Papillomavirus clusters with type 12 sequences although its branch length indicates its distant relationship to others in the same cluster at 0.06 substitutions/site, or 494 substitutions along the branch (72 closest relatives).

Alignments of the inferred N-Adenovirus, N-Herpesvirus and N-Papillomavirus to the extant closest relatives revealed the sequence changes distributed along the majority of coding regions including those encoding relevant infection proteins (Figure 8). In Adenovirus changes are observed in genes encoding Fiber, Hexon and Penton base as well as core proteins (Supplementary Figure S1 and Supplementary Tables 1) [31]. In Herpesvirus changes are observed in glycoproteins involved in interaction with host cell surface such as gB, gB, gH and gL (Supplementary Figure S2 and Supplementary Table 2) [32]. In Papillomavirus changes were observed in capsid protein L1 as well as oncoproteins E6 and E7 (Supplementary Figure S3 and Supplementary Table 3) [33].

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. 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.

4. Discussion

Further detailed characterization of variable regions of these ancient viral genomes might reveal Neanderthal adaptive signatures. Here 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, gD, gH and gL involved in cell adhesion and infectivity of Herpesviruses (Supplementary Figure S1) [34]. Changes in major surface viral proteins of Adenovirus and oncoproteins of Papillomavirus have been revealed as well (Figure 8, Supplementary Figures S2 and S3). 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 amplification 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.

Analysis of deamination in the terminal ends of reads assembled to reference sequences show that the expected pattern of ancient DNA 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 significantly distant from any extant viral genomes of the respective groups (Figure 7).

In the case of Chagyrskaya 7, as 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. Herpes simplex virus-1 (HSV-1) establishes latency preferentially in sensory neurons of peripheral ganglia. These viruses are the cause of oral and genital herpes. Human adenoviruses persistence and reactivation occurs in the gastrointestinal tract and the best-documented sites of latency include tonsillar and adenoidal T lymphocytes [35]. 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 [36]. 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.

The possibility that simple random similarities between Neanderthal reads and viral reference sequences could explain the mapping results obtained in Figure 1, random mock reference sequences of Adenovirus, Herpesvirus and Papillomavirus were generated. Size and base composition of random references were identical with actual counterparts. Mapping random references would reflect the random noise in the analysis. As shown, the assemblies obtained with real sequences have more reads mapped and higher coverage than the random sequences for all three viruses. 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). Also, in random references assemblies 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, the *bona fide* signal obtained by mapping of Neanderthal reads and actual viral genomes is above the level expected by random noise.

5. Conclusions

In this study we describe observations of immediate impact that may hold potential to foster new prospects of research in paleovirology. We provide provocative new data on the open question regarding the impact of viruses in the extinction of an ancestral human population, which in our view is of broad interest and a long-standing question. We inferred the probable sequence of Adenovirus, Herpesvirus and Papillomavirus associated with this Neanderthal population. Due to the nature and evolutionary age of these viruses they probably infected the individuals from which the DNA was isolated. The inference of the sequence of these paleoviruses revealed Neanderthal specific sequence changes that can be studied in greater depth, for each virus individually, to identify potential unique changes, especially in viral surface proteins. These might help to determine if these viruses were acquired by Neanderthals upon contact with AMH.

Supplementary Materials

The supporting figures and tables can be downloaded at BioRxiv.

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 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. Mafessoni, F.; Grote, S.; de Filippo, C.; Slon, V.; Kolobova, K.A.; Viola, B.; Markin, S.V.; Chintalapati, M.; Peyrégne, S.; Skov, L.; et al. A High-Coverage Neandertal Genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences* **2020**, *117*, 15132–15136, doi:10.1073/pnas.2004944117.
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. 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.
30. 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.
31. Kulanayake, S.; Tikoo, S.K. Adenovirus Core Proteins: Structure and Function. *Viruses* **2021**, *13*, 388, doi:10.3390/v13030388.
32. 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.
33. 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.

34. 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.
35. 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.
36. Gravitt, P.E. Evidence and Impact of Human Papillomavirus Latency. *Open Virol J* **2012**, *6*, 198–203, doi:10.2174/1874357901206010198.

