MSc Computer Science Thesis

In Silico Screening of Long-Read Sequencing Data for Endogenous Viral Elements in Alzheimer's Disease Patients And Centenarians

Name: Eduard Enkui Wanli Ma (4660668)
Student number: 4660668
Date: 10th of July, 2024
Program: Computer Science (Bioinformatics)

1st supervisor: Prof. Dr. Ir. M. Reinders
2nd supervisor: Dr. N. Tesi

Delft Bioinformatics Lab
Faculty of Electrical Engineering, Mathematics, and Computer Science
Van Mourik Broekmanweg 6
2628 XE Delft
The Netherlands

To mom and dad, my first teachers

To my brother, my first friend

Acknowledgments

First of all, I would like to express my sincere gratitude to Prof. Dr. Ir. Reinders and Dr. Tesi for giving me the opportunity to conduct this research and supervising my work. Their guidance and constructive feedback were of great value during every step along the way. Without their mentorship, the completion of this work would have been impossible.

I also would like to thank Dr. A. Salazar and Dr. Hulsman for the inspiring discussions that helped me better understand the outcomes of the research and kept me sharp in reaching conclusions. In addition, I extend my gratefulness to Dr. Lofi for serving on my thesis committee and evaluating my work.

Lastly, many thanks to all the colleagues and fellow students in the 100+ study group and the Delft Bioinformatics Lab who have shown interest in my research project.

Eduard Ma

In Silico Screening of Long-Read Sequencing Data Reveals Higher Incidence Of Endogenous Viral Elements in Alzheimer's Disease Patients Compared To Cognitively Healthy Centenarians

E. MA[1,2,*], N. TESI[1,2], AND M. REINDERS[1]

[1] *Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands*
[2] *Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands*
[*] *Correspondence: e.e.w.ma@student.tudelft.nl*

Compiled July 5, 2024

Ever since the origin of human life, we have been infected by a wide range of viruses. These pathogens have invaded our cells, leaving behind traces of their presence in our genome, known as endogenous viral elements (EVEs). Among the affected cells are neurons. The infectious hypothesis for Alzheimer's disease (AD) proposes that viral infections may serve as an environmental factor contributing to AD. In our study, we explored this hypothesis for the first time from an endogenous perspective by identifying EVEs in the long-read assembled genomes of both AD patients and cognitively healthy centenarians (CHCs). Using a custom-built data processing pipeline, our findings reveal that the genomes of AD patients harbor more EVEs than those of CHCs (p=3.24e-4, incidence rate ratio (IRR) = 1.27). Furthermore, we identified specific chromosomal regions with a higher incidence of viral integration in the AD cohort across different virus families. Notably, we found that remnants of the *Orthoherpesviridae* family tend to be located near genes that are differentially expressed in AD brains compared to healthy brains. Our data suggest that viral infections over time have increased the susceptibility to AD, underscoring the importance of preventive measures against infections.

Keywords: Paleovirology, Alzheimer's Disease, Centenarians, Endogenous Viral Elements, Long-Read Sequencing

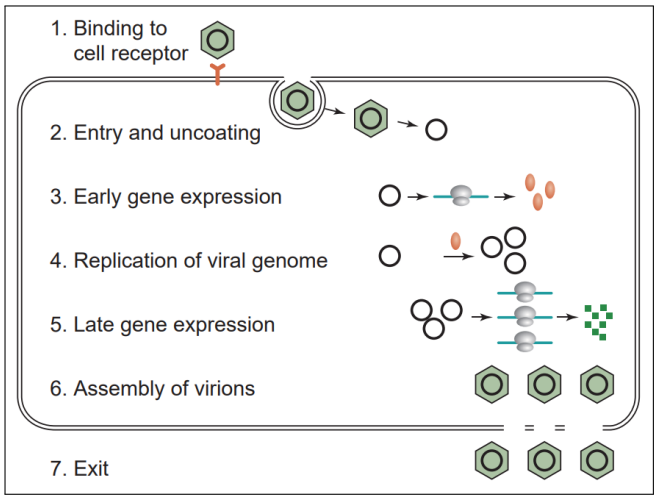

Fig. 1. The replication life cycle common to all viruses. A virus is drawn as a green hexagon. The host cell which is infected is represented by the double lined square area. Taken and adapted from figure 1.6 in [1].

INTRODUCTION

Viruses are the smallest, and perhaps paradoxically, the most abundant biological entities on our planet [2]. Not surprisingly then, there is an immense diversity within the virus kingdom. This diversity is best exemplified by the many different genome compositions in existence. Depending on the virus, its genome can consist of ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) and is either single-stranded (ss) or double-stranded (ds). Furthermore, specifically for RNA genomes, a distinction is made between whether a virion carries sense RNA (also known as positive RNA, hereafter mentioned as (+)RNA) that can be directly translated, or antisense RNA (also known as negative RNA, hereafter mentioned as (-)RNA) of which only its complementary copy can be translated [1]. Viruses always need a

Table 1. The influence of viruses on cognitive functions in the context of AD. △ = increased risk of cognitive impairment, □ = increased risk of AD. Taken and adapted from Table 1 in [8]

Virus Species	Influence
Herpes Simplex Virus Type 1 (HSV-1)	□
Herpes Simplex Virus Type 2 (HSV-2)	△
Varicella-Zoster Virus (VZV)	□
Epstein-Barr Virus (EBV)	□
Cytomegalovirus (CMV)	△
Human Herpesvirus 6 (HHV-6)	□
Hepatitis C Virus (HCV)	△
Influenza A Virus Subtype H5N1 (H5N1)	□
Influenza A Virus Subtype H1N1/09 (CA/09 H1N1)	□

host organism to survive and therefore play a pivotal role in our ecosystem. Their influence ranges from influencing domestic livestock populations and regulating oxygen and carbon levels in our oceans to shaping and manipulating the genomes of their hosts, for the better or worse from an evolutionary point of view [1, 3].

The manipulation of host genomes results from the general replication life cycle of a virus. Virions, i.e. individual virus particles, are relatively simple entities that lack many of the basic biochemical molecules needed for growth and replication, like enzymes that generate amino acids, nucleotides, ATP, etc. Hence, viruses are obligatory intracellular parasites, meaning that they can only replicate within a host cell. Despite the remarkable diversity in virus replication strategies, all viruses share a common replication life cycle in which the following tasks need to be performed (Figure 1): (1) bind to the appropriate cell receptor; (2) enter or invade the host cell and release of the viral genome (i.e., uncoating); (3) synthesize "early" proteins that are necessary for genome replication; (4) replicate the viral genome within the host cell, using its available resources; (5) express "late" proteins that are essential for packaging of the newly synthesized viral genome; (6) assembly of the (nucleo)capsid structure; and (7) escaping the host cell to spread to neighboring cells [1, 4].

A consequence of this life cycle is that viruses may integrate (part of) their genome into the host genome. During millions of years of virus-host interaction, some viruses managed to infect the gametes or cells of the early embryo in case of animals [5]. Once a viral genomic sequence is integrated into the genome of those cells, it will be passed on to all progeny cells. In this way, the viral genome is passed on vertically over generations following the principles of Mendelian genetics. Genomic insertions like these are called endogenous viral elements (EVEs) [6]. Today, EVEs make up about 5-8% of the human genome [7]. Viruses have therefore been and still are a huge source of horizontal gene transfer.

The first discoveries of viral elements in the human genome were derived from retroviruses, also called human endogenous retroviruses (HERVs) [9–11]. *Retroviridae* is a family of viruses that inserts a copy of its genome into the host cell's nuclear genome by creating a DNA copy of its RNA genome through a reverse-transcriptase enzyme [4]. Hence, by means of its natural way of replicating, including the mandatory reverse transcribing step, remnants of retroviruses are expected to be found in the genomes of their hosts. Infamous examples of a retrovirus today are the Human Immunodeficiency Virus 1 (HIV-1) and the Human papillomavirus (HPV), that cause Acquired Immunodeficiency Syndrome (AIDS) and cancer, respectively [4]. Because of their far-reaching pathological consequences, these integrations are thought to not spread in the host gene pool. Nevertheless, some integration events of other, less detrimental, retrovirus species in the past offered an evolutionary advantage and increased in frequency in the population. These integrations have become a HERV [12, 13]. An example of this symbiotic relationship is the *synctin* gene which is thought to have a retroviral origin and has been exapted in certain mammals, playing a crucial role in placental development [14, 15], or the recently discovered *RetroMyelin* gene which is essential for the production of myelin to insulate nerve fibers [16].

Retroviruses, however, are the only viruses for which genomic integration is a mandatory step in the replication cycle. The replication strategies of other viruses do not include integration into the host genome. Surprisingly though, many elements of non-retroviral origin, i.e. non-retroviral EVEs (nr-EVEs), have been discovered in the genome of animals and humans specifically, although they are much less common than endogenized retroviruses [6, 17]. Interestingly, their influence on the fitness of the host has been shown in some cases to have positive effects [18]. The integration of non-retroviral viruses is thought to be mediated by non-homologous recombination or through interaction with Class I transposable elements of the host cell, i.e. retrotransposons [6, 19, 20]. Overall, we now know that 5-8% of the human genome originated from a mixture of retro- and non-retroviral insertions, of which the majority is of retroviral origin due to its inherent integrative nature [10].

Even though the non-retroviral proportion in human genomes is small, there is one particular non-retroviral virus type that has received increased attention over the past decades in the field of AD, namely herpesviruses. Herpesviruses, i.e. the members of the *Orthoherpesviridae* family, have been around for millions of years and have infected human beings throughout their evolution constantly. Nine herpesviruses today infect humans and all of them establish a life-long latency after infection in neurons, lymphocytes, or other cell types which means that the viral DNA is harbored inside these cells for months, or even years, to eventually cause reactivation of the virus in the same individual [1]. In fact, it is estimated that the majority of our global population has been infected at least once with one of these nine viruses [1]. So why is this interesting to people working in the field of AD? AD is an inflammatory neurodenegerative disease and the most common form of dementia worldwide. On the molecular level, AD is hallmarked by the formation of extracellular amyloid beta (Aβ) plaques between neurons in the brain and neurofibrillary tangles made of hyperphosphorylated tau protein [21]. What exactly causes these aberrant molecular structures and why remains unclear. One fact that is widely accepted though is that AD is a multifactorial disease with many genetic as well as environmental contributory factors. One of these environmental factors has been hypothesized to be herpesvirus infections. Already back in 1991, traces of herpesvirus DNA were found back in the brains of AD patients and healthy controls [22]. Subsequent research revealed that in AD brains, 72% of the herpesvirus DNA was associated with Aβ plaques. In contrast, in aged control brains, which have a lower frequency of

amyloid plaques, only 24% of the viral DNA was associated with plaques. Moreover, experiments in mouse models demonstrated that herpesvirus infections upregulate Aβ producing enzymes β-secretase 1 and γ-secretase and cause tau hyperphosphorylation [23, 24]. Strikingly, it was later also shown that the Aβ plaques entrap virus particles in mice and are protective against herpesvirus induced encephalitis [25]. Much more research before and after these experiments all converged to the underlying thought that the formation of these Aβ and tau phosphorylation in the brain might be a response to protect individuals against acute infection in the short term, but contribute to plaque formation over the long term [25]. Long story short, the pile of data in favor of herpesvirus infection being a key environmental factor for the onset of AD is growing. These ideas have been extensively summarized as 'the infectious hypothesis': virus infections contribute to AD's pathogenesis [26]. Notably, the hypothesis nowadays includes many more virus species other than herpesviruses [8]. A summary list of the most important viruses is given in Table 1. We forward the reader to [8] for an exhaustive list.

Despite environmental influences, AD is predominantly hereditary, with evidence showing that the genetic profiles of individuals with AD differ from those of healthy control groups [27]. The connection between the infectious hypothesis and heritability may lie in the fact that herpesviruses and other non-retroviral viruses occasionally leave traces of their infections in their host's genome as EVEs, as previously discussed [28]. It would therefore be interesting to investigate whether the infectious hypothesis holds on the endogenous level, i.e., can we find back traces of these virus infections differently in the genomes of individuals with AD versus healthy controls. In order to investigate this, one needs numerous genomes of an AD cohort and a healthy control group, for sufficient statistical power, in addition to high quality data to make sure that these EVEs can be identified.

Especially the high quality of the sequencing data is important. Viral integrations can span in the range of hundreds to thousands of basepairs (bp) and often include repetitive elements [29]. Current state-of-the-art methods for short-read sequencing are limited to capturing only segments of around 250 bp [30]. Therefore, the repetitive nature of longer viral sequences is hard to capture. Screening approaches based on short-read sequencing have worked their way around this problem by skipping the assembly step. Their overall approach is to first align the short reads globally to the human genome in order to discard reads that map back to the human genome well, i.e. false positives. Only so-called discordant reads are kept and are screened for viral signals downstream by comparing them to a virus database [31]. Assembling the short reads that are punitively viral is then still a difficult task for longer integrations and these methods suffer in performance due to the smaller reads. Screening for EVEs in assemblies that were constructed from shorter read lengths thus leads to inevitable loss of signal. Long-read sequencing data would make up for this loss of signal by being able to capture bigger fractions of the viral remnants in the genome within one read and thus making the assembly of the host genome, including the viral integration, more accurate (Figure 2).

The 100+ study in The Netherlands has collected long-read sequencing data sets of hundreds of Dutch AD patients and CHCs, i.e. individuals that have reached the age of 100 or older with good cognitive and physical capabilities [32]. CHCs have managed to avoid AD development, leading to the belief that their genomes lack the genetic elements associated with the disease. Having genomic data of individuals at both extremes of the cognitive spectrum is ideal for investigating the infectious hypothesis at the genomic (endogenous) level.

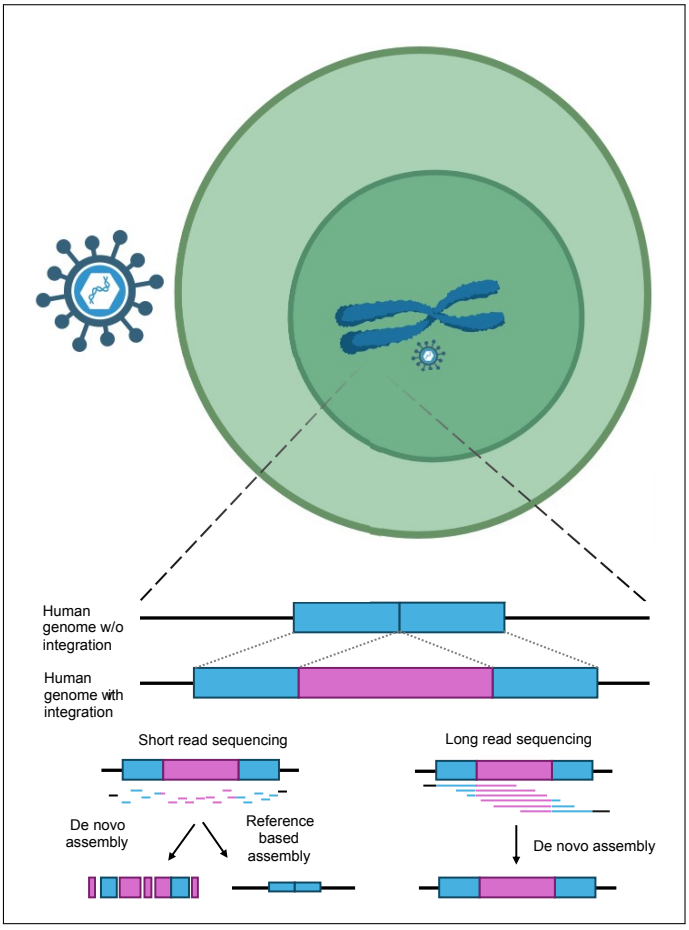

Fig. 2. Short- and long-read alignment for a viral integration. Long read sequencing reads are able to capture the genomic integration (pink), including the surrounding genomic elements (blue), way better than short read sequencing can due its large read size. De novo assembly can then reconstruct a viral element inside the host's genome. Reference based assembly of short-reads often discard reads of unknown origin and de novo assembly using short-reads brings along challenges of reconstructing viral elements from these smaller fragments such as gaps and eventually potential misassemblies.

In this research, we investigated the infectious hypothesis, for the first time, at the genomic level by examining differences in EVEs between AD patients and CHCs. To achieve this, we designed a BLASTx-based genome screening pipeline that directly analyzes human genome assemblies constructed from high-quality Pacific Biosciences (PacBio) long reads. This approach allows us to search for EVEs across the entire human genome of both AD patients and CHCs, rather than constructing EVEs from short reads as done in previous studies [31] (Figure 3). Considering that the infectious hypothesis now encompasses many more viruses beyond just herpesviruses, we scanned all genomes against the non-redundant RefSeq virus protein database [33]. The analysis is structured as follows:

1. **How many:** how many (remnants of) viral integrations are

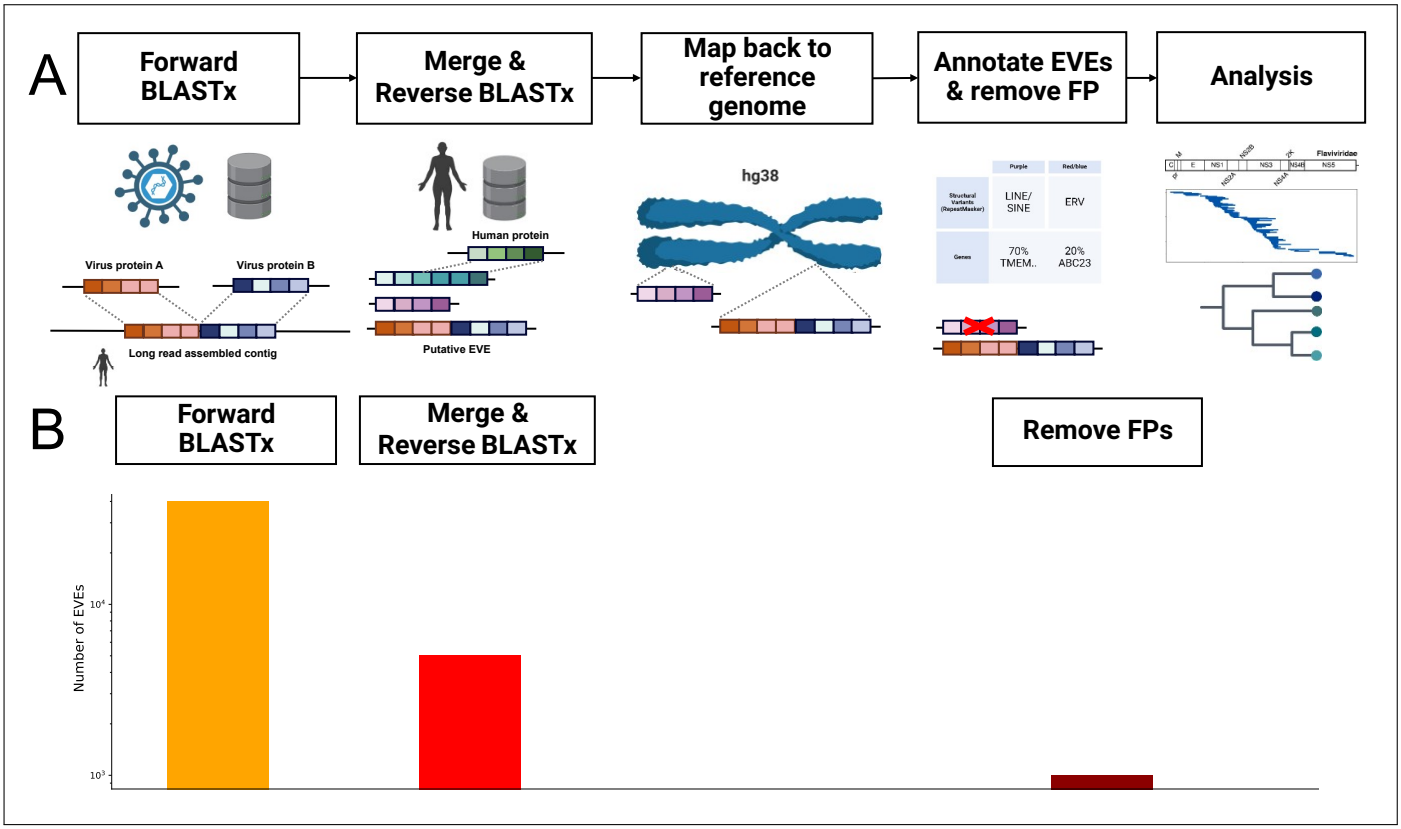

Fig. 3. **The EVE detection pipeline used in our work. A) Human sequences are queried against the a viral database. The putative EVEs are queried against a human protein database. The suspect EVEs that resemble human sequences are discarded. The contig coordinates are mapped back to positions on human chromosomes. Further filtering is performed through removing putative EVEs that overlap with known regions like LINEs and SINEs. Annotation of the regions with data like nearby gene information. B) Rough estimations of how many hits were found per filtering step in our pipeline for one genome.**

found back in the human genome?

