Using Genomic Analysis in Human Herpesvirus-6A and 6B to Determine Potential Glycoprotein-Based Differentiation

Joey Geddie and Carson Song

Trinity High School

Louisville, KY, US.

Table of Contents

Abstract

Human Herpesvirus 6A and 6B (HHV 6A and 6B) are two double-stranded RNA viruses in the Roseolovirus genus that infect a majority of the human population in infancy. The two viruses have very similar genetic makeup, sharing over 90% similarity, but have very different pathogenic qualities, including varying symptoms on both primary and secondary infection. To investigate this pathogenic disconnect, statistical analysis was performed on glycoprotein genes L, H, B, and Q1/Q2, which were acquired through the Integrated Microbial Genomes at Joint Genome Institute (IMG.JGI). All of the glycoprotein genes appear in both HHV 6A and 6B, and statistical analysis was performed on their counterpart gene in the other virus, all of which had different names. Two functions were performed through IMG.JGI in order to determine if these selected glycoproteins were potentially behind the pathogenic disconnect between the two viruses. BLAST was performed to find percent difference between genes and DotPlot was performed in order to identify potential visual mutations. Each function produced desired results without errors and allowed for conclusions to be made. Glycoproteins B, H, and L all had percent differences <5% and showed no apparent visual mutations, while glycoproteins Q1/Q2 had percent difference >25% and had visual mutations resembling a spliced gene. These results were in line with previous works, which leads to the suggestion that glycoprotein(s) Q1/Q2's protein synthesis and in vitro role in the viruses be investigated in order to further solidify their role in HHV 6A and 6B's pathogenesis.

Introduction:

The Human Herpesvirus is a genus which consists of eight different double-stranded DNA viruses that primarily infect humans. Of these eight, several are famous for their

pathogenic qualities and the ways they infect humans. Human Herpesvirus 3 and 2, more commonly known as chickenpox and genital herpes respectively, are well known throughout American culture. Human Herpesvirus 6 (HHV 6), The virus being covered in this experiment, however, has two subtypes which present interesting situations for study. The subtypes (creatively named Human Herpesvirus 6A and 6B) are 90% genetically similar, but present radically different pathogenic (disease-causing) qualities and symptoms which researchers have been unable to definitively list, much less explain (Kasolo). The two viruses' genomes should be of primary focus when trying to determine the cause(s) for pathogenic disconnect between them because their genomes are so similar. This pathogenic disconnect might be behind the links between these viruses and Multiple Sclerosis (Tyler). Therefore, their study must be performed in order to potentially help stop the horrific disease of MS.

The goal of this paper is to thoroughly explore said pathogenic disconnect. To do so, statistical analysis will be performed through the Integrated Microbial Genomes database (IMG.JGI). Two separate analytical functions will be performed on appropriately selected genes to determine if the selected genes are responsible for the two viruses pathogenic disconnect. These selected genes are the glycoproteins B, H, L, and Q1/Q2 complex. Initially, we desired to evaluate Q1 and Q2 separately, but the two genes are stored together in the IMG database as the U100 gene, so they had to be evaluated as one gene. This is not a problem, however, because many other papers evaluate them as one gene as well (Dominguez). All these proteins are vital for reproduction for the viruses as glycoproteins are the medium by which viruses gain entry into cells. The significant pathogenic disconnect between HHV 6A and 6B has a chance to be caused by these genes and therefore their study needs to be performed in order to help potentially stop the ongoing scourge of MS.

Research Question and Hypothesis

The research question being pursued is which genes cause the large variation in HHV-6A and HHV-6B symptoms? The hypothesis is if HHV-6A and HHV-6B have different symptoms, the glycoproteins associated with membrane fusion are responsible for the difference because they are directly responsible for which cells are infected.

Data

The data utilized is from the publicly available JGI IMG/MER database. The glycoproteins analyzed in our research include HHV-6A and HHV-6B glycoproteins B, L, H, and Q1/Q2. Percent differences found between HHV-6A and HHV-6B glycoproteins B, L, H, and Q1/Q2 are summarized by Table 1. All glycoproteins appear in both HHV-6A and HHV-6B as summarized by Table 2. Both HHV-6A and HHV-6B genomes were sequenced through whole genome sequencing.

Method

Isolating and studying the genome of HHV 6A and 6B is extremely difficult. To do so requires years of biochemistry expertise and equipment that can cost thousands of dollars. With this in mind, determining if these genes' differences are consequential was accomplished by statistical analysis through the integrated microbial genomes system (IMG.JGI). No other method allows for greater depth of research and more certainty in relation to results than statistical analysis, because statistical analysis allows the researcher to analyze the subject data by using functions and statistic data. Additionally, statistical analysis is easily crafted to become replicable to a degree that anyone can perform the method, while also allowing the researchers to explore deeply without increasing opportunities for errors.