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	Human Pathogenic 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

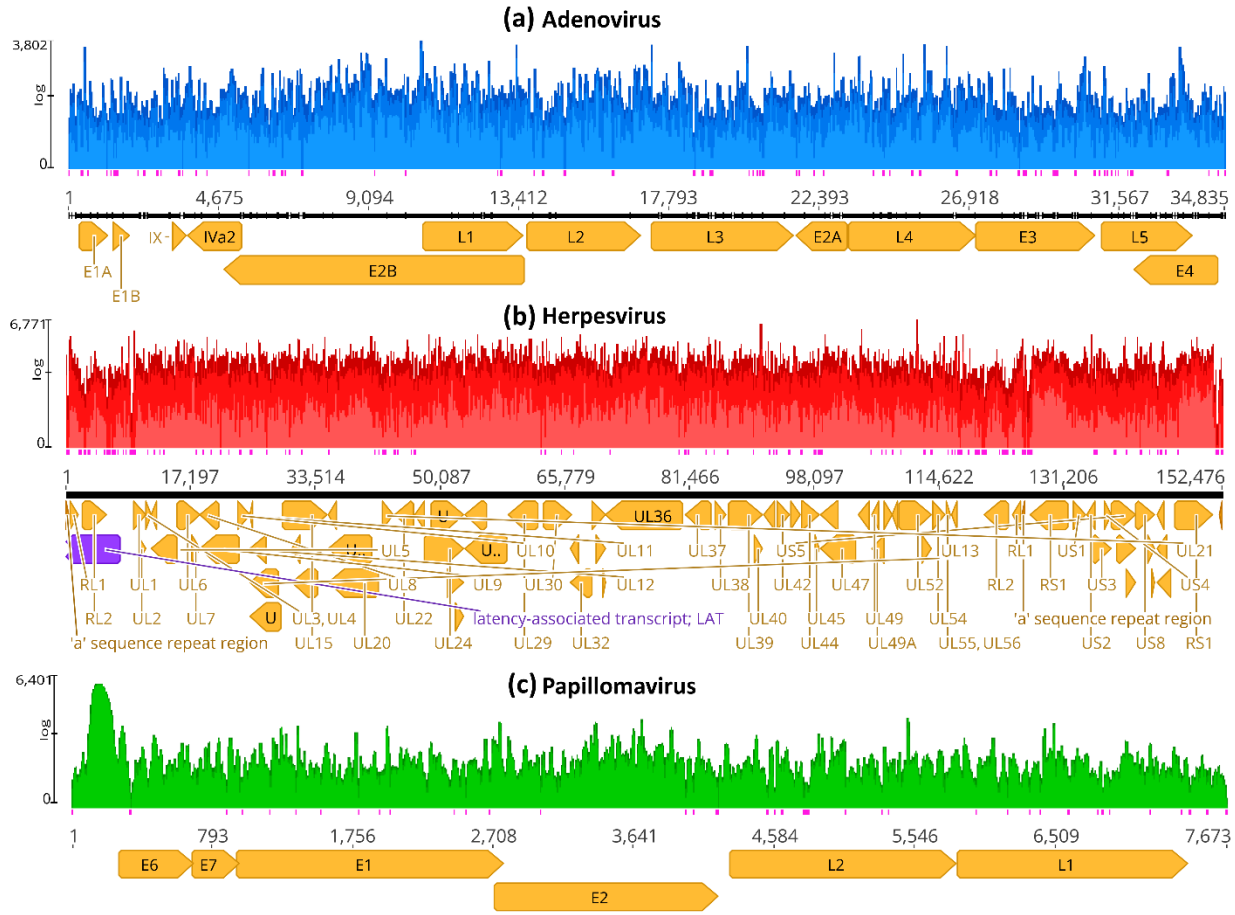

Figure 1. Mapping coverage of herpesvirus, papillomavirus and adenovirus using Neanderthal sample derived reads. Reference mapping was done using BBMap [24,25] (https://sourceforge.net/projects/bbmap/). In (a) is the mapping of 180,419 Neanderthal reads (mean coverage 102.2) to Adenovirus reference (KX897164). In (b), the mapping of 1,224,413 reads (mean coverage 171.2) to Herpesvirus reference (MN136523). In (c), the mapping of 23,998 Neanderthal reads (mean coverage 115.4) to Papillomavirus reference (X74466). Pink vertical bars below the assemblies indicate single strand regions and orange arrows indicate the protein coding genes.