2. **How long:** how long (in bp unit) are the (remnants of) viral integrations?

3. **Where:** where in the human genome are (remnants of) viral integrations located and what portion of the viral genome do they originate from?

Our findings show that (1) AD genomes contain more EVEs of a diverse range of vertebrate viruses in general; (2) AD genomes more frequently harbor *Borna-*, *Orthoherpes-*, and *Poxviridae* EVEs at specific chromosomal locations and the total EVE length is longer at these chromosomes compared to the CHC control group; (3) EVEs tend to be located near genes that are differentially expressed in AD brains compared to healthy brains, and; (4) a higher incidence of *Retroviridae* EVEs is observed in AD genomes compared to CHC genomes near AD associated single nucleotide polymorphisms (SNPs).

METHODOLOGY

Data Acquisition & Preprocessing

In this research, we included 249 individuals diagnosed with AD from the Amsterdam Dementia Cohort [34]. As controls, we studied 240 CHCs from the 100-plus Study [32]. All participants and/or their legal representatives provided written informed consent for participation in clinical and genetic studies. Blood samples of all individuals were taken, processed, and assessed for minimum quality requirements as described in [35]. These were subsequently long-read sequenced using PacBio Sequel IIe instruments with at least 1 SMRT cell. Whole-genome de novo assembly using hifiasm was performed and were aligned to GRCh38 using Holstege Lab's processing pipeline which is freely available at https://github.com/holstegelab/snakemake_pipeline [36]. Their pipeline outputs assemblies of the paternal, maternal, and primary haplotypes in the form of contigs. In this research we focused on the primary assemblies, i.e. the assemblies that collapsed both the maternal and paternal haplotypes.

Detection Pipeline

An overview of the pipeline is shown in Figure 3. Briefly, we: (1) screened contigs assembled from long-read sequencing data for subsequences that are similar to virus proteins using a local alignment algorithm; (2) merged viral-like regions that overlapped within the same genome, choosing the most significant virus protein as the representative virus sequence for that region, and compared all regions to human proteins to discard any that showed significant similarity to human proteins; (3) mapped the raw contig coordinates to human genome coordinates; (4) discarded regions that were situated within non-viral structures of the human genome and annotated our regions with data like nearby gene information; (5) and performed downstream analysis to infer any differences between the cohorts. The detection pipeline is freely available at: https://github.com/MaEduard/master_thesis

EVE Detection - Forward Screening

The contigs that resulted from the assembly process of all AD individuals and CHCs were screened *in silico* using the BLASTx approach against the non-redundant Refseq NCBI virus protein database following common standards in the field (https://ftp.ncbi.nih.gov/refseq/release/viral/, accessed March 2024) [5, 29, 37, 38].

We first identified putative EVEs by performing a DIAMOND BLASTx search, specifying our genomes as the query and the RefSeq NCBI library database as the database [39]. We call this the "forward" BLASTx search. Our assembly files are on the order of gigabytes, and the RefSeq database contains 683,238 protein sequences. Therefore, we chose DIAMOND BLASTx, as it is thousands of times faster than NCBI's default BLASTx algorithm [39]. The nucleotide query sequences (i.e. human assembled contigs) were translated in six reading frames (resulting in six protein sequences) and were compared against the viral protein sequence database using default settings. Human subsequences with high similarity to viral protein sequences (e-value $< 10^{-6}$) were extracted. Overlapping regions within one genome were merged and the alignment with the lowest e-value was chosen as representative for that merged region. Nucleotide sequences of the putative EVEs were then extracted using BedTools [40].

EVE Detection - Reverse Search

Putative EVEs that are highly similar to human proteins are considered false positive. We thus again perform a DIAMOND BLASTx alignment but this time with the general protein database as input (https://www.ncbi.nlm.nih.gov/genome/guide/human/, accessed January 2024) [39]. The nucleotide regions of our human genomes, that we suspect to be EVEs, are translated into proteins and these translated sequences are then compared to human proteins. We call this the "reverse" BLASTx search. Any subsequence that is highly similar to a human protein (e-value $< 10^{-6}$) and overlaps for at least 50% with that human protein, was considered false positive and was excluded from any downstream analysis. Of note, reverse BLASTx alignments that were similar to human proteins known as 'endogenous viral' were excluded in our reverse BLASTx filtering step.

Mapping

In order to identify the location of EVEs along the GRCh38 reference genome, the alignment files after the reverse BLASTx procedure were processed with an in-house script parsing the CIGAR strings of every assembly's alignment file (in BAM format). This step maps the contig coordinates to GRCh38 coordinates.

False Positive Removal And Annotation

Some identified EVEs may still have significant homology with repetitive regions in the genome not attributable to viral elements. We used RepeatMasker annotations to identify structural variants overlapping our discovered EVEs [41]. Any putative EVE that overlapped for more than 50% with an annotated human sequence was considered a false positive, except for sequences that RepeatMasker itself classified as "ERV" or "Borna-like". Additionally, we filtered our experimental data by only keeping virus families known to integrate in vertebrate species as reported by the International Committee on Taxonomy of Viruses (ICTV) Master species list (`ICTV_Master_Species_List_2023_MSL39.v1.xlsx`, https://ictv.global/msl, accessed May 2024). Of note, *Metaviridae*, classified as a vertebrate virus family in ICTV, but truly being a collection of retrotransposons (i.e. Ty3/Gypsy Long-Terminal Repeat (LTR) retroelements), was still taken into account in the analysis but we kept in mind its possible non-viral nature. [42]. Lastly, we annotate the EVEs using the NCBI taxonomy database and annotate EVEs being inside a gene or being intergenic based on the intersection of EVE and gene coordinates using BedTools intersection with default parameters (https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/, accessed March 2024) [40]. Using the same procedure, we annotate EVEs to be close to an AD SNP (100 kbp up- or downstream of the SNP) or not [27].

Statistical analysis

We tested whether the number of EVE was significantly different between AD individuals and CHCs using negative binomial regression analysis (NBR) [44]. Similarly, the lengths of EVEs was compared between AD individuals and CHCs using two-tailed Mann-Whitney U testing (MWU) [45]. Difference in proportions of the two cohorts was tested using two-tailed Fisher's exact test (FE) [46]. Wherever multiple testing was involved, we corrected our p-values following the Benjamini-Hochberg procedure (BH) with significance threshold set at 0.05 [47].

Implementation

The EVE detection pipeline was implemented as a mixture of bash and python scripts. All figures were created using matplotlib, Seaborn, Pygenometracks, and ETE Toolkit plotting packages [48–51]. Statistical analysis was performed through Scipy and PyMC statistical software packages [52, 53]. Wherever gene set enrichment analysis was performed, gProfiler was used with standard settings [54].

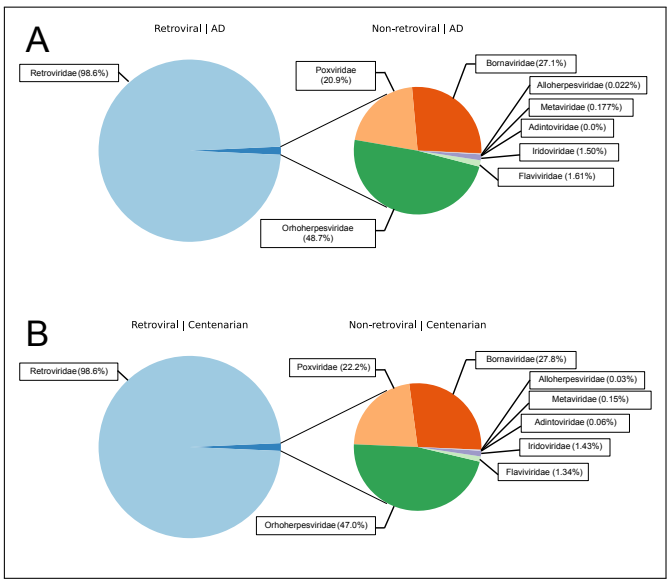

Fig. 4. A/B) The average proportions of different families in an AD individual (A) or CHCs (B). The blue, bigger circles highlights the retroviral proportion vs. the non-retroviral proportion. The smaller circle gives an overview of the proportions of non-retroviral EVEs discovered.

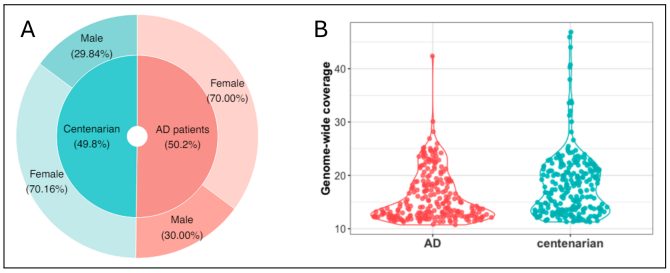

Fig. 5. **100+ study data characteristics outlining the relative sizes of the two cohorts and their sex proportions (A) and the distribution of genome coverage for the samples of both cohorts (B). Taken and adapted from figure 6 in [43]**

RESULTS

Demographics and Sequencing

The genomes of 249 individuals diagnosed with AD and 240 CHCs were sequenced using PacBio long-read technology and preprocessed as described previously. The majority of the participants in the study are female (Figure 5A). Additional information on the cohorts and samples used is available elsewhere [35]. Overall, the AD genomes were sequenced at a median coverage of 18.2x, with median read-lengths of 14.8 kbp; the CHC genomes were sequenced at a median coverage of 20.1x, with median read lengths of 14.7 kbp (Figure 5B).

AD genomes harbor more EVEs

Numerous EVEs were identified in the genomes of both CHCs and AD patients. On average, 1268 (±54) EVEs were found in AD genomes whereas 999 (±53) were found in CHCs. We thus see that in general the genomes of AD samples have a higher number of EVEs than CHC genomes (p=3.24e-4, IRR = 1.27, NBR).

In total, nine viral families, that are known to infect vertebrates, were identified including (+)RNA viruses (*Retro-, Flavi-*, and *Metaviridae*), (-)RNA viruses (*Bornaviridae*), as well as ds-DNA viruses (*Orthoherpes-, Alloherpes-, Pox-, Irido-* and *Adintoviridae*).

Retroviridae EVEs were predominantly identified in the genomes of both cohorts, as is expected given the obligatory integration step during infection that was previously discussed (Figure 4A & 4B, light blue portions). Zooming in, 1.4% of the EVEs correspond to non-retroviral hits which consist of a mixture of different virus families. In both cohorts, we observed sequences similar to viral proteins of the family *Orthoherpesviridae* as the second largest virus family. *Orthoherpesviridae* is the family of viruses that comprises the HHV-6A, HHV-6B, EBV, HSV-1, HSV-2, and CMV which were previously related to AD (Table 1 and Figure 4C & 4D, green). Additionally, a small fraction of the data found human sequences that were similar to *Borna-, Pox-, Irido-, Meta-, Alloherpes-, Adinto-* and *Flaviviridae* proteins.

To further investigate the differences between virus families, we analyzed what portion of a cohort carries an EVE of a specific virus family (Figure 6A). Due to their high abundance in the human genome, the *Retroviridae* family is present in all individuals across both cohorts, each individual carrying at least one endogenous viral element (EVE) from this family. For almost all non-retroviral virus families, we find that a larger fraction of the AD cohort is carrying at least one virus integration in their genome of the specified family, except for *Adintoviridae*. Statistical testing of these distributions did not show any significant difference after correction (FE, BH correction).

Looking at the counts of EVEs per family in an individual, i.e. the number of EVEs within one genome for a specified family, we observe more EVEs for *Retro-, Borna-*, and *Orthoherpesviridae* in AD genomes than in CHC genomes (p_{Retro}=3.00e-2 (IRR = 1.27), p_{Borna}=3.00e-2 (IRR = 1.29), $p_{Orthoherpes}$=8.49e-3 (IRR = 1.38); NBR, BH correction) (Figure 6B).

The total EVE length is longer in AD genomes than in CHC genomes

The difference between the two cohorts is also observed in the findings of the total EVE length analysis (Figure 6C). The identified EVEs in AD spanned a total length of 1,262,581 bp on average and ranged from 98 to 5372 bp in length with a median length of 791 bp. In contrast, the identified EVEs in CHCs covered on average 1,027,716 bp and ranged from 107 to 5372 bp length with a median length of 827 bp. Hence, the total EVE length per AD individual is significantly longer than the total EVE length of a CHC genome (p=1.91e-4, MWU). In general, we observe a lot of heterogeneity in EVE length among different virus families. The remnants of *Retroviridae* are preserved best. Lengths of other families are smaller and should be interpreted carefully. To exemplify, the EVE lengths of *Orthoherpesviridae* are mostly between 250-500 bp, whereas the genome length of herpesviruses are known to be at least around 125 kbp [55]. Nevertheless, we observe that the total EVE length of *Retro-, Borna-, Pox, Flavi-*and *Orthoherpesviridae* are significantly longer in the AD cohort compared to the CHC cohort (p_{Retro}=1.75e-3, p_{Borna}=2.86e-3, $p_{Orthoherpes}$=2.86e-3, p_{Pox}=1.32e-2, p_{Flavi}=1.26e-2; MWU, BH correction).

EVEs are unevenly spread across the human genome

To get a glimpse of where exactly on the chromosomes EVEs were discovered, we merged hits that overlapped into one interval (Figure 7A). Investigating any chromosome, except for the Y chromosome, shows that more EVEs were present in the AD cohort than in the CHC cohort. Moreover, we observe that for chromosome 13, 14, and 15 no abundantly shared hits were found at the start of the chromosomes, indicating that this region is less attractive for viral integration. A complete overview is given in Figure S1.

Multiple virus families are enriched on AD chromosomes

The difference in proportions between the two cohorts having an EVE of a specific virus family was analyzed per chromosome (Figure 7B, 7C). In both cohorts, the *Retroviridae* family was present on approximately all human chromosomes, except for the Y chromosome (Figure 7B & 7C, left column). Only males carry a Y chromosome and thus the 30% incidence of retrovirus integration on the Y chromosome in each cohort can be explained by the fact that approximately 30% of each cohort is male (Figure 5). The non-retroviral families were spread unevenly across chromosomes and show a different spread per family (Figure 7B & 7C, all but the left column).

Next, we examined the variations between the two cohorts per family and per chromosome (Figure 7D, green boxes). Specifically, we observed that the *Bornaviridae* family was enriched for the AD cohort on chromosome 9 (p=2.30e-2, odds ratio (OR)=1.91; FE, BH correction) and 10 (p=2.30e-2, OR=1.91; FE, BH correction). Furthermore, relatively more AD samples carry a *Poxviridae* EVE on chromosome 11 (p=2.9e-2, OR=1.86; FE, BH correction) and 19 (p=4.0e-2, OR=1.79; FE, BH correction). Lastly, *Orthoherpesviridae*, was found to be enriched at four chromosomes in the AD cohort compared to the CHC cohort, namely

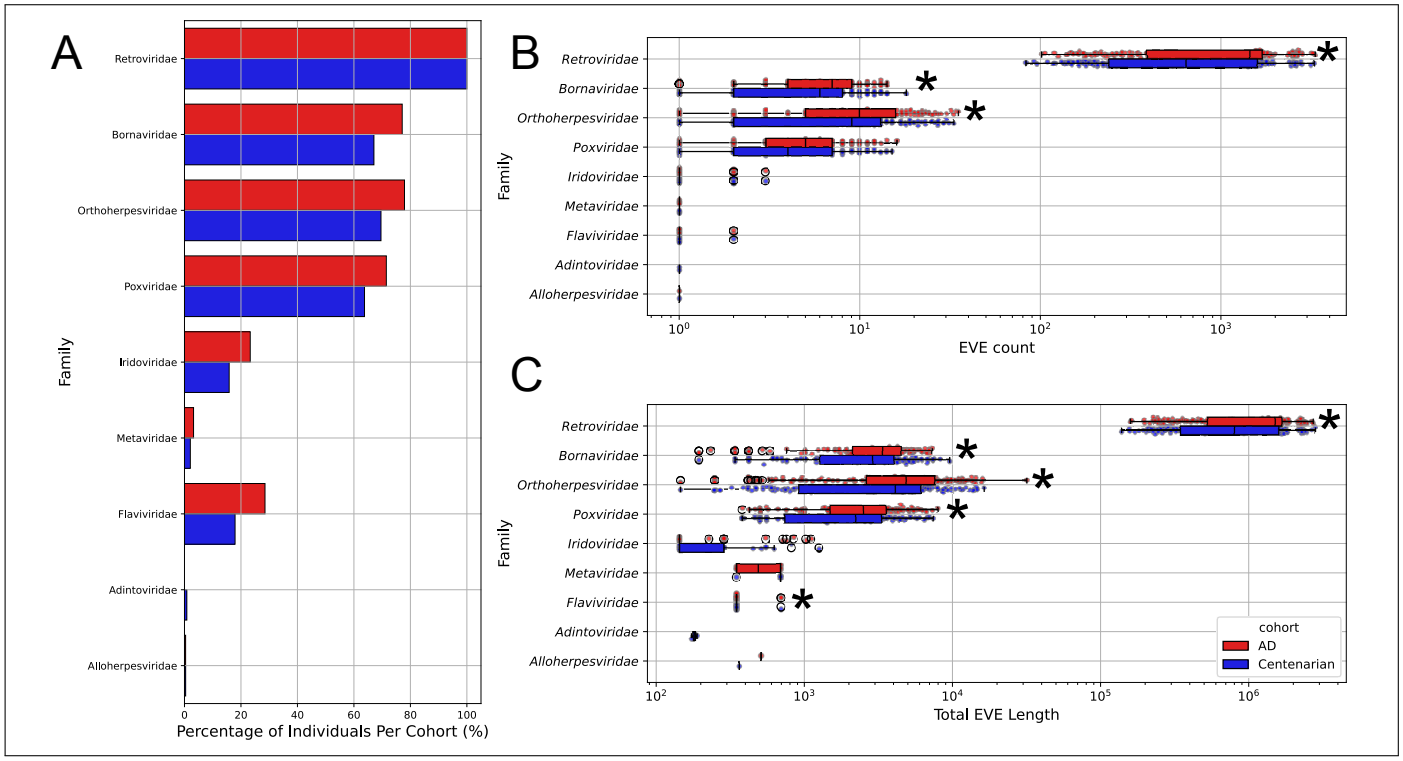

Fig. 6. A) Percentage of cohort having at least one EVE identified of the specified family. B) EVE count per family. Every dot represents an individual's total number of EVEs identified for the specified family. C) The total EVE length (bp) distribution per family. Every dot represents an individual and the total EVE length is the sum of all EVEs' length in an individual. Stars indicate statistical significance (p < 0.05) after false discovery correction. Box plots show median value as the middle of the box and the box itself shows the interquartile range which contains the middle 50% of the data.

on chromosome 1 (p=2.3e-2, OR=1.96), 7 (p=3.50e-4, OR=2.47), 17 (p=1.9e-2, OR=2.01), and 19 (p=1.2e-2, OR= 2.07) (FE, BH correction).

These findings narrowed down our analysis to specific chromosomes for different virus families. Per family we merged overlapping EVEs on a specific chromosome within a cohort and counted the number of samples that had an EVE in that merged region. We then intersected these regions between the cohorts and performed FE tests to investigate where enrichment was to be found. After BH test correction we were left with 10 regions for the non-retroviral families *Borna-*, *Pox-* and *Orthoherpesviridae*. For *Retroviridae* we found numerous of these regions throughout the genome and highlight in this research those regions close to AD associated SNPs, i.e. within a 100 kbp interval up- or downstream of the SNP [27]. An overview, including the statistics, is given in Table 2.