To perform content analysis, the international genetics database IMG.JGI was chosen because of ease of use, depth of information, reliability, and its status as the main database in the genetics field. According to Victor Markowitz et al., IMG.JGI is a useful tool in terms of the usability by the public, saying how "Comparative analysis of genomes is provided in IMG through a number of tools that allow genomes to be compared in terms of organism-specific statistics, genes and sequence conservation." All research was done with a login acquired through a previous project, but all functions and steps performed in this method are possible with a guest account. The functions chosen to accomplish the research goal were the DotPlot function and BLAST function. DotPlot is a statistical function that directly compares two genomes and generates an interactive figure that displays regions of the genome, which when interacted with shows a gene-by-gene map of the genome within the region; this function was chosen because of its ability to navigate the genomes of HHV 6A and 6B and search for large scale mutations such as mass deletion, insertion, or translocation. No other functions have a greater ability to qualitatively measure genomic variations than DotPlot. When performing DotPlot, select the basic settings before operating. This includes using 1 as reference, and nucleotide sequence-based comparison. This should be done because of the ease of use associated with default settings. When the selected genes or their mutated counterparts are located, open the links in a separate tab, and add the gene to the gene cart. BLAST function is a statistical function that directly compares two or more genes (up to 800) and produces the percent difference and shows where the genes are different by highlighting their FASTA codes. BLAST was chosen because of its ability to produce verifiable, quantitative statistics show relatedness for genes that can be applied to a broad extent. No other software was easier to apply and more

reliable under the circumstances. Any other software that was able to perform this type of analysis either costed thousands of dollars or was harder to access and operate.

These functions were be performed on the genes for glycoproteins L, H, B, and Q1/Q2, which are protein coding genes present in both HHV 6A and 6B (albeit in differing names) and related to virus replication and cell entry. Their HHV 6A versions were used as the basis and chosen as such because of their consistency in name, making the method significantly easier to execute. HHV 6A and 6B have over one hundred protein coding genes each, but the vast majority of these genes are putative (meaning their purpose has not been confirmed), hypothetical (meaning their purpose is wholly unknown) or have no information whatsoever, with their links on the IMG.JGI database leading to blank webpages. The glycoproteins L, H, B, and Q1/Q2 genes were chosen because of their commonality in both viruses, substantial available information, and vitality to HHV 6's reproduction. Data size is not an issue, since the presence or absence of even one gene can change the function of a virus. All three of these genes are related to how the two viruses replicate, so studying their differences in the two viruses is aligned with the research goal of uncovering the reason(s) why HHV 6A and 6B are so pathogenically different.

To begin the method, The first steps were to add the two viruses to the genome cart in IMG.JGI. The next step from there was to locate the selected genes from the protein-coding genes menu for HHV 6A on IMG.JGI. This was done because of consistent naming organization in the HHV 6A protein coding menu. From that step, the FASTA codes were copied and pasted into the BLAST menu on IMG.JGI one gene at a time. The genomes for the two viruses must be selected and added to the genome cart for BLAST to run. For this method, default settings were selected because of ease of use, consistency, and applicability (default settings are able to be

applied more broadly than other settings and thus helps replicability of method). After running, BLAST displays genes in the selected genomes most similar. Percentage differences for each gene BLAST identifies as similar were displayed. This was how gene counterparts were determined, as a gene's sequence data is far more important than a gene's geography within the genome (Caserta et al.).

The next step was to use the DotPlot function. HHV 6A and 6B were selected in the DotPlot menu and the function was performed. Each of the three selected genes were identified and apparent mutations documented in the codebook. Mutations were defined as any dissimilarity in the exact gene's corresponding coordinates in the other virus. The DotPlot functions' images made identifying the differences immediately noticeable and not difficult.