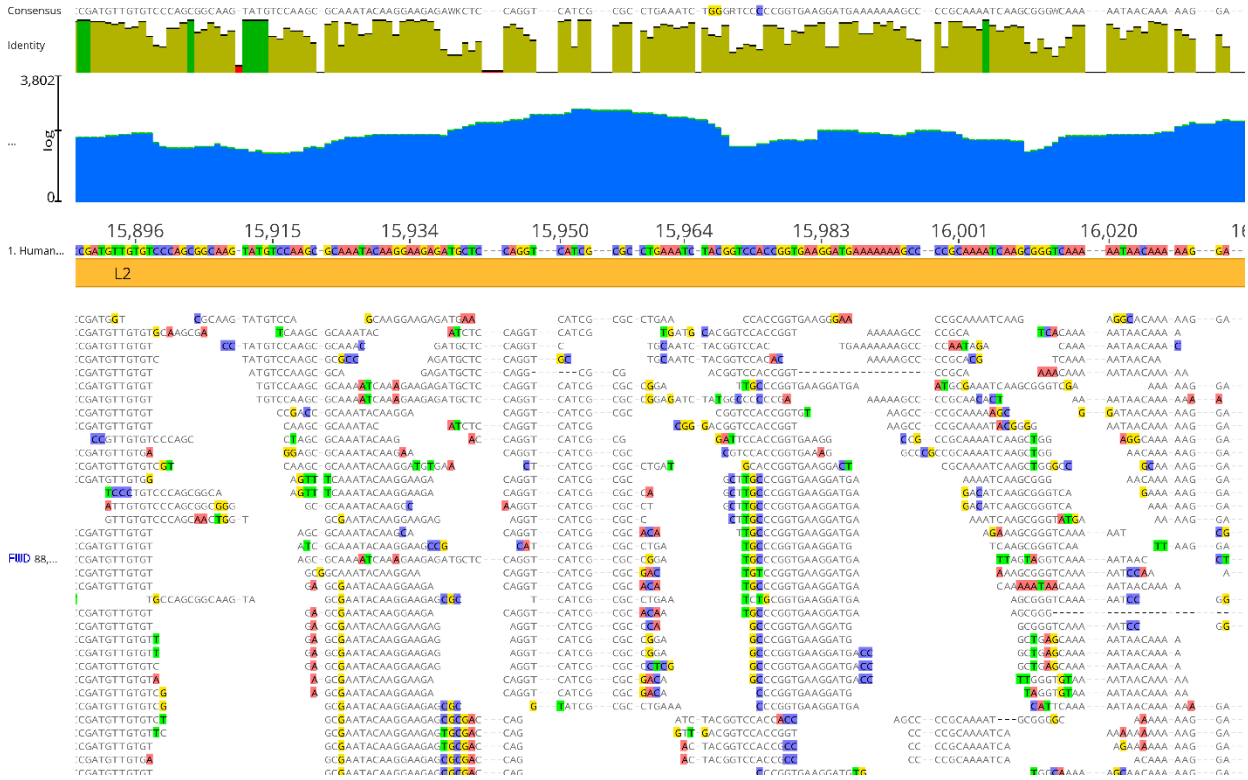

Figure 2. Detail of Neanderthal reads (Chagyrskaya 7) mapped to Adenovirus reference genome sequence (GenBank KX897164). Blue area indicates coverage and dark green, olive and red bars above coverage indicate identity. Colored residues are those that differ from reference. The consensus (50% strict rule) is above identity bars.