The next sections discuss these differences in more detail for *Borna-*, *Pox-*, *Retro-* and *Orthoherpesviridae*. For every section, we first outline the characteristics of the virus family on the whole genome level and then zoom into the differences observed previously. Also, we show where approximately the EVEs originated from in the virus genome.

Orthoherpesviridae

The *Orthoherpesviridae* family encompasses the species of herpesviruses that are most interesting to us due to their association with AD. On average, we found 11 (±0) EVEs of this family in an AD genome compared to 9 (±0) EVEs in a CHC genome. The remnants covered 4204 bp on average in the AD cases whereas in CHCs, the average coverage was 3003 bp. Members of the *Orthoherpesviridae* family are divided in two subgroups, $\alpha-$ and β-herpesviruses. Almost all EVEs resembled proteins of α-herpesviruses (Figure 8C). Specifically, the Chelonid α-herpesvirus envelope protein represented the majority of the EVEs in both AD individuals and CHCs. This envelope protein is part of one of the conserved domains in the α-herpesvirus genome (Figure 8A). Protein comparison analysis showed that the Chelonid envelope protein is phylogenetically close to the human β-herpesvirus 6A and 6B envelope proteins. $\beta-$herpesvirus proteins were found sporadically in both cohorts (Figure 8B). Lastly, we saw that most herpesvirus EVEs were intergenic, meaning that they were situated between genes. The rest was either located inside introns or inside exons (Figure 8D).

AD individuals carry longer herpesvirus EVEs and show enrichment of herpes EVEs near genes that are differentially expressed in AD brains compared to healthy controls

We observed differences per chromosome in proportions of the two cohorts that harbored a herpesvirus EVE. Specifically, for chromosome 1, 4, 17, and 19 we observed proportionally more AD individuals with herpesvirus EVEs than CHCs (Figure 7D and Table 2).

On chromosome 1 the total EVE lengths were observed to be longer for AD individuals than CHCs (Figure 9A) (p=3.45e-3; MWU, BH correction). Furthermore, we listed all the genes closest to the herpesvirus remnants on chromosome 1 for which we found more AD samples than CHCs : RAB42, ODF2L, TGFBR3, TRIM33, RHEX, STUM, ZNF678. Gene set enrichment analysis using gProfiler did not show strong enriched pathways for

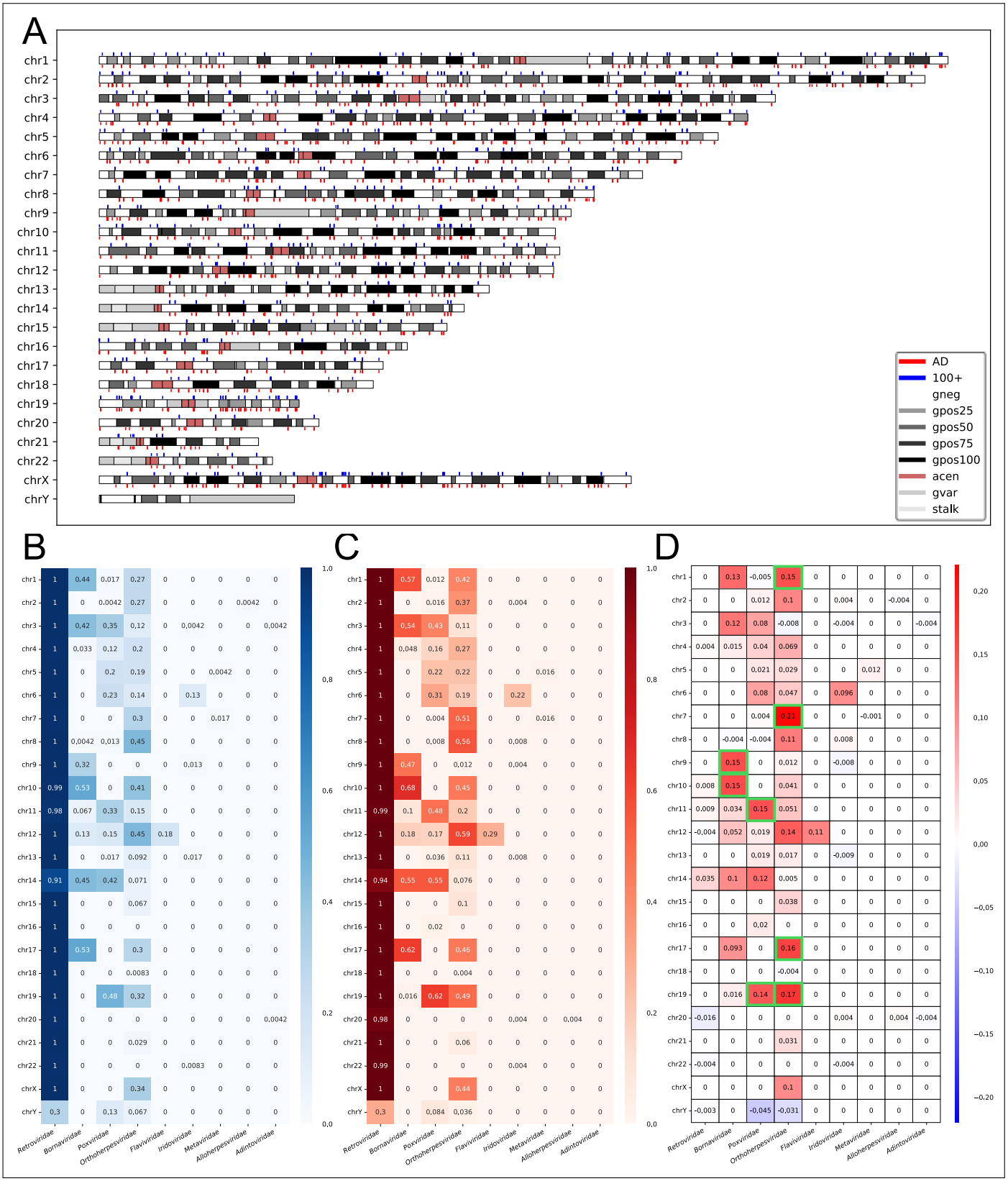

Fig. 7. **A) Karyotype plot of the human genome including common integration sights in AD (red, bottom of each chromosome) and CHCs (blue, top of each chromosome). EVE intervals were merged separately for each cohort and displayed if at least 20% of the population ($N_AD + N_{CHC}$) had an integration in that interval. Cytogenetic bands colors (all colors but red and blue) indicate gene density on the chromosome. Of note, the EVEs on the chromosome Y did not meet the 20% threshold and are thus not displayed on this plot. B/C) Distribution plot showing per chromosome and per virus family the fraction of the CHC (B) or AD (C) cohort having at least one EVE on the specified chromosome for the specified family. D) Difference in distributions of B and C between the two cohorts. Positive (red) values indicate enrichment in AD genomes whereas negative (blue) values indicate the enrichment in CHC genomes. Green boxes highlight significant differences ($p < 0.05$) between the cohorts based on FE test after BH correction.**

Table 2. Overview of the significantly enriched regions for different virus families. Every region is specified through a chromosome number and start (left) and stop (right) coordinates. P-values and odds ratio are reported for the FE tests after BH correction. SNP gene represents the gene in which a SNP associated to AD lies if the specified region is in close proximity (-100 kbp or +100kbp of the SNP) [27].

Family	Chromosome	Coordinates	N_{AD}	N_{CHC}	Odds Ratio	p	Closest Gene	SNP gene
Orthoherpesviridae	1	28587945-28588796	60	25	2.730	1.57e-3	TAF12	-
	7	5021640-5022178	88	54	1.882	2.39e-2	RBAK-RBAKDN	-
	17	50413398-50413999	92	59	1.797	5.33e-2	ACSF2	-
	19	50347788-50348322	89	53	1.96	2.08e-2	NASPA	-
Bornaviridae	9	37086819-37087689	79	52	1.680	3.54e-2	EBLN3P	-
	9	96008246-96009126	80	49	1.845	1.99e-2	ERCC6L	-
	10	2208886-2220995	159	110	2.088	1.16e-3	EBLN1	-
Poxviridae	11	78085840-78086304	117	79	1.806	1.02e-2	NDUFC2	-
	19	18578415-18578915	152	108	1.915	3.70e-3	UBA52	-
	19	20865623-20866179	4	17	0.214	1.41e-2	ZNF66	-
Retroviridae	6	32560043-32566098	27	2	14.472	1.36e-4	HLA-DRB1	HLA-DQA1
	7	12292625-12295056	56	21	3.026	3.40e-3	VWDE	TMEM106B
	17	49172477-49175501	165	123	1.868	3.94e-2	B4GALNT2	ABI3

this set of genes (Figure S4 and S5). Strikingly, however, all but the RHEX gene showed significant differential expression in AD brains for different regions according to the AMP-AD study [56] (Table S1). The only region on the chromosome itself that was enriched after BH correction lied in the interval of 28587945-28588796 (Figure 9E chr1, green arrows & Table 2). Upon closer inspection of this genomic region, we observe that the remnants were situated within the last out-of-frame exon of the TATA-Box Binding Protein Associated Factor 12 (TAF12) gene (ENST00000685589.1) (Figure S3). TAF12 is differentially expressed in the cerebellum and the temporal cortex in both males and females [56, 57].

What's more, genes nearby EVEs on chromosome 7 that were more common in the AD cohort resulted in the following list of genes: RBAK-RBAKDN, OSBPL3, ANLN, CYP3A43, AKR1B1. Gene set enrichment analysis of these genes in gProfiler showed weak enrichment for the steroid metabolic process (GO:0008202) (Figures S8 and S7). Consulting the Agora AD association database [57], it turned out that all genes have significant differential expression in at least one brain region in AD samples and all gene loci have a significant brain expression Quantitative Trait Locus (eQTL), except for RBAK-RBAKDN for which no data was available [56] (Table S2). In addition, the total EVE length distributions on chromosome 7 indicate that the EVEs on chromosome 7 are longer (Figure 9B) (p=1.9e-4; MWU, BH correction). We obtained one region significantly enriched on chromosome 7 for the AD cohort after BH correction (Figure 9E chr7 green arrows & Table 2). The interval lies within intron 2 out of 7 of the RBAK-RBAKDN gene (ENSG00000272968.5) (Figure S6). Lastly, it is situated within a repetitive LTR sequence as classified by RepeatMasker in the reference genome GRChg38, indicating its mode of integration [6, 19, 20].

On chromosome 17, gene set enrichment analysis of the genes closest to the merged region (SLFN12L, KRTAP9-7, KRT17, ACSF2, RNF213-AS1) showed weak enrichment for the keratin filament pathway (GO:0045095) (Figure S11 and Figure S10). The last three genes show different RNA expression profiles in AD brains compared to healthy controls [57] (Table S3). Zooming in further, no region was significantly enriched for AD. The only region close to significance was located in the interval 50413398-50413999 (Figure 9E chr17, green arrows & Table 2). It is situated in an intergenic region upstream of ACSF2. ACSF2 has been observed to have higher expression in many AD brain regions compared to controls [56, 58]. Chromosome 17 in general had a

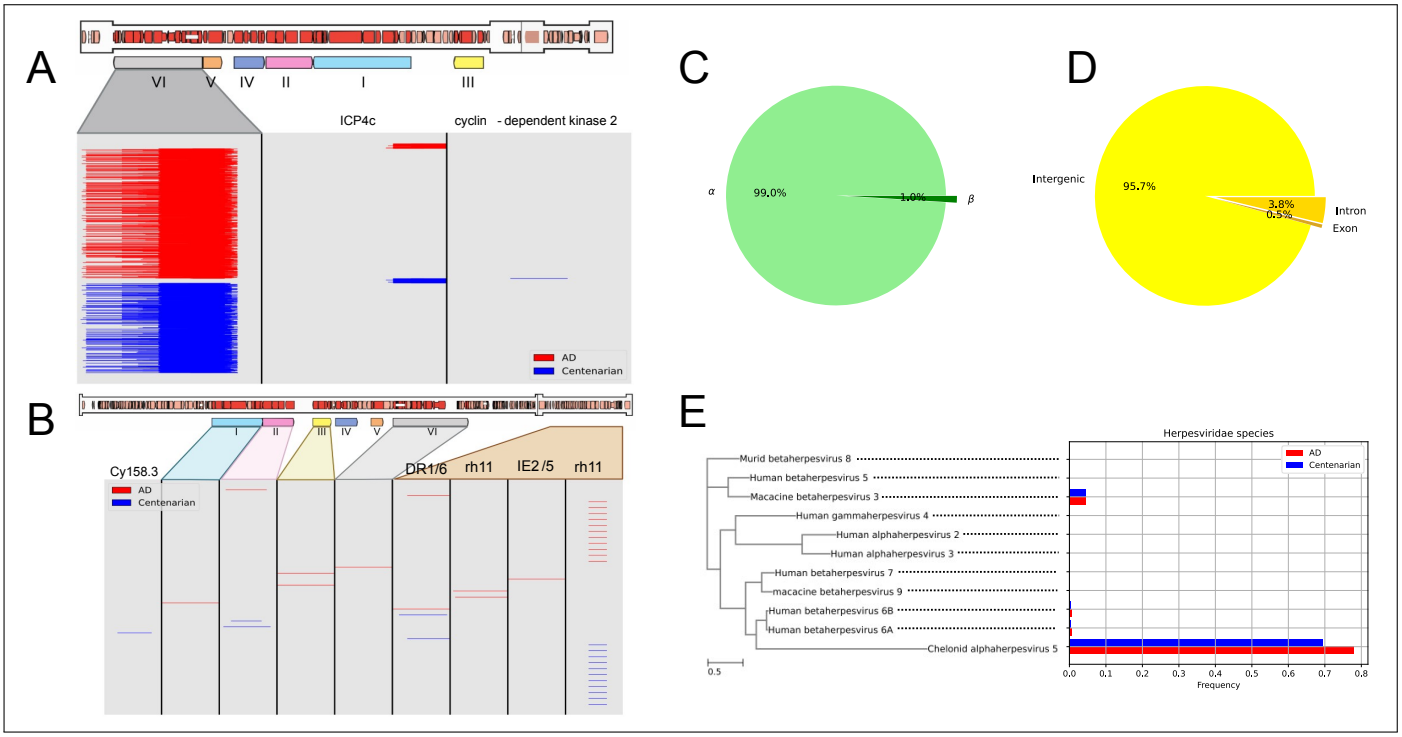

Fig. 8. A/B) Coverage plot of all *Herpesviridae* sequences in representative figures of the α-herpesvirus and β-herpesvirus genome in A and B, respectively. Roman letters indicate the six conserved domains in the herpesvirus genome: (I) Uracil-DNA glycosylase (II) Helicase-primase (III) DNA packaging terminase (IV) Major capsid protein (V) DNA polymerase catalytic subunit (VI) Envelope glycoproteins. Proteins that map back to non-conserved regions are annotated in the figure. Rough estimates based on literature show where they come from. Cy158.3 could not be mapped back. Schematics of the genomes were taken and adapted from [55]. C) Proportions of α- and β-herpesvirus found in the two cohorts. D) Proportions of intergenic, exonic, and intronic herpes EVEs in the two cohorts E) Simplified phylogenetic relationship between different herpesvirus species and their frequency in our genomes. Branch length indicates average number of substitutions per site. Tree was created using glycoprotein B protein sequence of every species with FastTree [39]. All branches have local bootstrap confidence > 0.7.

longer total EVE length distribution for the AD cohort compared to the CHCs (Figure 9C) (p=4.96e-3; MHU, BH correction).

Lastly, genes nearby herpesvirus integrations that had higher incidence of AD samples than CHCs on chromosome 19 were: ZNF93, ZNF682, ZNF626, ZNF98, NAPSA, ZNF813. All these zinc finger (ZNF) genes show significant differential expression in the brains of AD patients and all are reported to be a brain eQTL [57] (Table S4). Gene set enrichment analysis shows significant results of several pathways of which the strongest is "Factor: TFCP2; motif: ACCGGTTNAAACYGGT; match class: 1" (TF:M03949_1) (Figure S13 and S12). Additionally, another pathway for these genes that was enriched is KEGG:05168, termed "Herpes Simplex virus 1 infection". Length distributions shows that herpes EVE length is longer in the AD cohort (Figure 9D). After analyzing all regions with EVEs on chromosome 19, we observed one statistically enriched region, namely 50347788-50348322 (Figure 9E chr19, green arrows & Table 2). The closest gene to this region is NASPA (ENST00000253719.7). NAPSA has been shown to have a higher expression in the cerebellum and parahippocampal gyrus of AD brains compared to healthy controls and has an eQTL [57].

Bornaviridae

The members of the *Bornaviridae* family carry a linear (-)RNA molecule of approximately 9 kbp as their genome. In our data, we find on average 6 ($\pm$0) bornavirus elements back in the individuals of the AD cohort whereas we observe 5 ($\pm$0) in the individuals of the CHC cohort. The AD cohort's sequence length of these remnants span on average 2567 bp. For the CHCs, we observe a lower average sequence length of 1984 bp. Most remnants resembled the conserved nucleoprotein (Figure 10A). The other bornavirus EVEs resembled glycoproteins (Figure 10B). What stands out is that the nucleoprotein EVEs were better conserved than the glycoproteins.

Bornavirus EVEs have a weak link to AD

The exploratory analysis showed that we observed more AD individuals with a bornavirus-like element on chromosome 9 and 10 compared to the CHC cohort (Figure 7D). Interestingly, AD individuals have a longer total EVE length on both chromosomes compared to CHCs (p_{chr9}=6.45e-3, p_{chr10}=4.54e-3; MWU, BH correction). Zooming in further, for chromosome 9, gene set enrichment analysis on the three nearby genes (RUSC2, EBLN3P, ERCC6L2) to AD enriched regions did not show significant results. All three did show a difference in RNA expression in AD brains, however [57] (Table S5). For chromosome 10, gene set enrichment analysis of genes nearby the EVEs (EBLN1 and TBC1D12) with higher AD incidence than CHC incidence also showed no results. The only link found was that TBC1D12 has a a brain eQTL and showed significant RNA expression changes several regions of AD brains [57]. Three regions on those chromosomes were observed to be significantly enriched (Figure 10D green arrows & 2). Two of three, chr9:37086819-37087689 and chr10:22208886-22209950, map back to previously discovered bornavirus EVEs (EBLN3P and EBLN1). The third region, chr9:96008246-96009126, is situated inside the last intron of the ERCC6L2 gene (ENSG00000182150.20) (Figure S15). ERCC6L2 is

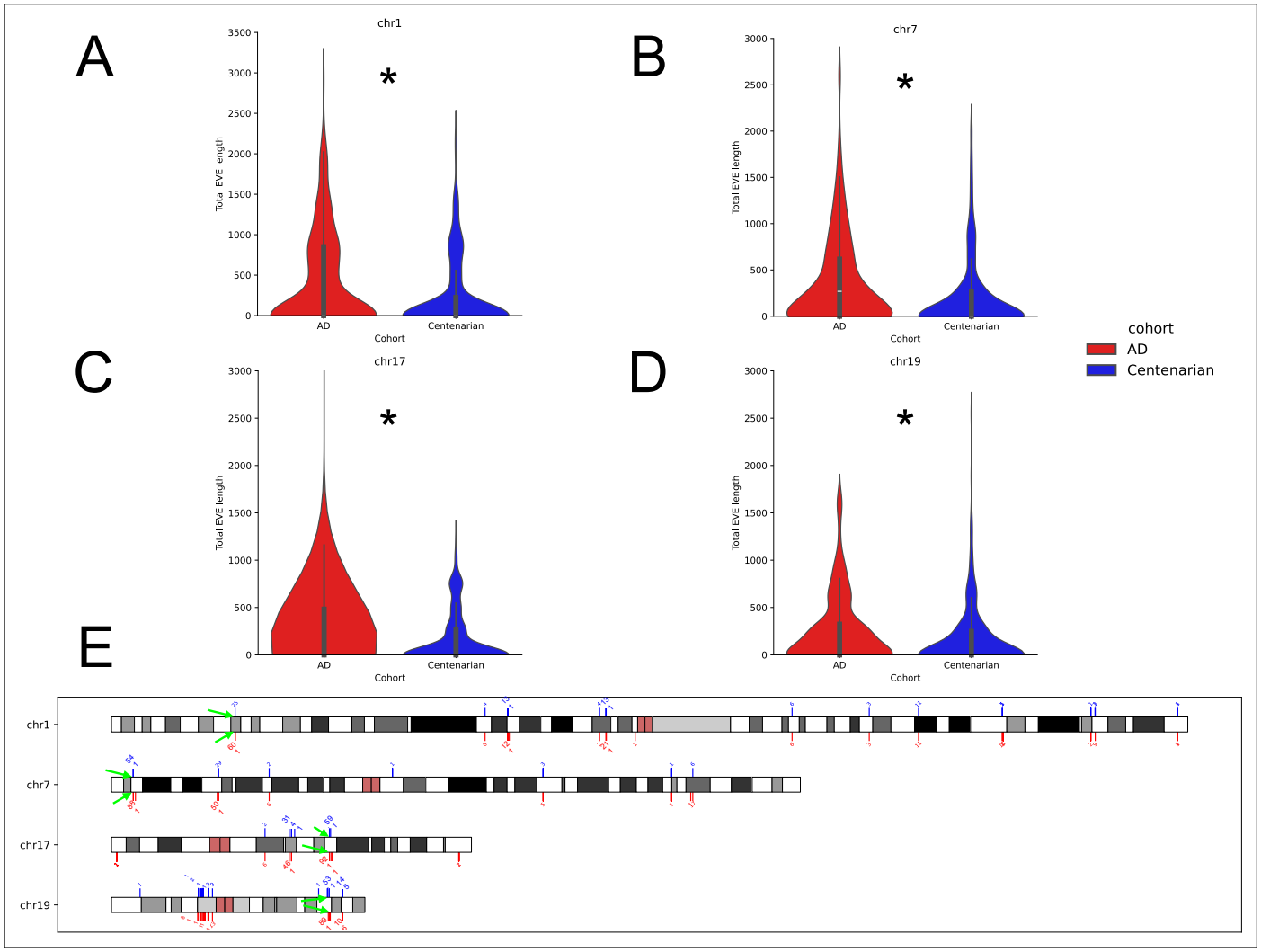

Fig. 9. **A/B/C/D) Total EVE length of herpesvirus EVEs distributions for chromosome 1 (A), 7 (B), 17 (C), and 19 (D). Stars indicate significance after BH correction. Chromosome locations of herpesvirus EVEs and the EVE length distribution on those chromosomes. E) Chromosome locations of herpesvirus EVEs including number of samples that have an EVE in that region. Discussed (significantly) enriched regions are indicated with green arrows.**

a brain eQTL and is differentially expressed in different regions of the brain [56, 57].