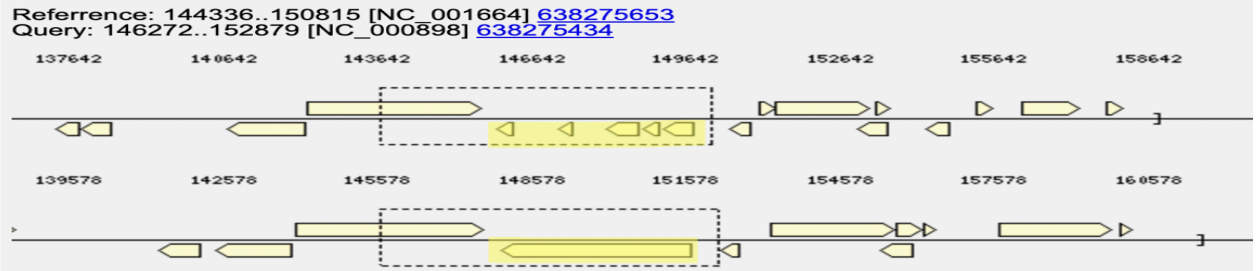
One scholarly study that supports this avenue of research is "Limits and Patterns of Cytomegalovirus Genomic Diversity in Humans," by Nicholas Renzette et al. This paper, published in *Proceedings of the National Academy of Sciences of the United States of America,* performs in depth statistical analysis over genomic patterns in Cytomegalovirus (HCMV), another herpesvirus targeting humans. In this paper, Renzette et al. goes through a more complex process of statistical analysis, doing so by generating primary genetic data from samples of HCMV and organizing the data using an extremely complex biochemical software called adegenetR (Renzette et al.). The latter half of the paper, however, is similar to this one, which is solely running functions on genetic data in order to determine relatedness and causality. Renzette et al. even used a DotPlot function as well, even though it was a more dated version (Renzette et al.). Besides the differences in software, primary data generation, and virus selection (which is relatively, though not wholly, negligible according to Ursula Gompels et al., who wrote how HCMV and HHV6 have similar protein-coding gene organization), this paper is

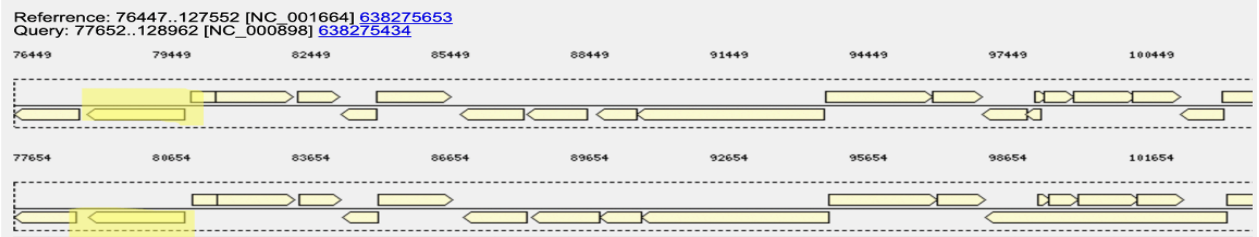
no different in terms of goals and alignments. This paper has the ability to put forth proper results and conclusions to be used by the community.

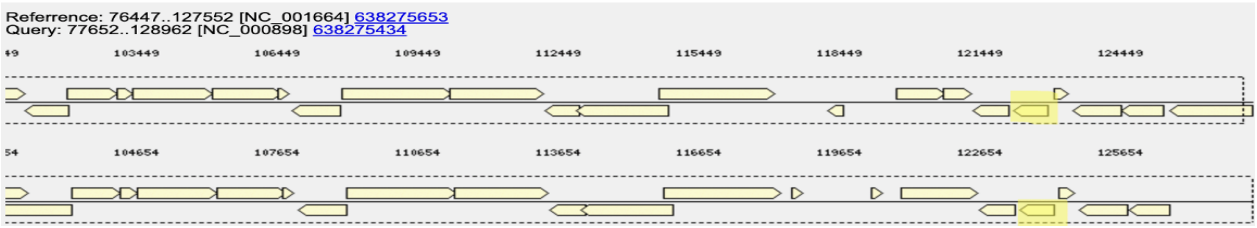
Respecting and honoring the codes of research in this work was one of the most important goals behind this research. Ethics were not of a great concern though. Because human subjects were not used in any capacity, there were no worries over protecting their rights and wellbeing. Plagiarism was also avoided at every step, and this was ensured as we maintained proper citations and accreditation each source we used. The main places in my research where the allure of breaking the code of ethics was strongest were the falsification and fabrication of data. Because the data sets for this statistical analysis were so foreign and complex, it could have been really easy to falsify data to make the research process easier. This was prevented, however, by documenting each step in my research process by executing my method and by using graphs and images to serve as "checks and balances."