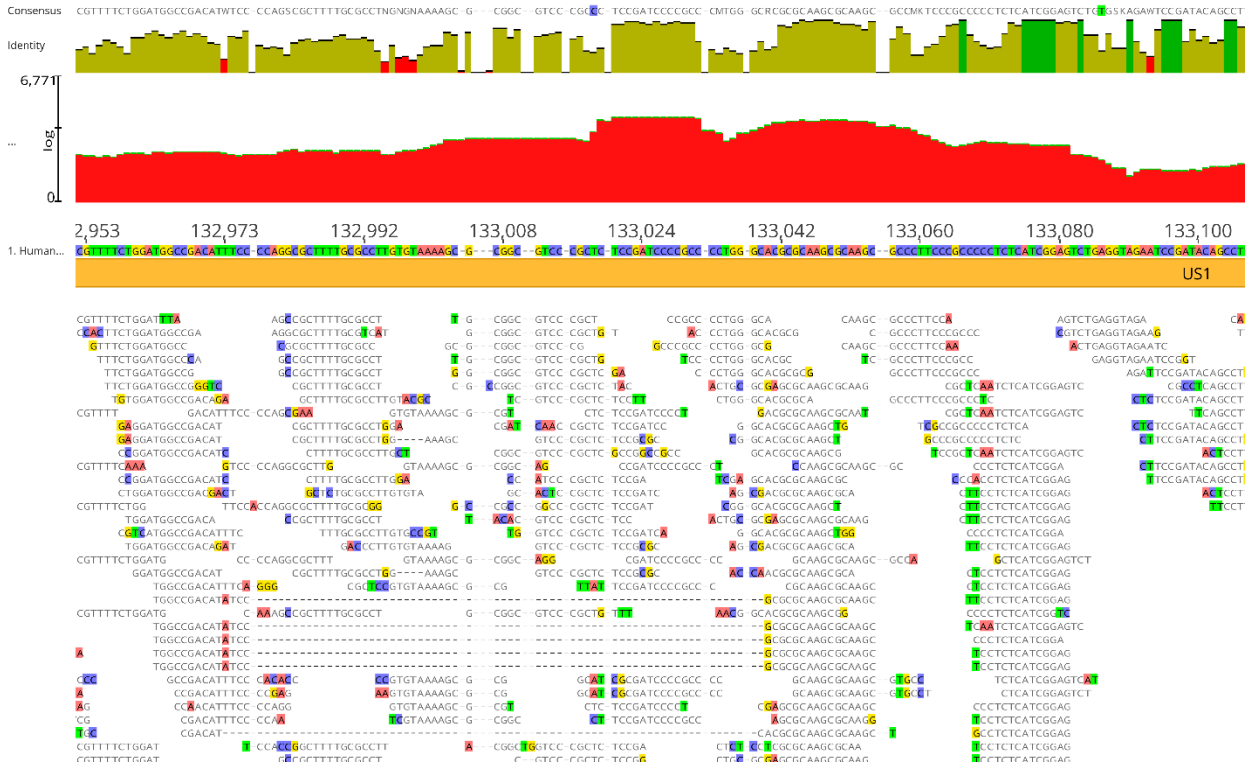

Figure 3. Detail of Neanderthal reads (Chagyrskaya 7) mapped to Herpesvirus reference genome sequence (GenBank MN136523). Red area indicates coverage and dark green, olive and red bars above coverage indicate identity. Colored residues are those that differ from reference. The consensus (50% strict rule) is above identity bars.