Poxviridae

The *Poxviridae* family consists of viruses that carry a linear ds-DNA genome of at least 100 kbp. We observed on average 5 poxvirus-like remnants in AD genomes compared to 4 in the CHC cohort. The AD pox-like elements spanned on average 1869 bp. CHC genomes contained pox-like elements of length 1481 bp on average. In contrast to previous discussed EVEs, poxvirus-like elements identified in our genomes all resembled non-conserved proteins. Visual inspection reveals that MC132 is the best conserved domain in length compared to the other EVEs. Ubiquitin is the only other protein that was found back almost in full length (Figure 11A).

Poxvirus EVEs have a weak link to AD

We observed two chromosomes for which we identified more AD samples having a poxvirus-like element compared to the number of CHC samples, namely chromosome 11 and 19 (Figure 7D). Only chromosome 11 showed a longer total EVE length for AD individuals compared to CHCs (p=1.52e-2; MWU, BH correction) (Figure 11B, 11C).

The closest gene to all AD enriched regions was NDUFC2 (ENSG00000151366) on chromosome 11. chr11:78085840-78086304 was the only region which had significantly different number of AD cases compared to CHC cases (Figure 11D chr 11 green arrows & Table 2). NDUFC2 is differentially expressed on the RNA and protein level in AD brains compared to healthy controls and the locus is a brain eQTL [56, 57].

For chromosome 19, gene set enrichment analysis on the nearby genes (UCA1, UBA52, ZNF714) for which AD had higher incidence of integration did not show any link to AD. UBA52 and ZNF714 are differentially expressed in AD brains [57]. Two regions were significantly enriched on the chromosome itself, namely chr19:18578415-18578915 and chr19:20865623-20866179 (Figure 11D green arrows & Table 2). The first region is 866 bp downstream of UBA52 (ENSG00000221983) (Figure S17). UBA52 has been identified as a potential biomarker for AD. Also, UBA52 expression plays a crucial role in protein folding and downregulation of UBA52 plays an important role in Parkinson's disease development [59]. The second region is intergenic and close

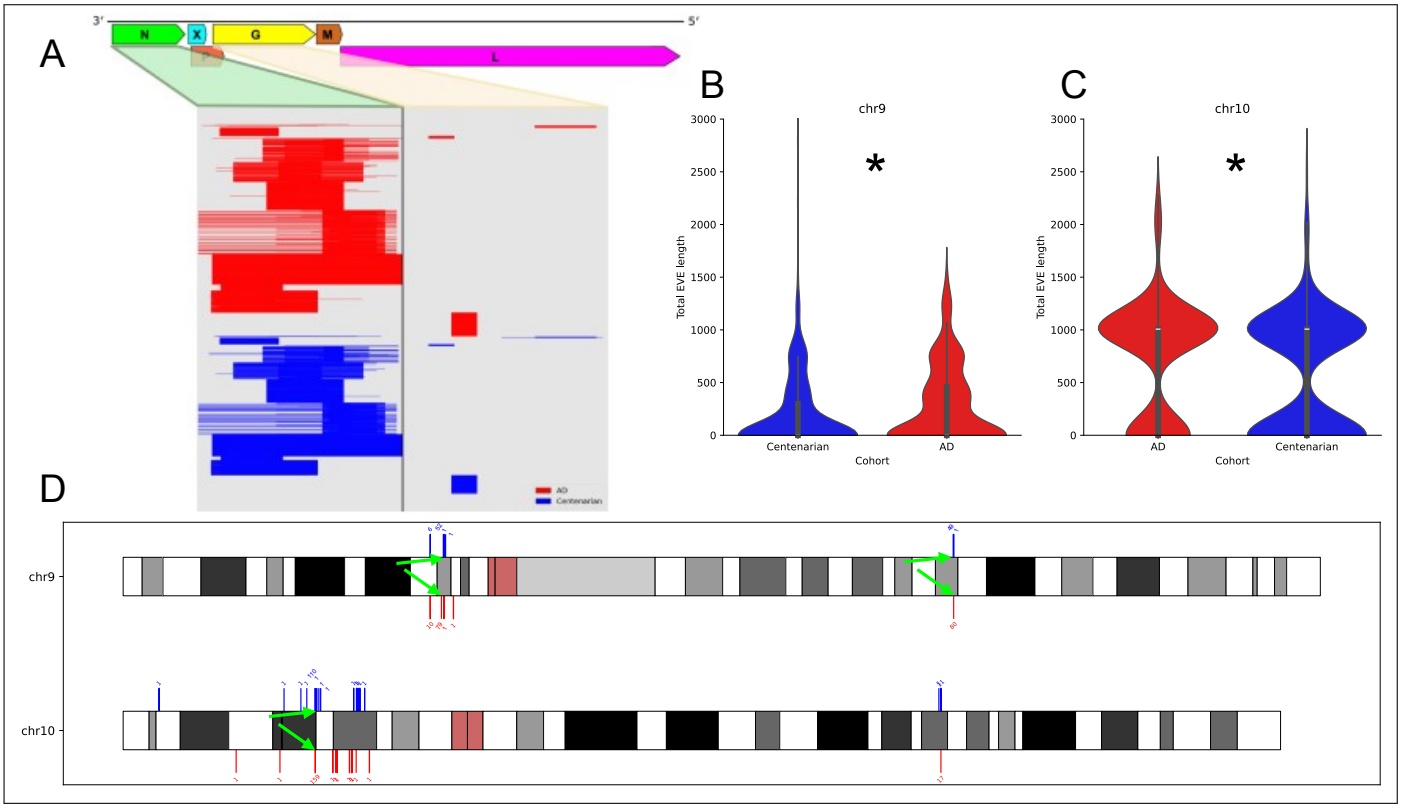

Fig. 10. **A) Coverage plot of the our recovered bornavirus-like EVEs. The conserved domains are shown where N is the nucleocapsid protein and G the glycoprotein. B/C) Density distribution of the total EVE lengths observed on chromosome 9 (B) and chromosome 10 (C) for the bornavirus EVEs. Black star indicate significant difference (MWU, BH correction). D) Chromosome locations of Bornavirus EVEs including number of samples that have an EVE in that region. Discussed (significantly) enriched regions are indicated in green.**

to ZNF66 (ENSG00000160229) (Figure S18). ZNF66 is differentially expressed in AD cases and part of the family of genes that were close to herpesvirus integrations [57]. Of note, this is the only region that has been shown to be enriched for the for CHC cohort.

Retroviridae

The *Retroviridae* family is the most abundant virus family in our genomes, in line with what was found previously [10]. We observe that an AD genome on average has 1250 retroviral-like EVEs (±53) compared to 985 (±52) EVEs in CHC genomes. The identified EVEs in AD covered on average 1253766 bp. In CHCs, on the other hand, the identified EVEs covered on average 1021125 bp.

Although we did not observe differences in the proportions of cohorts having retrovirus-like integrations per chromosome because of their widespread abundance on almost all chromosomes (Figure 7D), we did find numerous sub-chromosomal regions which had significant higher number of retroviral EVE incidence for both AD or CHCs. We focused specifically on regions within 100 kbp upstream or downstream of SNPs associated with AD [27]. Three of these regions, on chromosome 6, 7, and 17 showed significant enrichment for the AD cohort and are discussed below (Figure 12E, green arrows & Table 2).

Retroviridae EVEs are more abundant near AD associated SNPs and have longer total EVE length

For all three chromosomes, total EVE length was longer in AD individuals (p_{chr6}=4.81e-3, p_{chr7}=4.96e-3, p_{chr17}=p=4.70e-3 MWU, BH correction).

On chromosome 6 and 7, the remnants of these retrovirus integrations mostly resembled the conserved pol region of the retrovirus genome (Figure 12A). The region on chromosome 6 corresponds to an intergenic region (Figure S19). The closest gene to this region is HLA-DRB1 (ENSG00000196126). The closest SNP is inside the HLA-DQA1 gene at position 32615322.

The region on chromosome 7 corresponds to an intergenic region with the closest gene being VWDE (ENSG00000146530.15) approximately 36 kbp downstream of our region (Figure S20). The closest SNP is inside TMEM106B at position 12229967.

The region on chromosome 17 contains more remnants similar to the envelope protein of the retrovirus compared to the other two previously discussed regions (Figure 12). The region overlaps with the out-of-frame exon of the B4GALNT2 gene. The closest SNP is inside the ABI3 gene on position 49219935.

Lastly, a gene set enrichment analysis was performed on all regions throughout the genome for which we found significant AD enrichment that had an FE odds ratio of 10 or higher, resulting in a list of 77 genes (Supplementary section .1). Weak signal was found for pathways related to neuron migration and axonal development (Figure S22 & S23).

DISCUSSION

Ever since the first discovery of virus DNA inside the brains of AD patients, the so-called 'infectious hypothesis' of AD has been under heavy dispute. True or not, the fact remains that virus infections have influenced and shaped our genome throughout human history. Evidence of these influences can be traced back by investigating our DNA for subsequences similar to virus

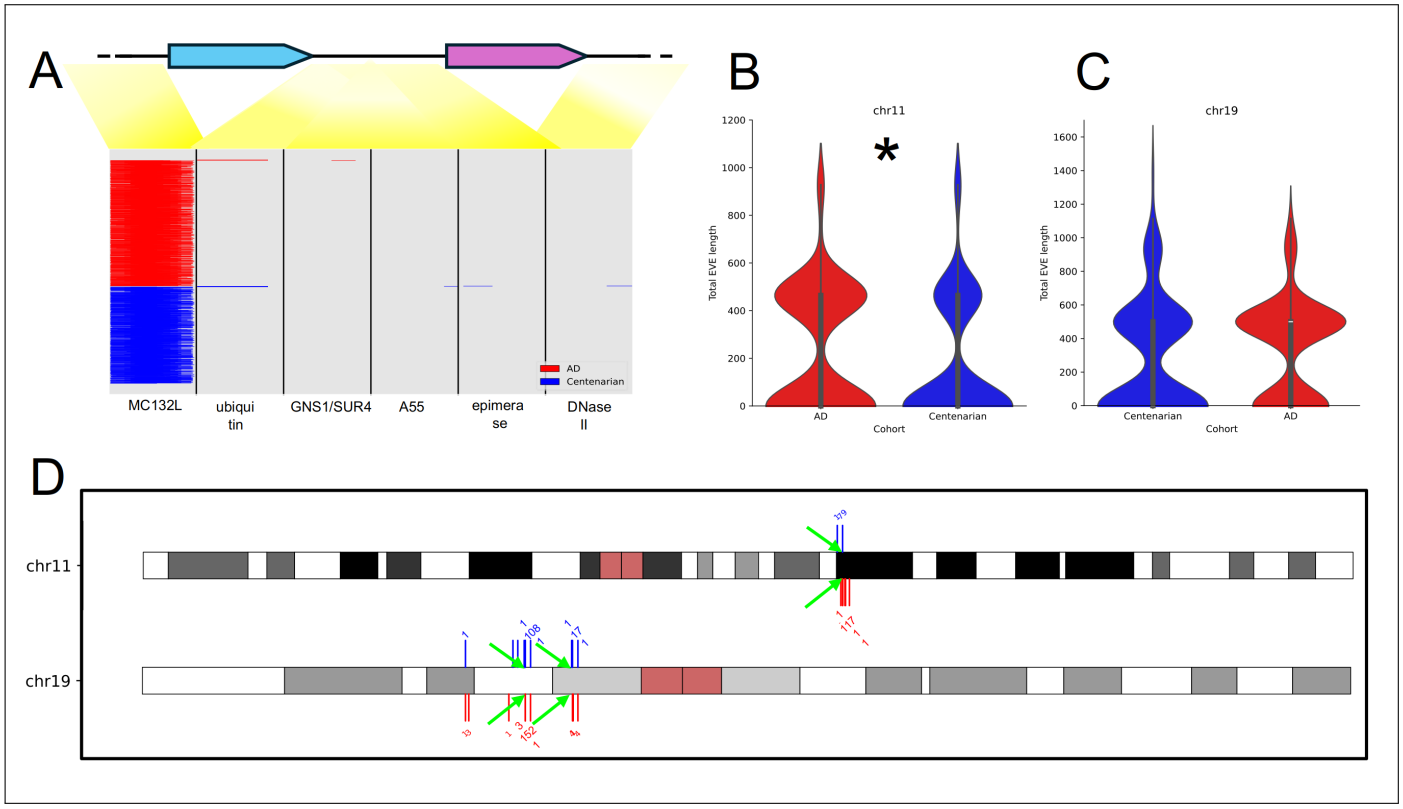

Fig. 11. **A) Coverage plot of the our recovered poxvirus-like EVEs. All found EVEs were similar to proteins from non-conserved domains as indicated by the yellow gradient between the conserved open reading frames. Protein names per region are annotated below every region B/C) Total EVE length distributions for the poxvirus EVEs on chromosome 11 and 19 in B and C, respectively. Star indicates significant difference between the distributions (MWU, BH corrected). D) Chromosome locations of EVEs including the number of samples from each cohort that have an EVE in the specified regions. Significantly enriched regions are indicated with green arrows. Note, chromosome 19 has been scaled for visualization purposes.**

proteins. Assuming that viruses might be an environmental contributory factor for the onset of AD, we hypothesized that traces of these virus integrations are to be found back in the genomes of AD patients. We therefore provide a new and unique perspective on the infectious hypothesis of AD by screening the genomes of AD patients and healthy CHCs for remnants of virus proteins from past infections.

In order to do so, we developed a BLASTx alignment based pipeline to screen, for the first time to the best of the authors' knowledge, human assemblies constructed from PacBio long-read sequencing data against all known vertebrate viruses in the Refseq NCBI virus protein database [33].

We observed that the genomes of the AD cohort contained more EVEs than CHC genomes in general and that a higher fraction of people with AD carry an EVE of one of the three non-retroviral families, *Orthoherpes-*, *Borna-*, and *Poxviridae* on specific chromosomal locations compared to the CHC control group. In addition, for most of the chromosomes that we discussed, the total EVE length was longer in the AD cohort than the CHC cohort.

Specifically for the herpesvirus family, we find that the EVEs tend to be close to genes that are differentially expressed in the brains of AD patients compared to healthy control groups. Most notably, we also observed relatively more AD samples to have herpesvirus EVEs on chromosome 19 located nearby ZNF genes. Gene set enrichment analysis showed that these genes are involved in the "Herpes Simplex Virus 1 infection" pathway (KEGG:05168) (Figure S12). Zooming in, we observed that these proteins play a role in the activity of Zinc Finger Antiviral Protein according to the KEGG database (Figure S24) [60]. ZAP is an important protein for the cell's intrinsic immune response [61]. Our data thus indicates that longer herpesvirus EVEs nearby these antiviral genes might hinder the individual's immune response to virus infections.

A similar but weaker link was found for *Pox-* and *Bornaviridae* where we saw sporadically that genes close these EVEs had different expression levels in AD brains. Also, we observed that the remnants of the *Retroviridae* family were enriched for the AD cohort in regions close to SNPs associated with AD. EVEs present in the genomes of Alzheimer's disease AD patients possibly influence the expression of nearby genes, and potentially even the genes in which SNPs associated with AD are located.

Together, these results are indicative in favor of the infectious hypothesis by shedding light on the infectious history of the ancestors of both groups.

Interestingly, we observed a low number of regions for which the CHC cohort had a higher incidence of samples compared to the AD cohort but these were not different significantly except for one poxvirus EVE on chromosome 19 (Figure 11D). This could hint to the phenomenon that some EVEs might be beneficial for ageing.