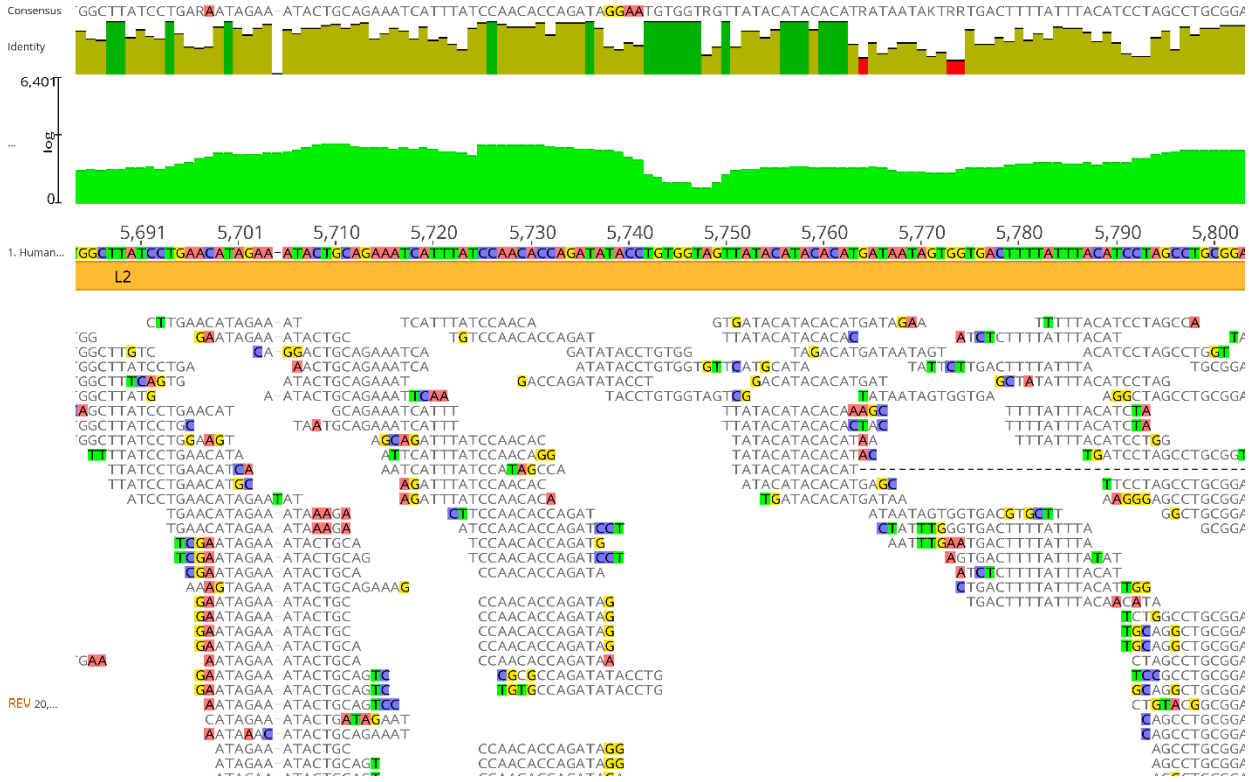

Figure 4. Detail of Neanderthal reads (Chagyrskaya 7) mapped to Papillomavirus reference genome sequence (GenBank X74466). Green area indicates coverage and dark green, olive and red bars above coverage indicate identity. Colored residues are those that differ from reference. The consensus (50% strict rule) is above identity bars.

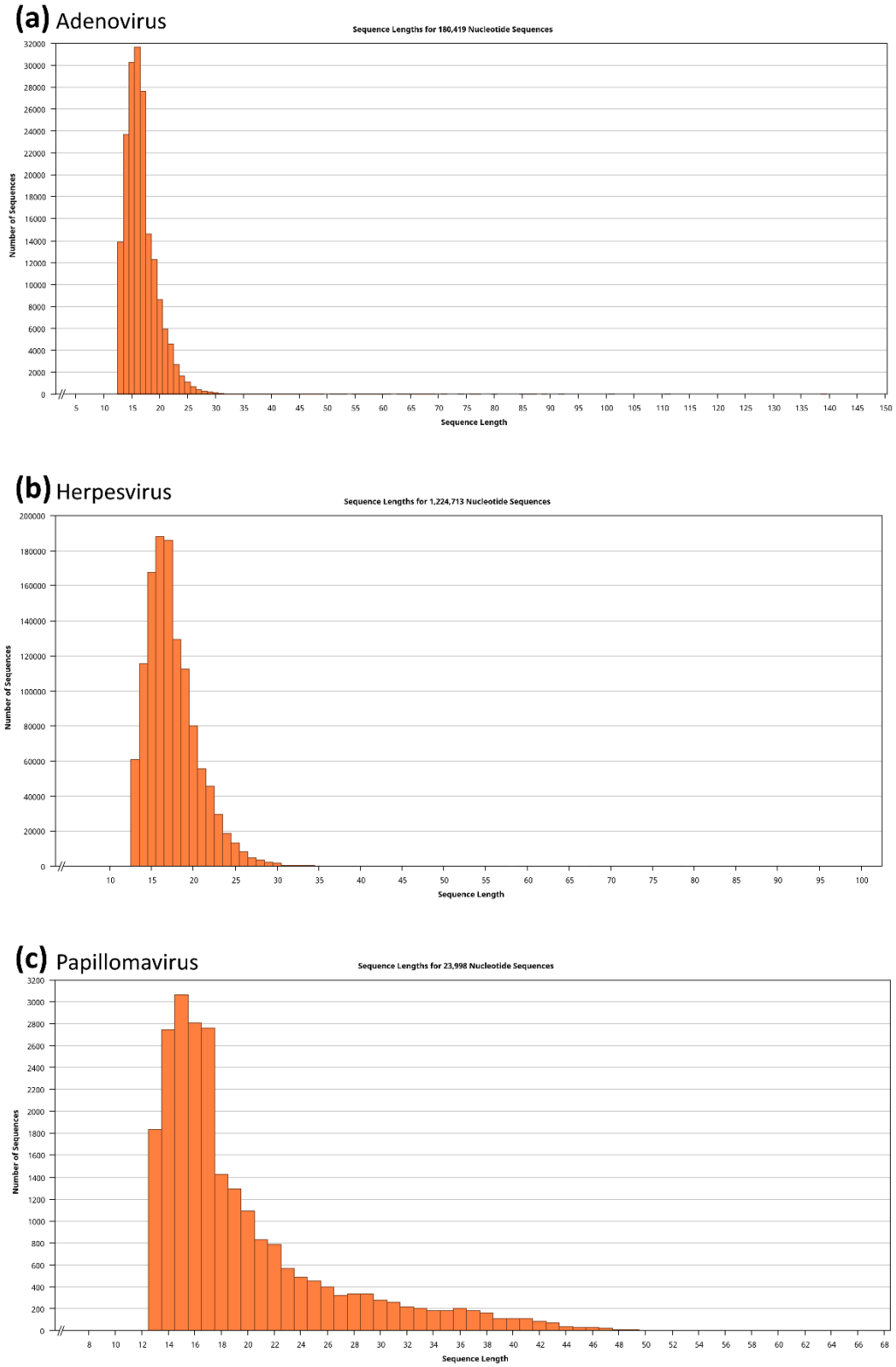

Figure 5. Neanderthal read length distributions. Read lengths used for mapping of (a) Adenovirus (from 13 to 139 bases), (b) Herpesvirus (from 13 to 93 bases) and (c) Papillomavirus (from 13 to 61 bases).

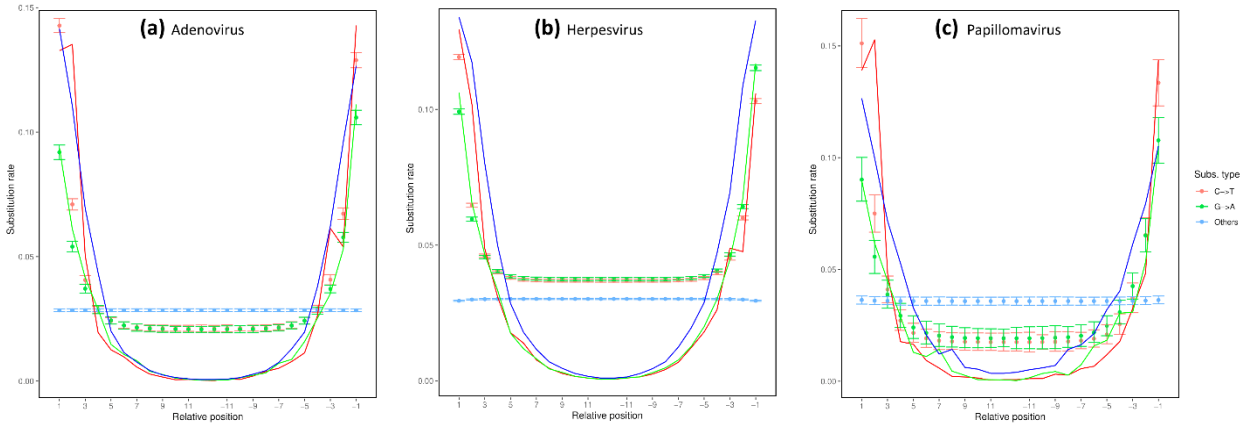

Figure 6. Ancient origin of reads used in assemblies the Adenovirus assembly (a), the Herpesvirus assembly (b) and the Papillomavirus assembly (c). A assemblies analyzed by mapDamage program [26]. Discrete points per relative position indicate the substitution rate per read position of Chagyrskaya 7 and lines indicate the likelihood curves. In all three assemblies the C-T deamination occurs at the terminal overhangs as expected in ancient DNA reads.