Our results should be interpreted with care, however. First of all, we observe a higher prevalence of herpesvirus remnants than most studies do, and suspect that false positive results are

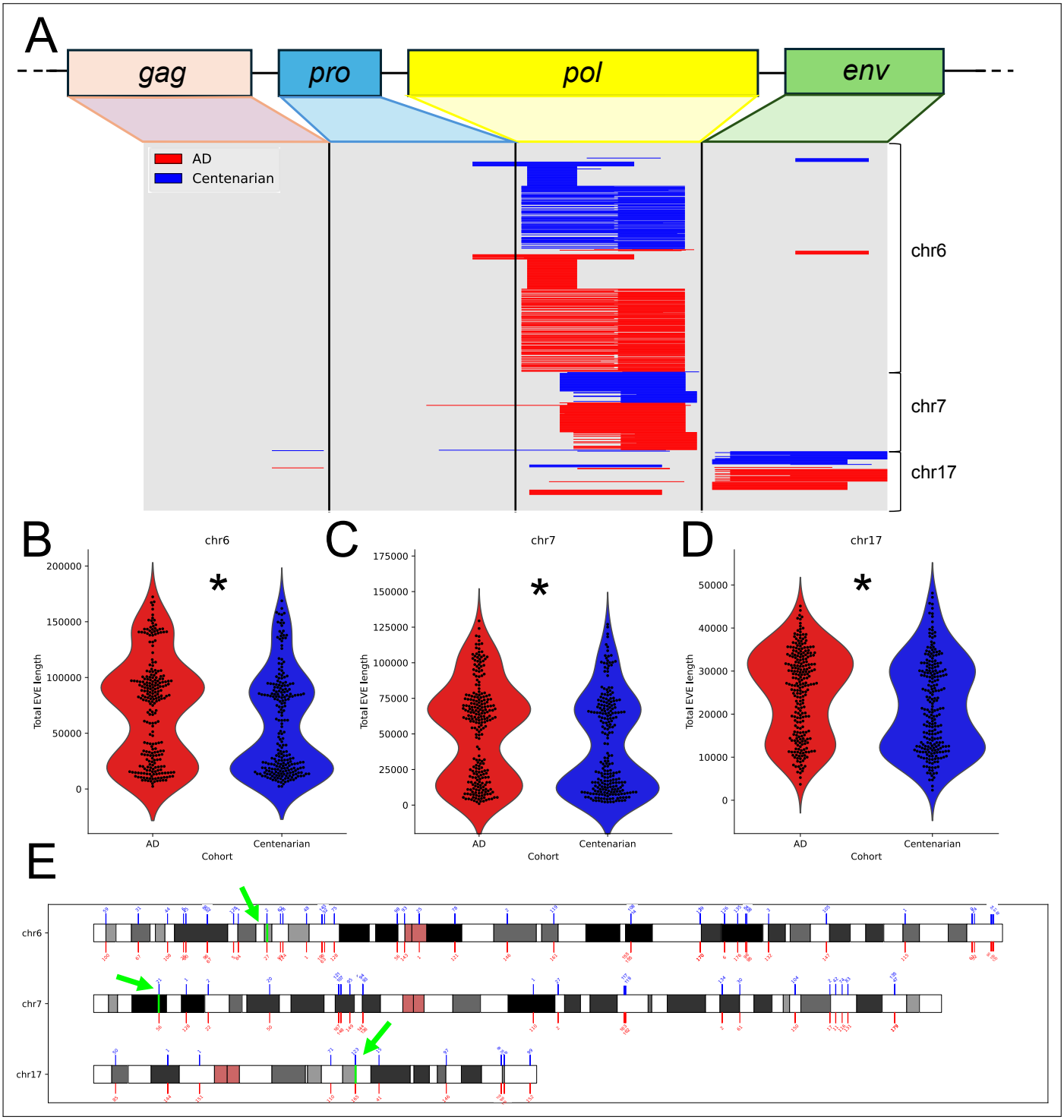

Fig. 12. **A) Retroviral genome recovery of the three chromosomal regions discussed. B/C/D) Total EVE length distribution for chromosome 6 (B), 7 (C), and 17 (D). Stars indicate significance (MWU, BH correction). E) EVE sights on chromosome 6, 7, and 17 including the number of samples that were found to have an EVE in the specific region. SNP interval of 200 kbp is indicated in green and pointed at with green arrows.**

still present. For example, HHV-6 remnants are thought to be detected in approximately 1% of the human population and mostly in the telomeres of chromosomes [28, 62]. Our findings mostly showcase sequence similarity to vertebrate non-human envelope protein of the Chelonid α-herpesvirus as the most similar virus protein to a specific human subsequence. One interpretation of these results is that infections of viruses similar to this herpesvirus happened. But, we also know that envelope proteins are a conserved type of protein in the virus kingdom, hinting at the fact that the remnant stems from a different virus family. One of those might be retroviruses since they are much more present in the human genome compared to non-retroviral viruses. Further validation would help answering this question. A validation technique that could be added to our pipeline

would extracting the subsequence in the genome and use a sequence classifier like RepeatMasker to validate our findings [41]. Leaving the exact viral origin aside, we believe the subsequences to be structural variants of viral origin, retro- or non-retroviral.

Our pipeline detected previously discovered bornavirus-like elements in the human genome [18, 38, 63]. Interestingly, we did not observe these remnants to be present in all of our genomes (353 out of 489 carried bornavirus-like remnants). This is in contrast to previous research that suggests these remnants to be ubiquitously present. We suspect to have observed heterogeneity in the sequence composition of these bornavirus-like elements causing our pipeline to not always detect these sequences as some bornavirus-like elements have stronger similarity to the virus proteins than others. Extraction of the subsequences at the genomic locations of previously discovered bornavirus-like elements and further downstream analysis can supplement our analysis and validate the differences we observe.

Lastly, the findings of *Poxviridae* are new to the field of EVEs but need careful validation before jumping into the conclusion. Members of the *Poxviridae* family replicate inside the cytosol and thus chances of integrations are very low [1]. We emphasize here that the subsequences that we report were queried against all virus sequences in the non-redundant Refseq database. The *Poxviridae* proteins resembled these subsequence the best based on lowest e-value and we observe differences in incidence between the cohorts. Previously discovered EVEs were also thought to not integrate at all but were identified nevertheless [7]. It would be interesting to do a similar experiment on vertebrate genomes other than human genomes to see if a poxvirus like sequence can be identified as well. Findings of orthologous sequences would then be a strong indication of valid findings.

Another opportunity for deeper analysis would be to include open reading frame extraction and poly-A tail inference in our pipeline as is mostly done in other EVE research [7, 38]. Hence, another filtering step that could be applied would be to use tools like BlastAlign to infer putative open reading frames and extract sequences around an EVE to explore poly-A tail presence [64]. Moreover, one could use the expressed sequence tag database and perform a BLASTn experiment on the extracted EVEs as was done in [16] to infer whether EVEs are expressed on the RNA level. These steps would also enhance the interpretation of the influence of these EVEs on increased AD onset risk.

Additionally, most virus signal was found to be inside LTR repetitive elements according to the GRChg38 chromosomal locations (see Supplementary section 1). While this is not unexpected, given the mode of non-retroviral integration, it could also indicate retrotransposal or retroviral origin [6]. Removal of (retro-)transposable elements was performed only based on previously discovered retrotransposons rather than classifying these regions in our genomes ourselves to save computational resources. Now that we have the regions of interest in place, classification of these subsequences with repetitive element classifiers could shed more light on the origin of these sequences. Leaving the discussion about the origin in the middle for now, we observe differences in structural variant composition between the two cohorts at least, which is important to note on itself.

Besides validation of our findings, a follow-up experiment would be to look at the contrasting perspective, exogenous virus detection. To the best of the authors' knowledge, the infectious hypothesis has only been investigated by detecting viral genomic sequences exogenously, i.e. outside of the host's genome. The general trend in these findings is that virus signal is increased in AD samples compared to the healthy control group [8, 22, 26]. Given the high quality raw sequencing data in blood and brains of our two cohorts, it would be interesting to see if a similar signal can be detected as previously discovered.

Finally, we note that we are highly biased to the Dutch samples. Stronger support for the infectious hypothesis would be created if this study could be performed on data from various populations across the globe.

Nevertheless, our findings underscore the diversity and differences in the EVE landscape of individuals, even within the Dutch population. Our genomes can be viewed as documents describing our current phenotype and archiving the virus interactions we have had in the past.

In conclusion, this research, for the first time, reports differences in identified EVEs in the AD and CHC genomes, shedding new light on the infectious hypothesis and providing observations in favor of it. The study highlights the importance of preventative measures against infectious diseases, demonstrating not only short-term benefits but also long-term benefits for ourselves and possibly for our descendants.

Acknowledgments. The authors would like to thank Dr. Salazar and Dr. Hulsman for their constructive feedback and valuable insights. Furthermore, we thank both the Bioinformatics lab and the 100+ study for technical support and data sharing.

Disclosures. The authors declare no conflicts of interest. All non-referenced figures were created through biorender.com or are taken, adapted, and cited accordingly. The results published here are in part based on data obtained from Agora, a platform initially developed by the NIA-funded AMP-AD consortium that shares evidence in support of AD target discovery. Agora is available at: doi:10.57718/agora-adknowledgeportal.

Source code and Data availability. Due to privacy reasons, data used cannot be disclosed. The source code used is available at https://github.com/MaEduard/master_thesis.

FULL REFERENCES

1. N. H. Acheson, *Fundamentals of molecular virology* (John Wiley & Sons, 2011).
2. A. Mushegian, "Are there 1031 virus particles on earth, or more, or fewer?" J. bacteriology **202**, 10–1128 (2020).
3. L. Cai, M. G. Weinbauer, L. Xie, and R. Zhang, "The smallest in the deepest: the enigmatic role of viruses in the deep biosphere," National Sci. Rev. **10**, nwad009 (2023).
4. B. Alberts, "Pathogens and infection," in *Molecular biology of the cell,* (Garland science, New York, 2017), chap. 23, pp. 1263–1296.
5. A. Aswad and A. Katzourakis, "Paleovirology: the study of endogenous viral elements," Virus evolution: current research future directions pp. 273–292 (2016).
6. A. Katzourakis and R. J. Gifford, "Endogenous viral elements in animal genomes," PLoS genetics **6**, e1001191 (2010).
7. A. Katzourakis and M. Tristem, "Phylogeny of human endogenous and exogenous retroviruses," Retroviruses primate genome evolution **186**, 203 (2005).
8. T. Piekut, M. Hurła, N. Banaszek, *et al.*, "Infectious agents and alzheimer's disease," J. Integr. Neurosci. **21**, 73 (2022).
9. M. A. Martin, T. Bryan, S. Rasheed, and A. S. Khan, "Identification and cloning of endogenous retroviral sequences present in human dna." Proc. National Acad. Sci. **78**, 4892–4896 (1981).
10. L. Es, "Initial sequencing and analysis of the human genome," Nature **409**, 860–921 (2001).
11. M. Tristem, "Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database," J. virology **74**, 3715–3730 (2000).
12. A. Aswad and A. Katzourakis, "Paleovirology and virally derived immunity," Trends ecology & evolution **27**, 627–636 (2012).
13. F. P. Ryan, "Human endogenous retroviruses in health and disease: a symbiotic perspective," J. Royal Soc. Med. **97**, 560–565 (2004).
14. S. M. Rawn and J. C. Cross, "The evolution, regulation, and function of placenta-specific genes," Annu. review cell developmental biology **24**, 159–181 (2008).
15. C. Lavialle, G. Cornelis, A. Dupressoir, *et al.*, "Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation," Philos. Trans. Royal Soc. B: Biol. Sci. **368**, 20120507 (2013).
16. T. Ghosh, R. G. Almeida, C. Zhao, *et al.*, "A retroviral link to vertebrate myelination through retrotransposon-rna-mediated control of myelin gene expression," Cell **187**, 814–830 (2024).
17. M. Horie, T. Honda, Y. Suzuki, *et al.*, "Endogenous non-retroviral rna virus elements in mammalian genomes," Nature **463**, 84–87 (2010).
18. K. Fujino, M. Horie, T. Honda, *et al.*, "Inhibition of borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome," Proc. National Acad. Sci. **111**, 13175–13180 (2014).
19. D. B. Hristova, K. B. Lauer, and B. J. Ferguson, "Viral interactions with non-homologous end-joining: a game of hide-and-seek," J. Gen. Virol. **101**, 1133–1144 (2020).
20. T. Miyasaka, K. Oguma, and H. Sentsui, "Distribution and characteristics of bovine leukemia virus integration sites in the host genome at three different clinical stages of infection," Arch. virology **160**, 39–46 (2015).
21. P. Scheltens, B. De Strooper, M. Kivipelto, *et al.*, "Alzheimer's disease," The Lancet **397**, 1577–1590 (2021).
22. G. A. Jamieson, N. J. Maitland, G. K. Wilcock, *et al.*, "Latent herpes simplex virus type 1 in normal and alzheimer's disease brains," J. medical virology **33**, 224–227 (1991).
23. M. A. Wozniak, R. F. Itzhaki, S. J. Shipley, and C. B. Dobson, "Herpes simplex virus infection causes cellular β-amyloid accumulation and secretase upregulation," Neurosci. letters **429**, 95–100 (2007).
24. A. Zambrano, L. Solis, N. Salvadores, *et al.*, "Neuronal cytoskeletal dynamic modification and neurodegeneration induced by infection with herpes simplex virus type 1," J. Alzheimer's Dis. **14**, 259–269 (2008).
25. W. A. Eimer, D. K. V. Kumar, N. K. N. Shanmugam, *et al.*, "Alzheimer's disease-associated β-amyloid is rapidly seeded by herpesviridae to protect against brain infection," Neuron **99**, 56–63 (2018).
26. M. Wainberg, T. Luquez, D. M. Koelle, *et al.*, "The viral hypothesis: how herpesviruses may contribute to alzheimer's disease," Mol. psychiatry **26**, 5476–5480 (2021).
27. C. Bellenguez, F. Küçükali, I. E. Jansen, *et al.*, "New insights into the genetic etiology of alzheimer's disease and related dementias," Nat. genetics **54**, 412–436 (2022).
28. S. Kojima, A. J. Kamada, and N. F. Parrish, "Virus-derived variation in diverse human genomes," PLoS Genet. **17**, e1009324 (2021).
29. Z. J. Whitfield, P. T. Dolan, M. Kunitomi, *et al.*, "The diversity, structure, and function of heritable adaptive immunity sequences in the aedes aegypti genome," Curr. Biol. **27**, 3511–3519 (2017).
30. T. Hu, N. Chitnis, D. Monos, and A. Dinh, "Next-generation sequencing technologies: An overview," Hum. Immunol. **82**, 801–811 (2021).
31. H. Bowles, R. Kabiljo, A. Al Khleifat, *et al.*, "An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data," Front. bioinformatics **2**, 1062328 (2023).
32. H. Holstege, N. Beker, T. Dijkstra, *et al.*, "The 100-plus study of cognitively healthy centenarians: rationale, design and cohort description," Eur. J. Epidemiol. **33**, 1229–1249 (2018).
33. N. A. O'Leary, M. W. Wright, J. R. Brister, *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," Nucleic acids research **44**, D733–D745 (2016).
34. W. M. Van Der Flier and P. Scheltens, "Amsterdam dementia cohort: performing research to optimize care," J. Alzheimer's Dis. **62**, 1091–1111 (2018).
35. A. Salazar, N. Tesi, M. Hulsman, *et al.*, "An aluyb8 mobile element further characterises a risk haplotype of tmem106b associated in neurodegeneration," (2023).
36. H. Cheng, G. T. Concepcion, X. Feng, *et al.*, "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm," Nat. methods **18**, 170–175 (2021).
37. U. Palatini, E. Pischedda, and M. Bonizzoni, "Computational methods for the discovery and annotation of viral integrations," in *piRNA: Methods and Protocols,* (Springer, 2022), pp. 293–313.
38. V. A. Belyi, A. J. Levine, and A. M. Skalka, "Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes," PLoS pathogens **6**, e1001030 (2010).
39. B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using diamond," Nat. methods **12**, 59–60 (2015).
40. A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," Bioinformatics **26**, 841–842 (2010).
41. A. Smit, R. Hubley, and P. Green, "Repeatmasker open-4.0. 2013–2015," (2015).
42. C. Llorens, B. Soriano, M. Krupovic, and I. R. Consortium, "Ictv virus taxonomy profile: Metaviridae," J. Gen. Virol. **101**, 1131–1132 (2020).
43. N. Tesi Sr, A. Salazar, Y. Zhang, *et al.*, "Characterising tandem repeat complexities across long-read sequencing platforms with treat," bioRxiv pp. 2024–03 (2024).
44. C. I. Bliss and R. A. Fisher, "Fitting the negative binomial distribution to biological data," Biometrics **9**, 176–200 (1953).
45. T. W. MacFarland, J. M. Yates, T. W. MacFarland, and J. M. Yates, "Mann–whitney u test," Introd. to nonparametric statistics for biological sciences using R pp. 103–132 (2016).
46. R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of p," J. royal statistical society **85**, 87–94 (1922).
47. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. Royal statistical society: series B (Methodological) **57**, 289–300 (1995).
48. J. D. Hunter, "Matplotlib: A 2D graphics environment," Comput. Sci. & Eng. **9**, 90–95 (2007).
49. M. L. Waskom, "seaborn: statistical data visualization," J. Open Source Softw. **6**, 3021 (2021).
50. L. Lopez-Delisle, L. Rabbani, J. Wolff, *et al.*, "pygenometracks: reproducible plots for multivariate genomic datasets," Bioinformatics **37**, 422–423 (2021).
51. J. Huerta-Cepas, F. Serra, and P. Bork, "Ete 3: reconstruction, analysis, and visualization of phylogenomic data," Mol. biology evolution **33**,

1635–1638 (2016).
52. P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, “Scipy 1.0: fundamental algorithms for scientific computing in python,” Nat. methods **17**, 261–272 (2020).
53. O. Abril-Pla, V. Andreani, C. Carroll, *et al.*, “Pymc: a modern, and comprehensive probabilistic programming framework in python,” PeerJ Comput. Sci. **9**, e1516 (2023).
54. U. Raudvere, L. Kolberg, I. Kuzmin, *et al.*, “g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” Nucleic acids research **47**, W191–W198 (2019).
55. D. Gatherer, D. P. Depledge, C. A. Hartley, *et al.*, “Ictv virus taxonomy profile: Herpesviridae 2021,” J. Gen. Virol. **102**, 001673 (2021).
56. S. K. Sieberts, T. M. Perumal, M. M. Carrasquillo, *et al.*, “Large eqtl meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions,” Sci. data **7**, 340 (2020).
57. A. K. Greenwood, J. Gockley, K. Daily, *et al.*, “Agora: an open platform for exploration of alzheimer’s disease evidence: genetics/omics and systems biology,” Alzheimer’s & Dementia **16**, e046129 (2020).
58.
59. S. Tiwari, A. Singh, P. Gupta, and S. Singh, “Uba52 is crucial in hsp90 ubiquitylation and neurodegenerative signaling during early phase of parkinson’s disease,” Cells **11**, 3770 (2022).
60. M. Kanehisa, M. Furumichi, Y. Sato, *et al.*, “Kegg for taxonomy-based analysis of pathways and genomes,” Nucleic acids research **51**, D587–D592 (2023).
61. K. Q. de Andrade and C. C. Cirne-Santos, “Antiviral activity of zinc finger antiviral protein (zap) in different virus families,” Pathogens **12**, 1461 (2023).
62. G. Morissette and L. Flamand, “Herpesviruses and chromosomal integration,” J. virology **84**, 12100–12109 (2010).
63. K. N. Myers, G. Barone, A. Ganesh, *et al.*, “The bornavirus-derived human protein ebln1 promotes efficient cell cycle transit, microtubule organisation and genome stability,” Sci. Reports **6**, 1–12 (2016).
64. R. Belshaw and A. Katzourakis, “Blastalign: a program that uses blast to align problematic nucleotide sequences,” Bioinformatics **21**, 122–123 (2005).

1. SUPPLEMENTARY

Gene Name	Link
RAB42	https://agora.adknowledgeportal.org/genes/ENSG00000188060/evidence/rna
ODF2L	https://agora.adknowledgeportal.org/genes/ENSG00000122417/evidence/rna
TGFBR3	https://agora.adknowledgeportal.org/genes/ENSG00000069702/evidence/rna
TRIM33	https://agora.adknowledgeportal.org/genes/ENSG00000197323/evidence/rna
RHEX	https://agora.adknowledgeportal.org/genes/ENSG00000263961/summary
STUM	https://agora.adknowledgeportal.org/genes/ENSG00000203685/evidence/rna
ZNF678	https://agora.adknowledgeportal.org/genes/ENSG00000181450/evidence/rna

Table S1. Gene names of nearby genes of the herpes-like EVEs on chromosome 1 and the corresponding links to the Agora knowledge hub for their statistically significant relationship to AD.

Gene Name	Link
BRAK-RBAKDN	https://agora.adknowledgeportal.org/genes/ENSG00000272968
OSBPL3	https://agora.adknowledgeportal.org/genes/ENSG00000070882
ANLN	https://agora.adknowledgeportal.org/genes/ENSG00000011426
CYP3A43	https://agora.adknowledgeportal.org/genes/ENSG00000021461
AKR1B1	https://agora.adknowledgeportal.org/genes/ENSG00000085662

Table S2. Gene names of nearby genes of the herpes-like EVEs on chromosome 7 and the corresponding links to the Agora knowledge hub for their statistically significant relationship to AD.

Gene Name	Link
SLFN12L	https://agora.adknowledgeportal.org/genes/ENSG00000205045
KRTAP9-7	https://agora.adknowledgeportal.org/genes/ENSG00000180386
KRT17	https://agora.adknowledgeportal.org/genes/ENSG00000128422/evidence/rna
RNF213-AS1	https://agora.adknowledgeportal.org/genes/ENSG00000263069
ACSF2	https://agora.adknowledgeportal.org/genes/ENSG00000167107

Table S3. Gene names of nearby genes of the herpes-like EVEs on chromosome 17 and the corresponding links to the Agora knowledge hub for their statistically significant relationship to AD.