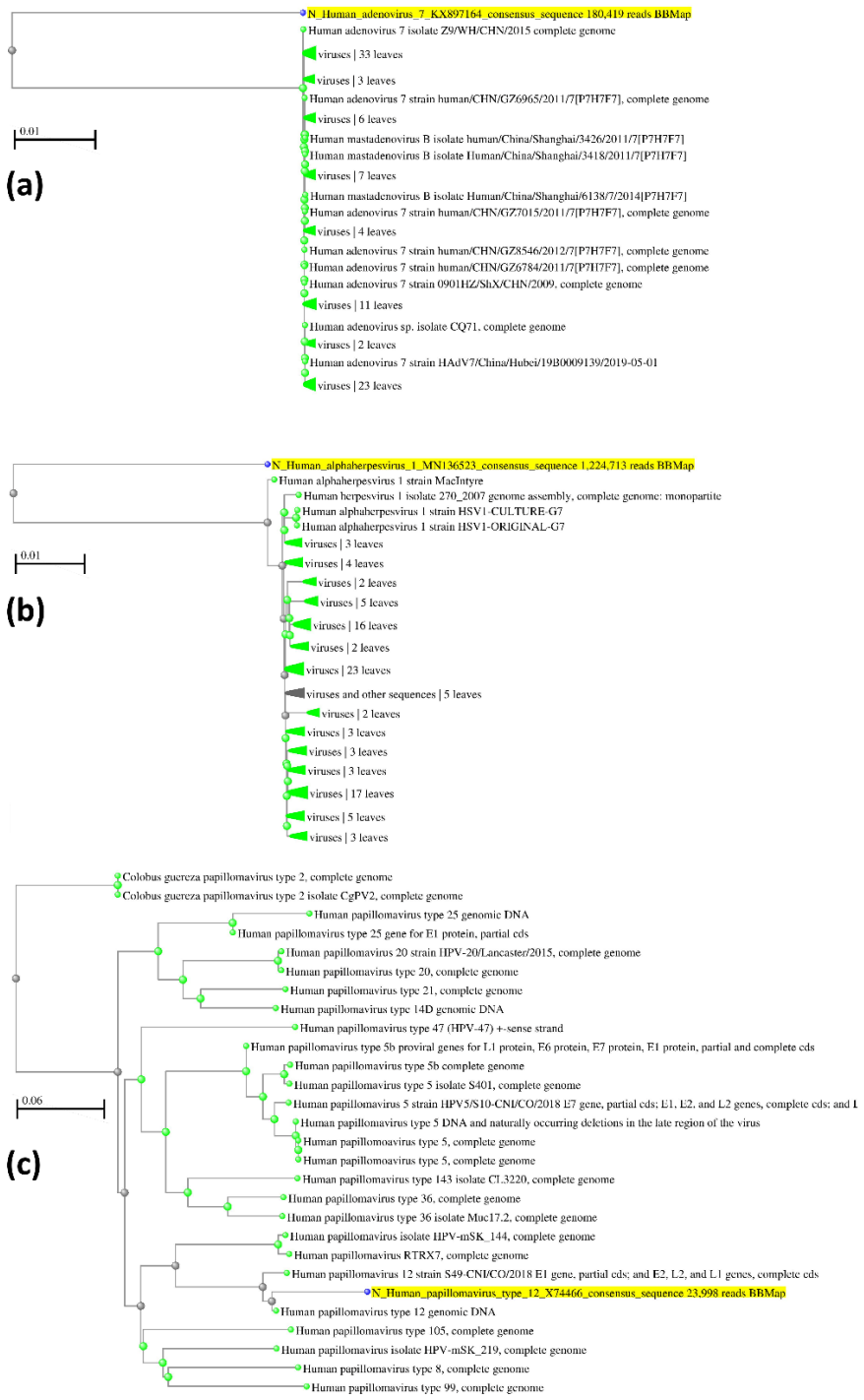

Figure 7. Neighbor-Joining trees of reconstructed ancient viral genomes and extant viral genomes. Yellow indicates the query (consensus of BBMap assemblies). The query sequences were aligned to the closest relatives in GenBank as searched using MegaBlast [27]. Bars in grey boxes indicate the scale of substitutions per position. The genealogies of viral sequences were inferred from Chagyrskaya 7 of N-Adenovirus and its 100 closest relatives (a), N-Herpesvirus and its 709 closest relatives (b) and N-Papillomavirus and its 72 closest relatives (c).

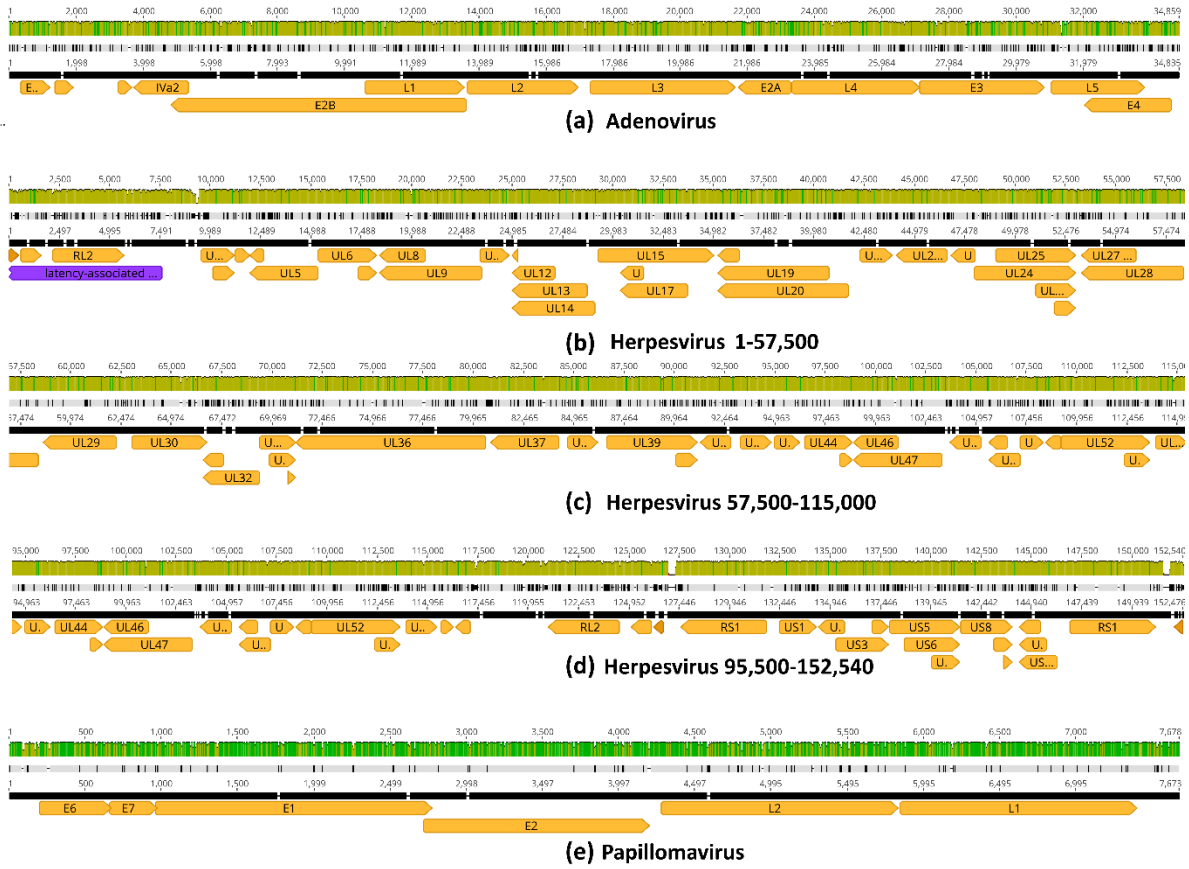

Figure 8. Alignment of reconstructed Neanderthal viral genomes to extant closest relatives. In (a) N-Adenovirus vs Adenovirus (94.4% global identity). In (b-d) N-Herpesvirus vs Alphaherpersvirus 1 (94.1% identity). In (e) N-Pappilomavirus vs Papillomavirus (94.5% identity). Pairwise alignments were performed using Geneious aligner. Olive green bars indicate 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.

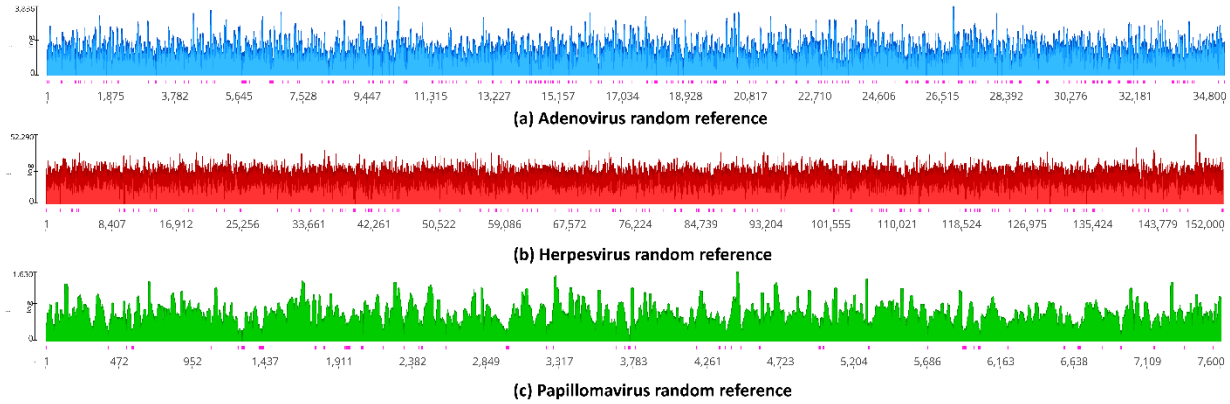

Figure 9. Assembly by mapping of Neanderthal reads against random reference sequences. Random reference sequences of Adenovirus, Herpesvirus and Papillomavirus were obtained by UGENE version 44 [29]. Random sequences have the same size and base frequencies as actual reference sequences. Random reference 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.