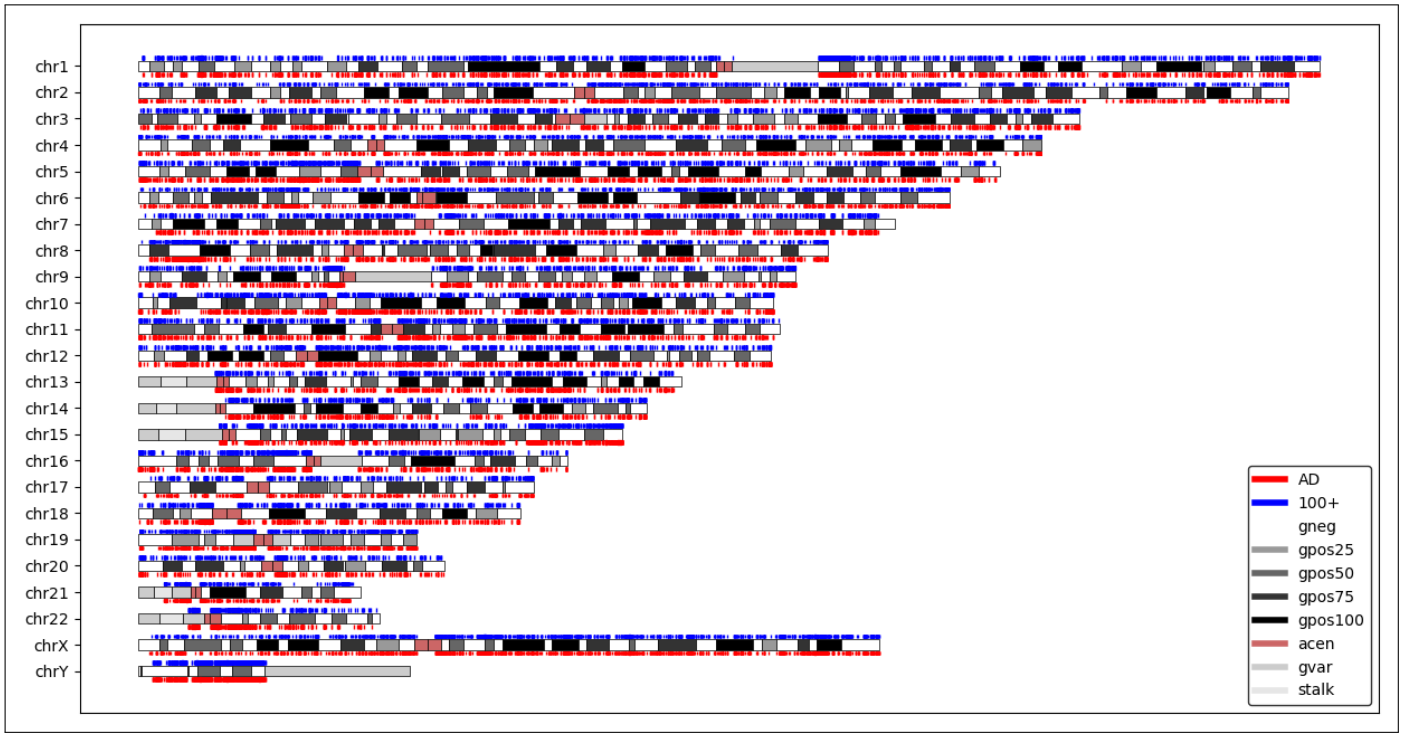

Fig. S1. All identified EVEs in the human genome of the AD and CHC cohort.

Fig. S2. EVE locations along human genome chromosomes.

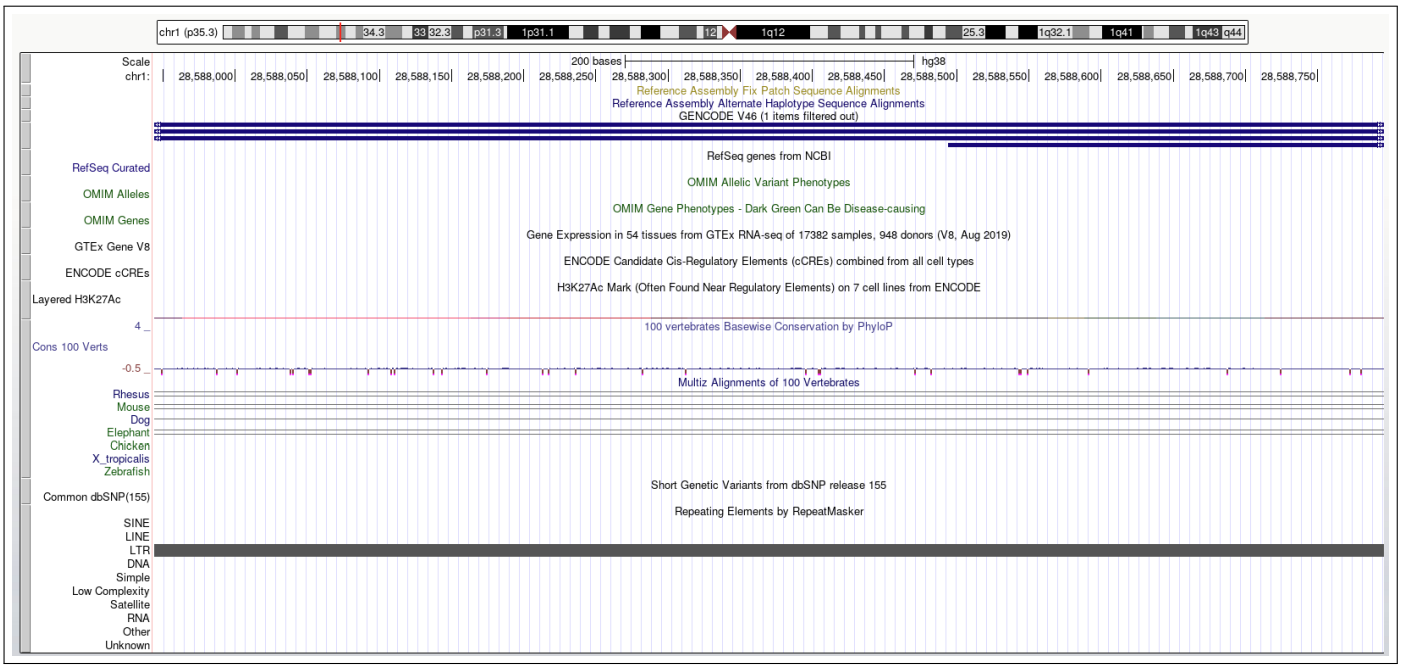

Fig. S3. **UCSC genome track of chr1:28,587,093-28,589,648**

Fig. S4. **gProfiler plot output, for nearby genes of enriched herpesvirus regions on chromosome 1**

Fig. S5. **gProfiler table output, for nearby genes of enriched herpesvirus regions on chromosome 1**

Gene Name	Link
ZNF93	https://agora.adknowledgeportal.org/genes/ENSG00000205045
znf682	https://agora.adknowledgeportal.org/genes/ENSG00000197124
ZNF626	https://agora.adknowledgeportal.org/genes/ENSG00000188171
ZNF98	https://agora.adknowledgeportal.org/genes/ENSG00000197360
NAPSA	https://agora.adknowledgeportal.org/genes/ENSG00000131400
ZNF813	https://agora.adknowledgeportal.org/genes/ENSG00000198346

Table S4. Gene names of nearby genes of the herpes-like EVEs on chromosome 19 and the corresponding links to the Agora knowledge hub for their statistically significant relationship to AD.

Gene Name	Link
ERCC6L2	https://agora.adknowledgeportal.org/genes/ENSG00000182150
EBLN3P	https://agora.adknowledgeportal.org/genes/ENSG00000281649
RUSC2	https://agora.adknowledgeportal.org/genes/ENSG00000198853

Table S5. Gene names of nearby genes of the borna-like EVEs on chromosome 9 and the corresponding links to the Agora knowledge hub for their statistically significant relationship to AD.

.1. *Nearby genes to Retroviridae EVEs that showed significant enrichment for AD and had odds ratio > 10*

AADACL3, ABI1, ADGRL2, AKAP10, ANGPT1, APOBEC1, ASCC3, ATG4C, BEX3, C14orf177, C6orf226, C8orf33, CCDC179, CDRT15P1, COL4A5, COMMD1, DANT2, DNAH11, EIF2AK2, FAM181B, FBXO17, FPR3, GIMAP4, GPR83, H2AC12, HDAC2-AS2, HLA-DRB6, HMGCS1, KCCAT333, KCNJ6, LINC01093, LINC01242, LINC01278, LINC01356, LINC01615, LINC01815, LINC02062, LINC02591, LOC100506990, LOC101927623, LOC644669, LOC646029, LOC646813, LOC729732, MEI4, NDN, NELL2, OR11H2, OR52K2, PCLO, PDE7A, PRELP, PRMT8, PSMC1, PTGDR, RAB39A, RBFA, ROBO1, RPS4Y2, RPSAP58, RSPO3, SETMAR, SLC25A44, SLIT2, SMIM19, SPIN3, SYNE1, NCOR1P4, TGIF2LX, TMCC1-AS1, TTTY9A, UFM1, XCR1, ZBBX, ZIM2, ZNF678, ZPLD1

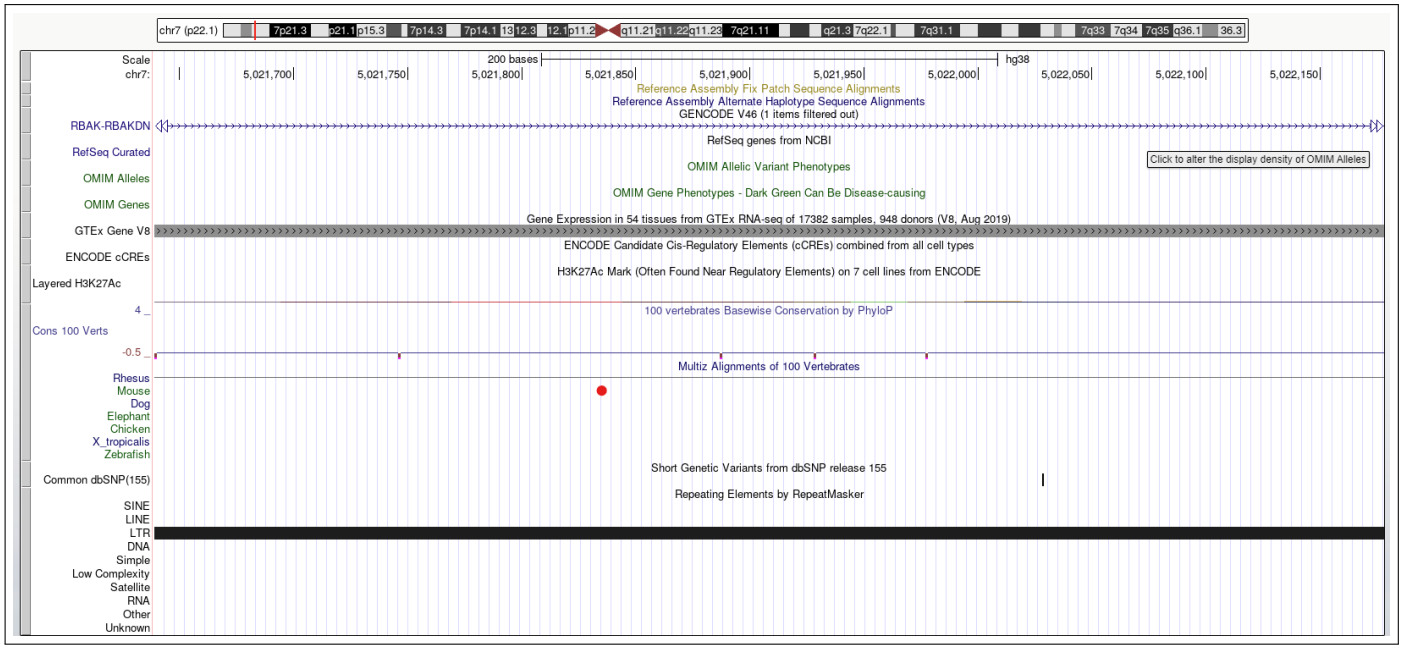

Fig. S6. **UCSC genome track of chr7:5,021,640-5,022,178**

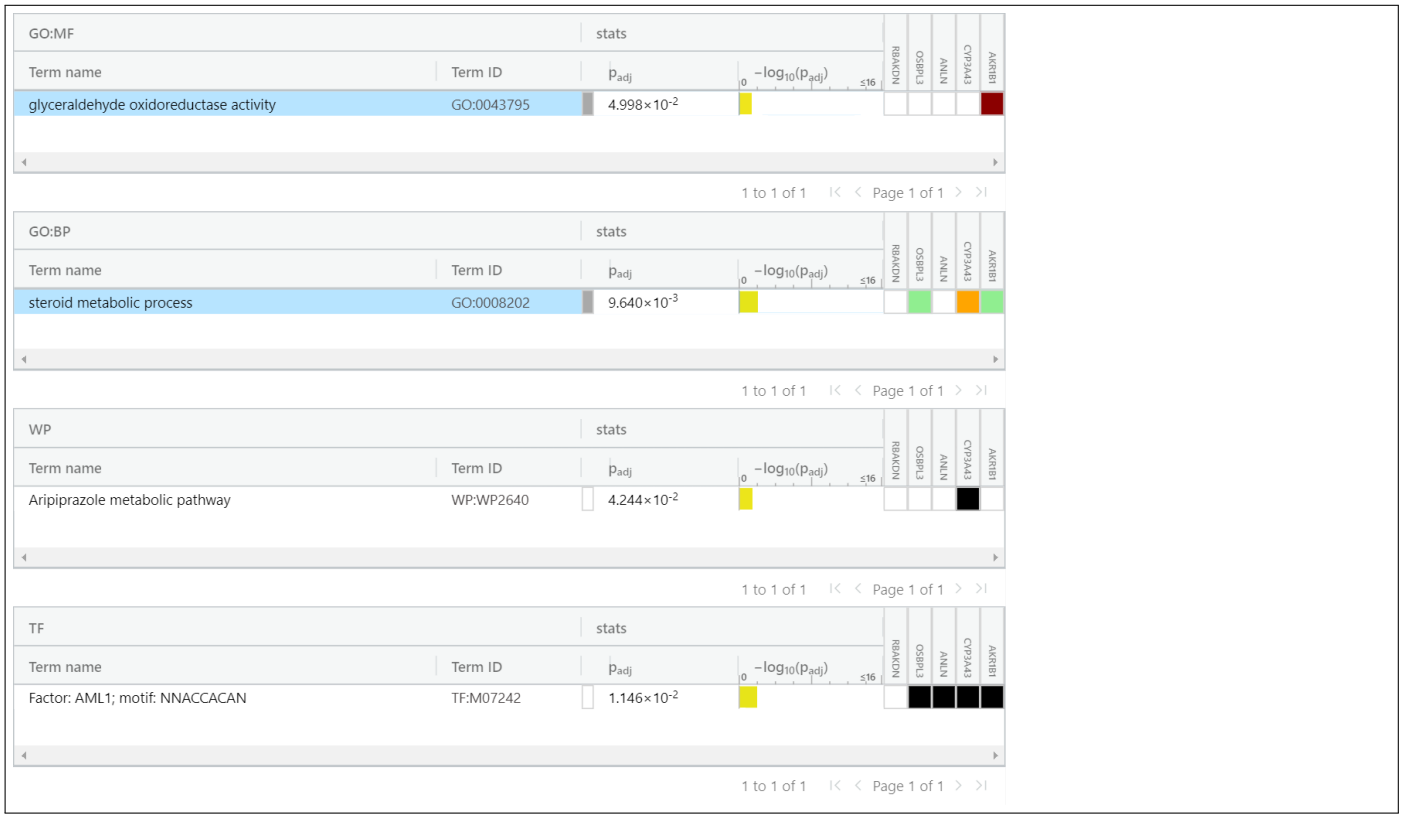

Fig. S7. **gProfiler table output, for nearby genes of enriched herpesvirus regions on chromosome 7**

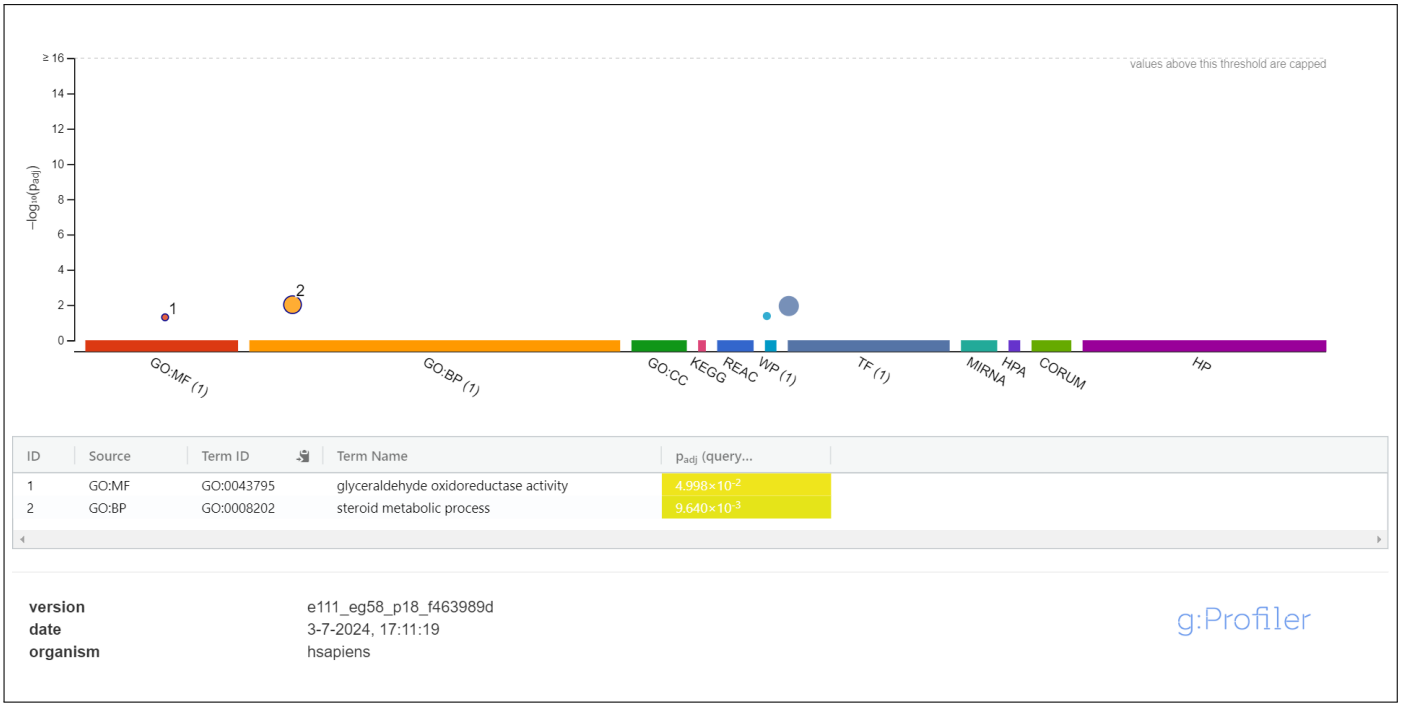

Fig. S8. gProfiler plot output, for nearby genes of enriched herpesvirus regions on chromosome 7

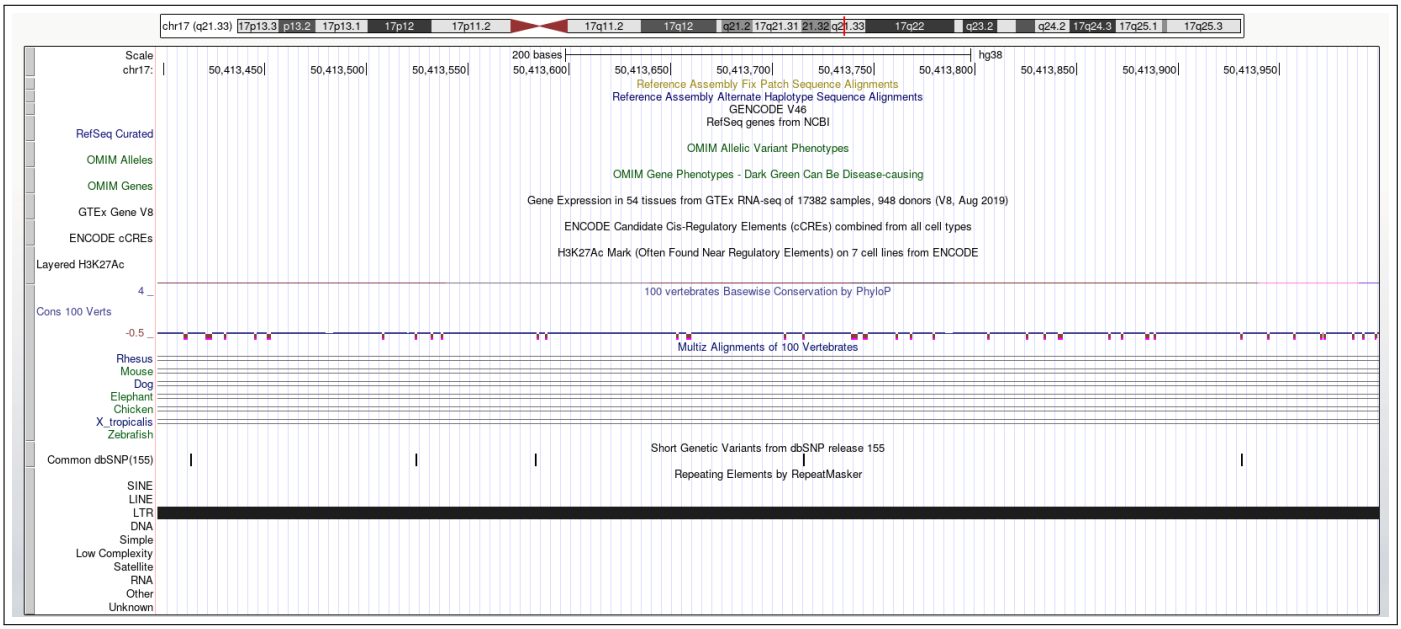

Fig. S9. UCSC genome track of chr17:50,413,398-50,413,999

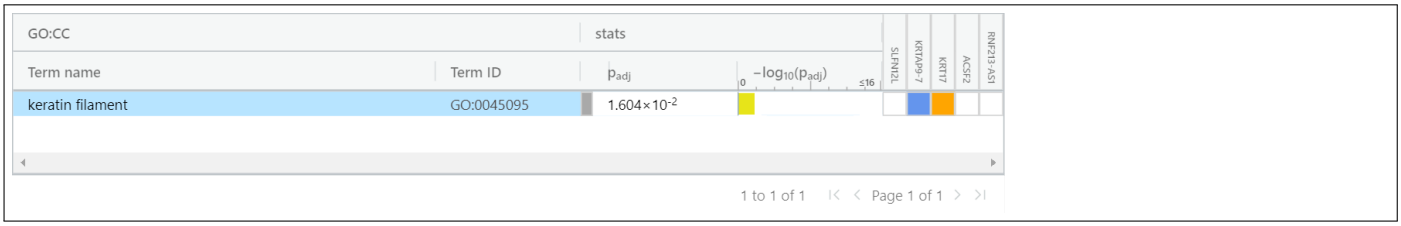

Fig. S10. gProfiler table output, for nearby genes of enriched herpesvirus regions on chromosome 17

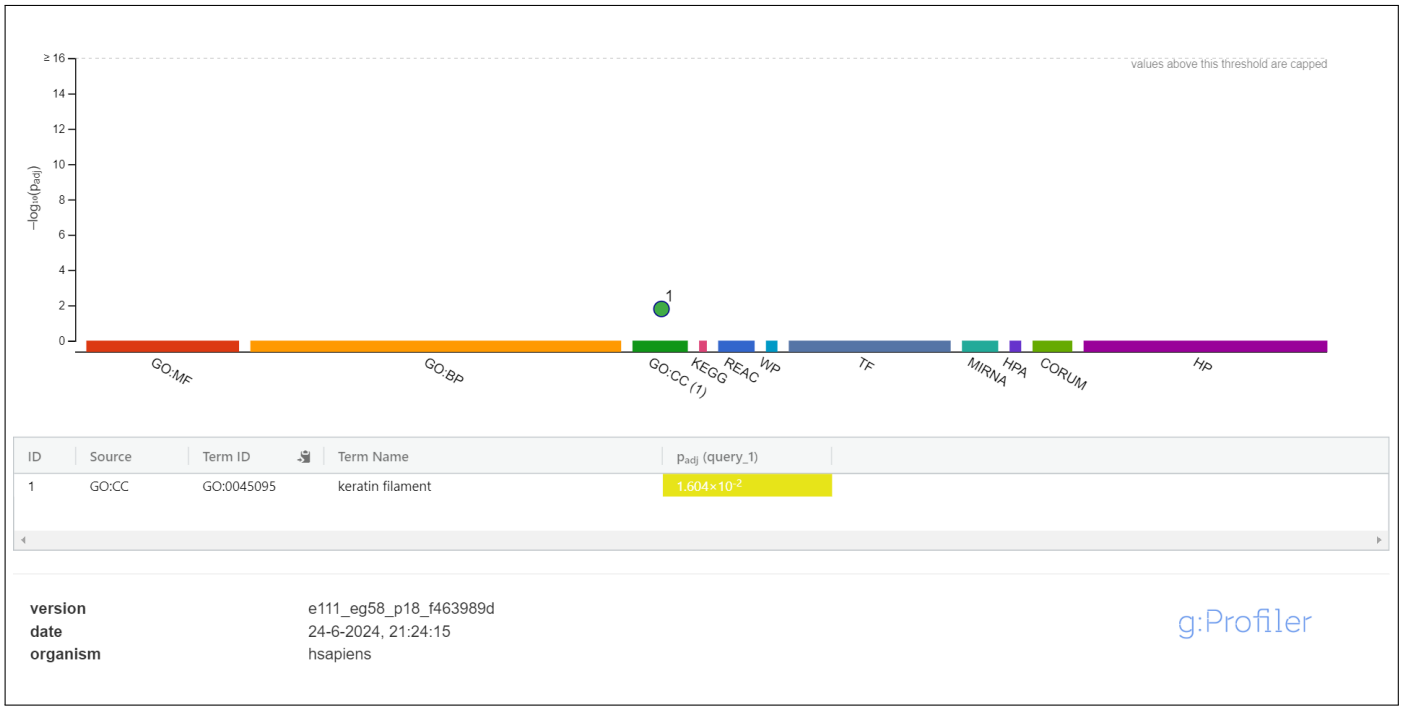

Fig. S11. **gProfiler plot output, for nearby genes of enriched herpesvirus regions on chromosome 17**

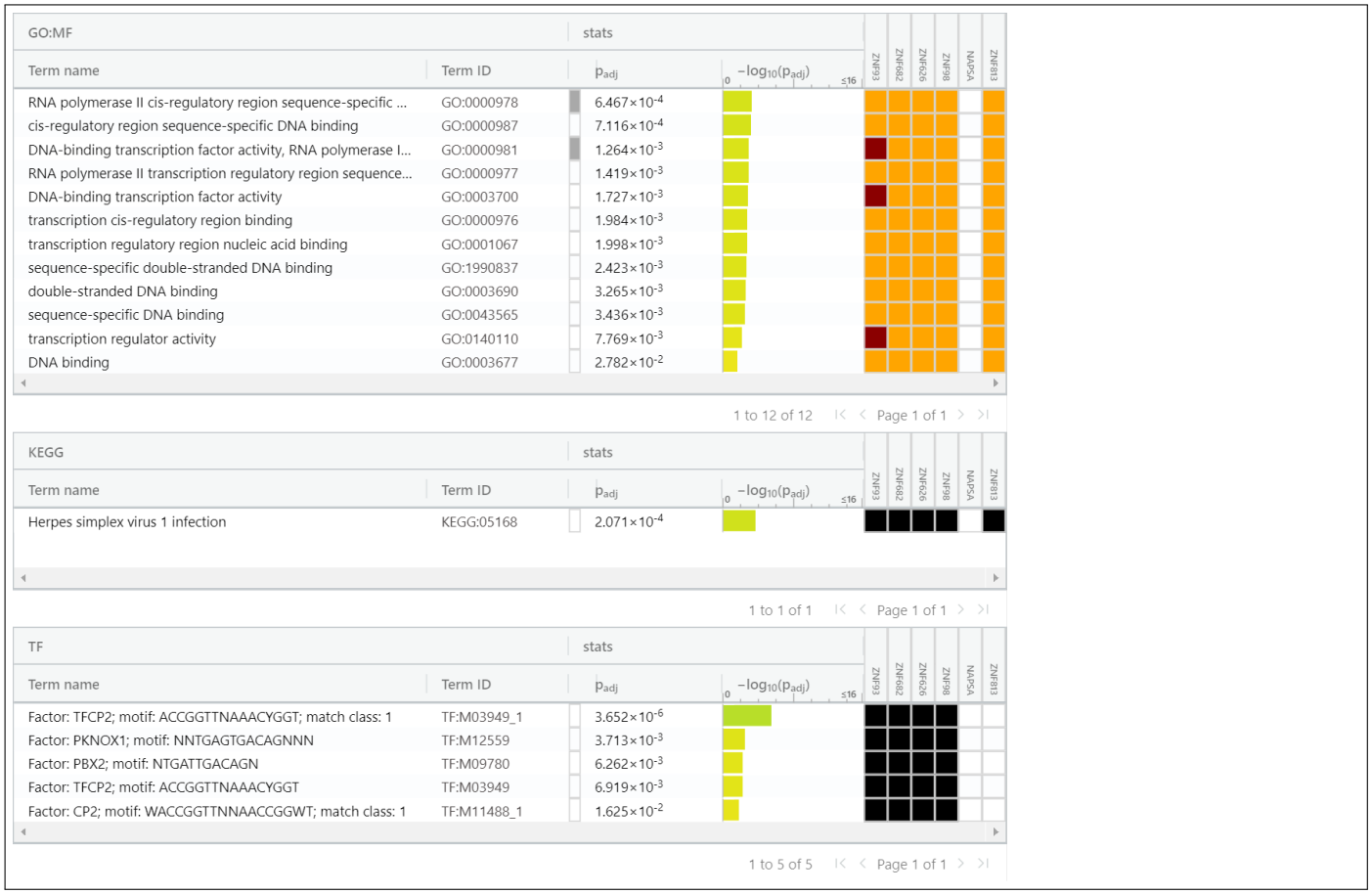

Fig. S12. **gProfiler table output, for nearby genes of enriched herpesvirus regions on chromosome 19**

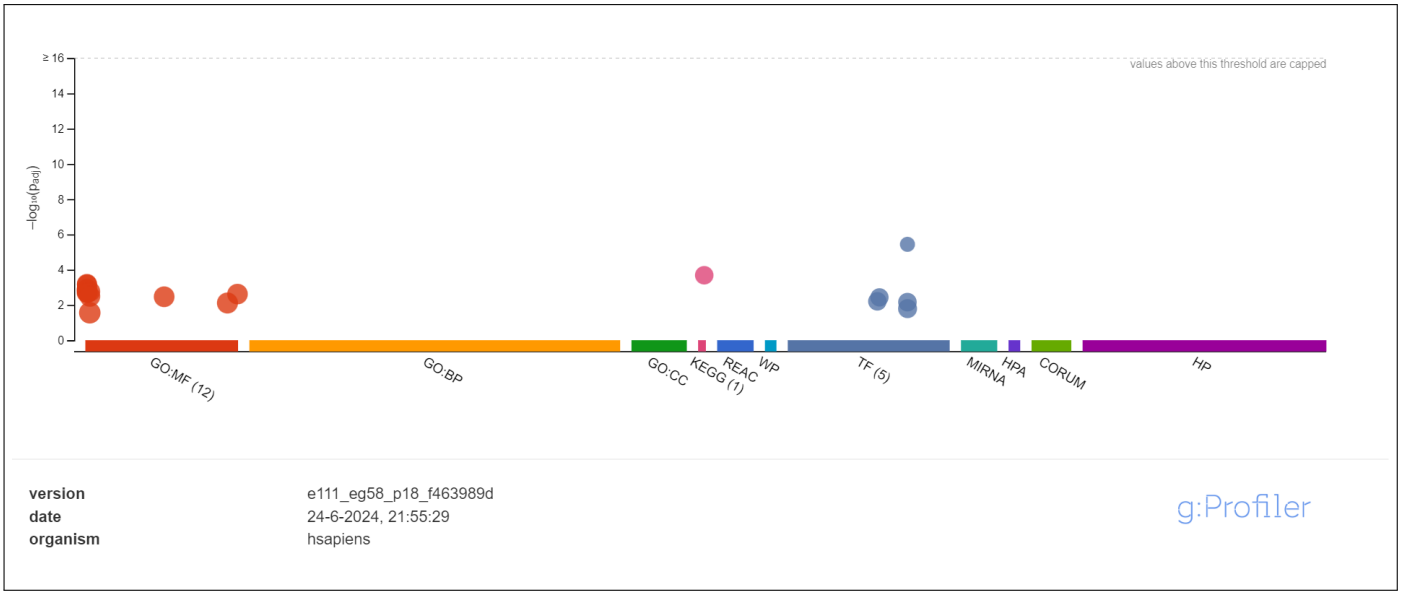

Fig. S13. gProfiler plot output, for nearby genes of enriched herpesvirus regions on chromosome 19

Fig. S14. UCSC's genome broswer track of chr9:37086819-37087689

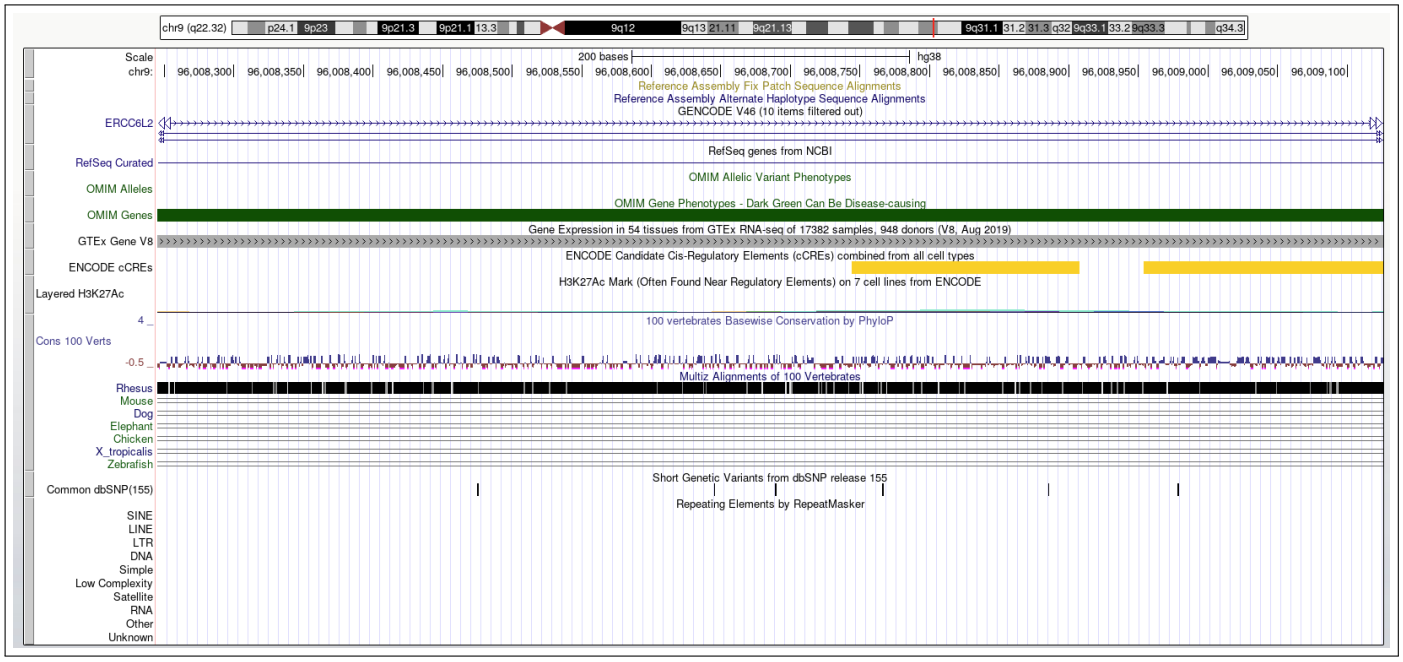

Fig. S15. UCSC's genome broswer track of chr9:96008246-96009126

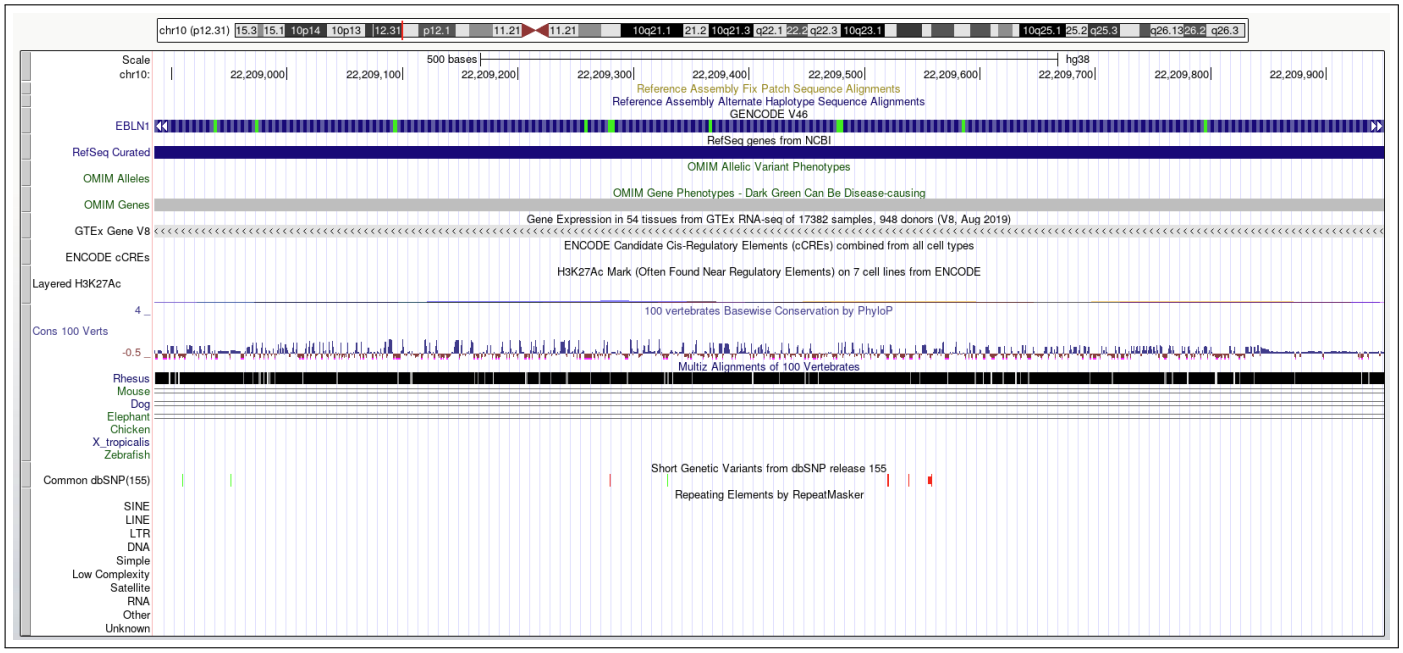

Fig. S16. UCSC's genome broswer track of chr10:22208886-22209950

Fig. S17. UCSC's genome broswer track of chr19:18578415-18578915

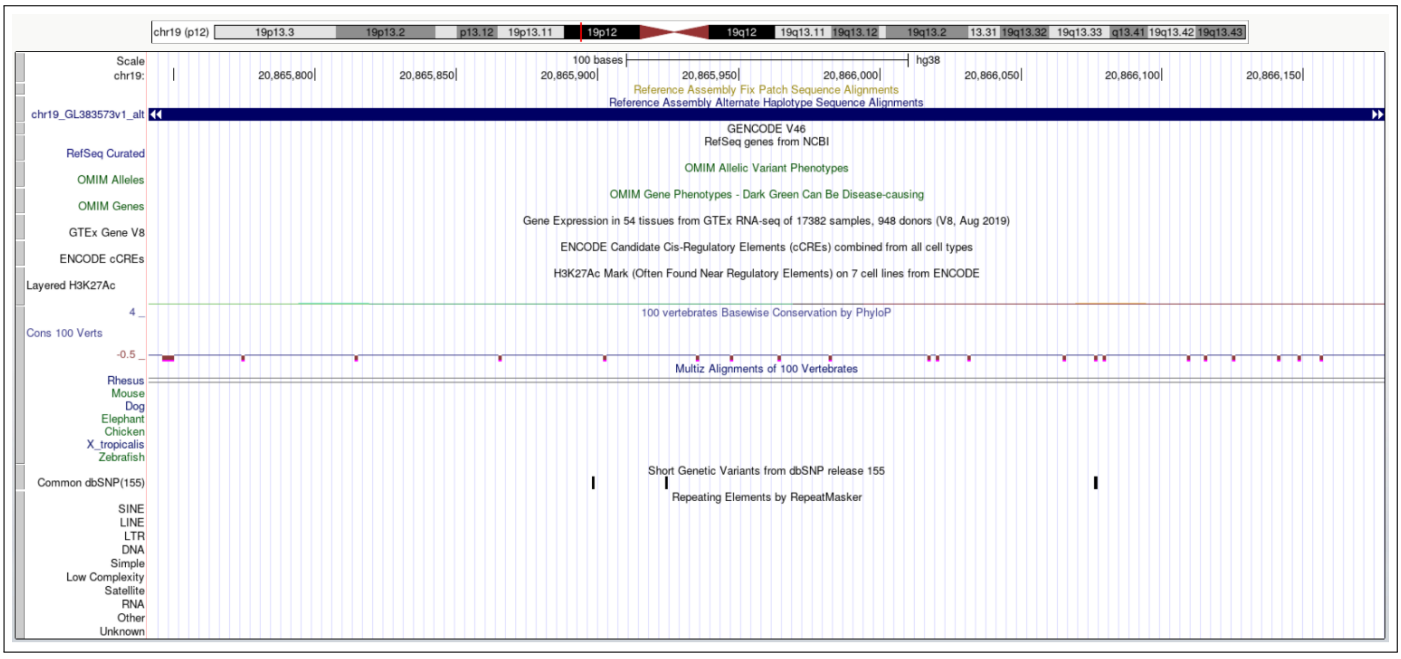

Fig. S18. UCSC's genome broswer track of chr19:20865623-20866179

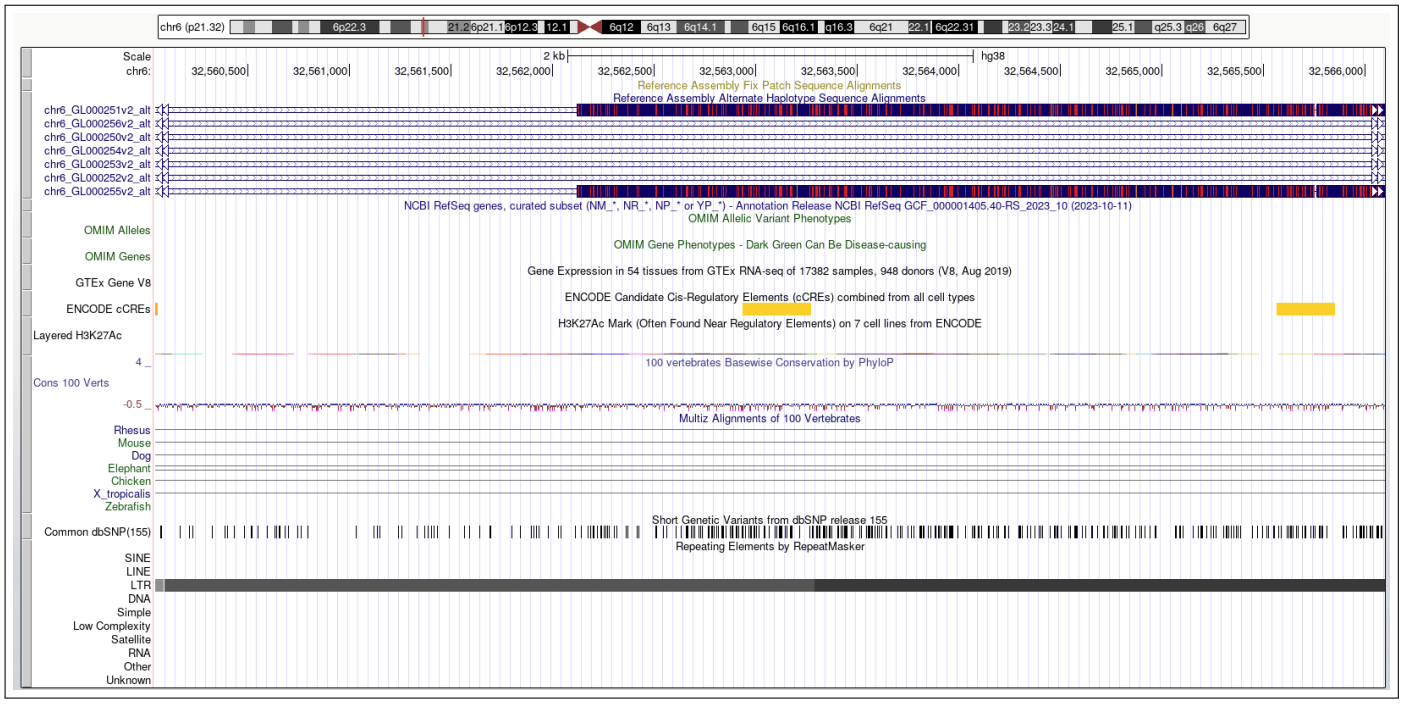

Fig. S19. UCSC's genome broswer track of chr6:32560043-32566098

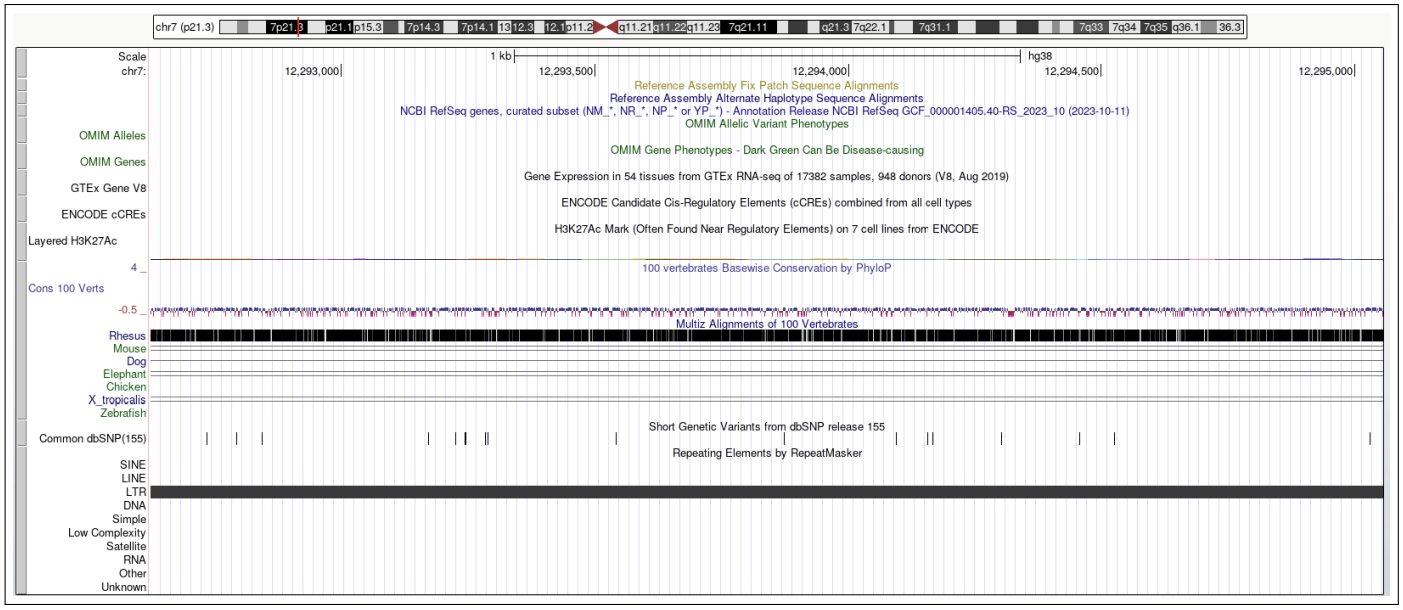

Fig. S20. UCSC's genome broswer track of chr7:12292625-12295056

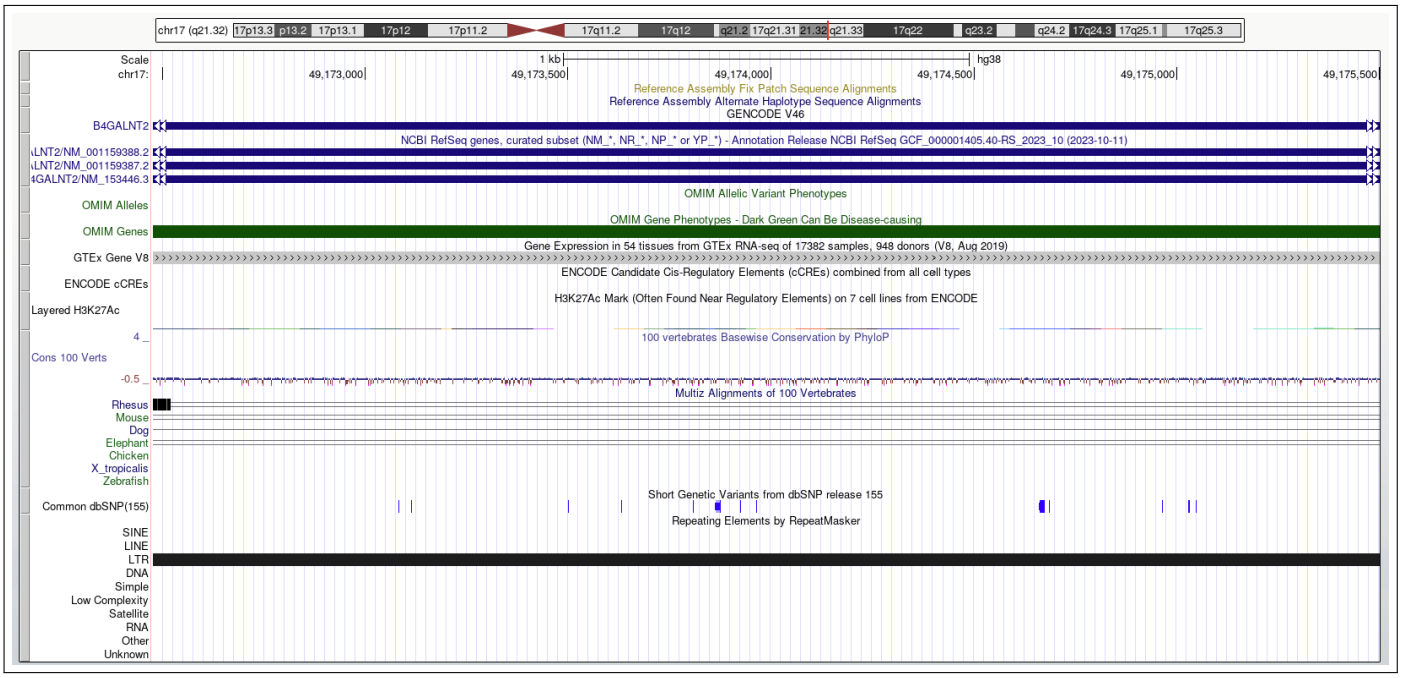

Fig. S21. **UCSC's genome broswer track of chr17:49172477-49175501**

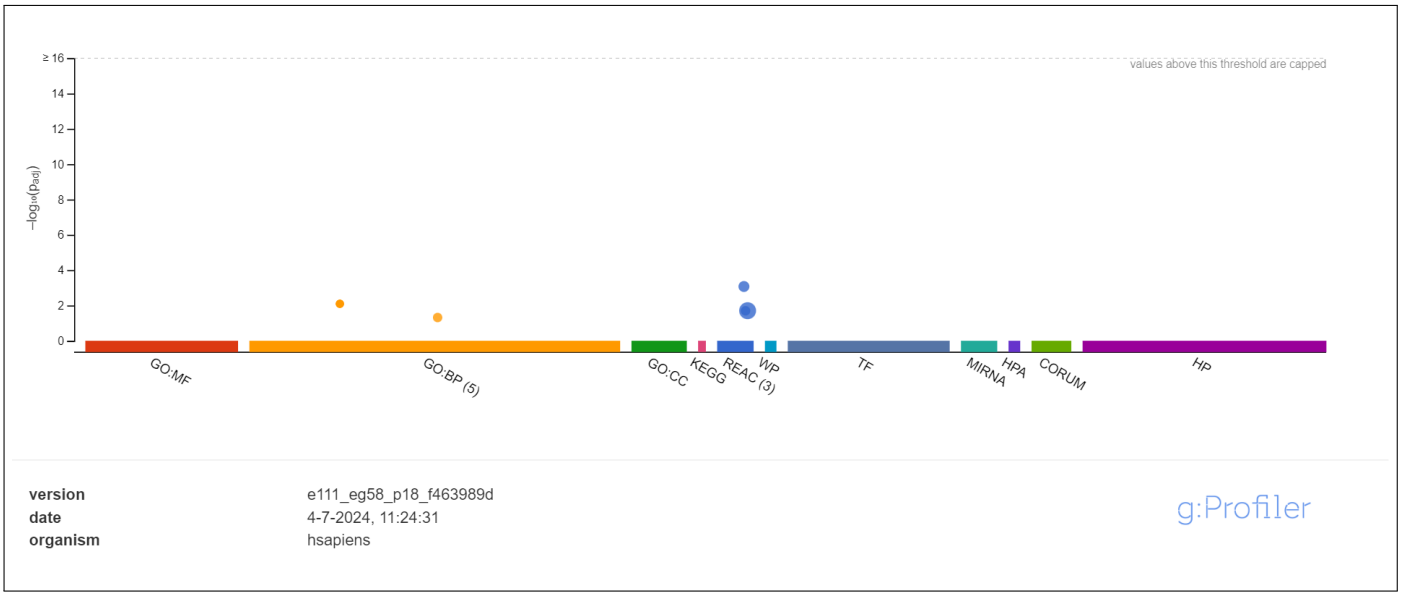

Fig. S22. **gprofiler graph of *Retroviridae* enriched genes**

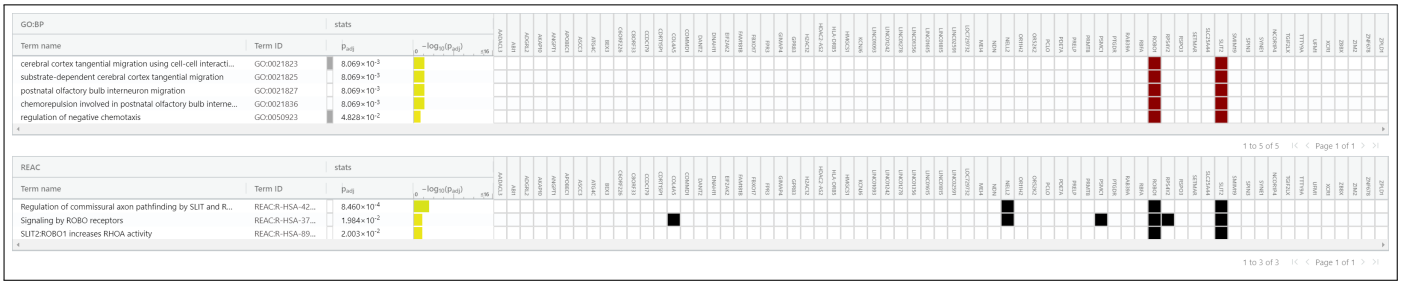

Fig. S23. **gprofiler table of *Retroviridae* enriched genes**

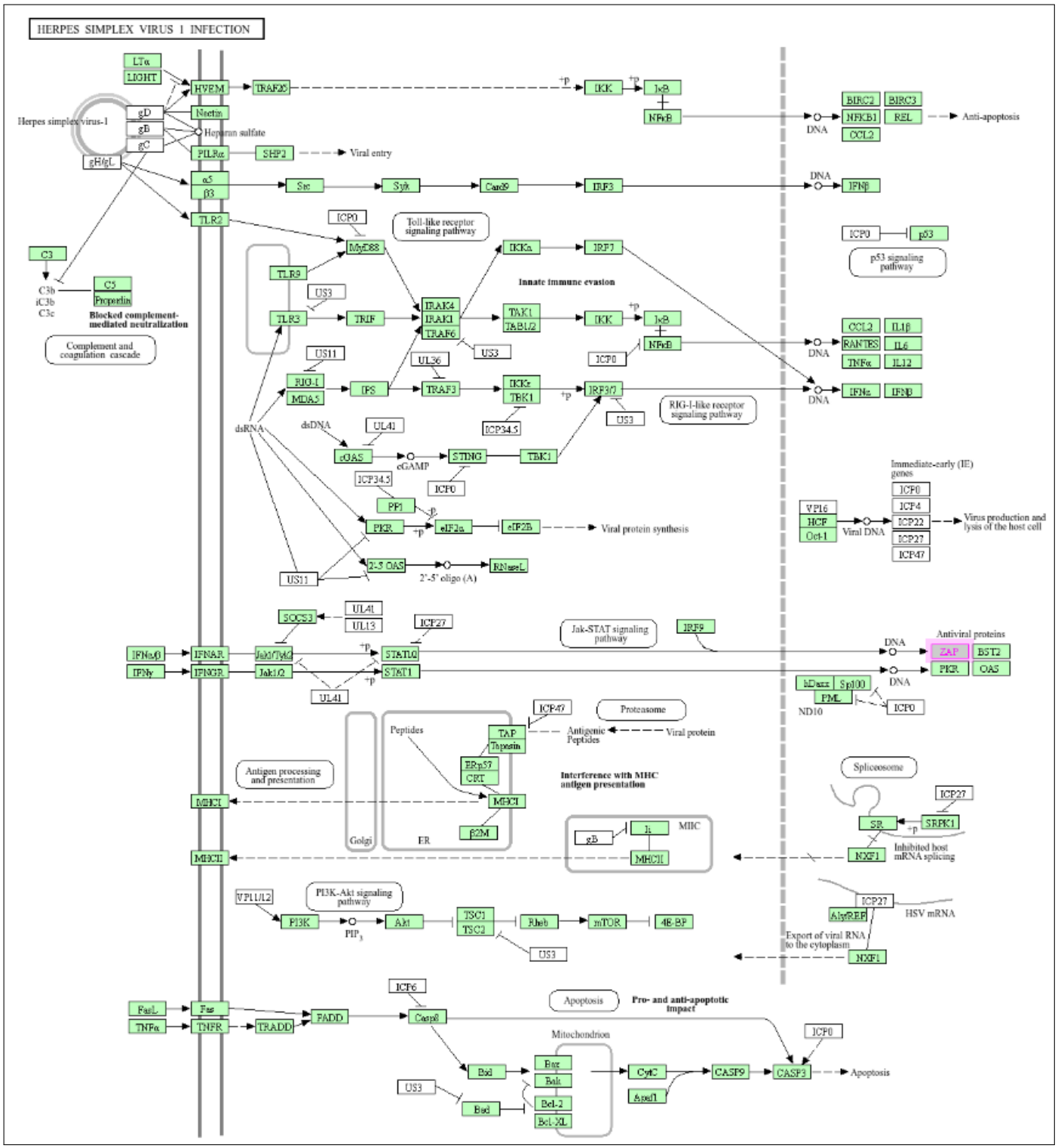

Fig. S24. KEGG pathway analysis of the "Herpes Simplex Virus 1 Infection" pathway (KEGG:05168). The ZNF genes mentioned were linked to the antiviral protein ZAP which is highlighted in pink.