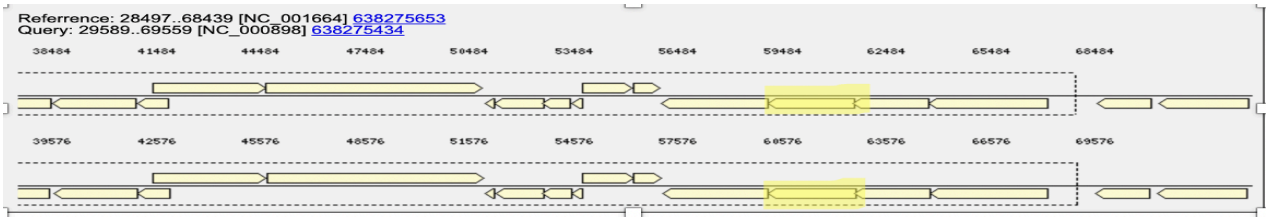
Results

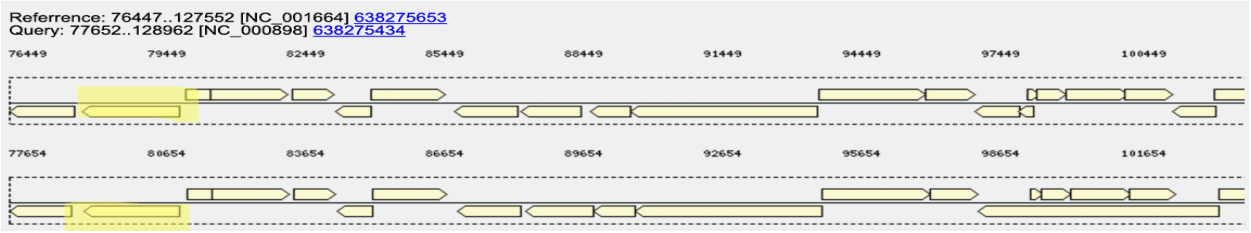
The two statistical functions provided ample evidence to make judgements on the genes' relation. The DotPlot function did not show any visual mutations in glycoproteins L, H, and B. Glycoprotein Q1/Q2, however, displayed a very strange set of results. There appear a set of five spliced genes in HHV 6A that correspond to the "spliced envelope glycoprotein" in HHV 6B. A spliced gene is when one gene can encode multiple gene products thanks to the combining of several genes. This evidence suggests that the U100 or Glycoprotein Q1/Q2 is spliced together with some of its other gene neighbors in HHV 6A into one gene in HHV 6B. The data generated by the DotPlot function is also supported by the results of BLAST.











Pictured: Gene neighborhood for each selected gene, with the selected gene and its counterpart(s) highlighted in yellow. From top to bottom: Glycoprotein Q1/Q2, L, H, and B.

The BLAST results fell in line nearly exactly with the DotPlot results. Glycoproteins B, H, and L all displayed single digit levels of percent difference with their counterparts in the other virus. Additionally, their counterpart genes indicated by DotPlot were similarly displayed by BLAST results. Glycoprotein Q1/Q2, however, had a very significant percent difference that indicated a large amount of mutation. What piqued interest, however, was the difference subject sequence length between Q1/Q2 and its counterpart. U100, in HHV 6A, had a subject length of 189 while its HHV 6B counterpart (spliced envelope glycoprotein) had a subject length of 616. This supports the results from DotPlot, which indicated that several gene neighbors of U100 in HHV 6A are spliced into one gene in HHV6B.

Glycoprotein	Percent Difference
GlyP B	4%
GlyP H	6%
GlyP L	6%
GlyP Q1/Q2	26%

Gene Name	Genome ID	Genome Name	Query Start Coord	Query End Coord	Subject Start Coord	Subject End Coord	Bit Score	E-value	Identities	Subject Length
U100, glycoprotein gp82/105	638276214	Human herpesvirus 6A	1	189	1	189	387	4e-142	189/189 100%	189
spliced envelope glycoprotein	638276215	Human herpesvirus 6B	61	189	1	129	201	4e-64	95/129 74%	616

Pictured: Top: Matrix displaying percent difference for each glycoprotein selected. Bottom: BLAST search results displaying percent difference and subject length difference between U100 and spliced envelope glycoprotein.

Results A: Comparing with Previous Attempts

These sets of results are more definitive compared to this team's previous attempts. Previous attempts were muddied because of misconceptions related to the glycoprotein gQ1/Q2 complex. On the first trial, it was believed that both genes would be logged separately inside the

genome database. When results did not fall in line with this misconception, there was simply an asterisk placed and due to time constraints, not investigated further. In actuality, the genes once previous results are aligned with their proper gene products, however, the results compare very well and paint consistent levels of relation.

Results B: Comparing with Published Works

The first action taken after processing data results was to do research on results and see how they fell in line with the knowledge put forth by researchers in the past. Dominguez et al. found in their paper, "Human Herpesvirus 6B Genome Sequence: Coding Content and Comparison with Human Herpesvirus 6A," that the U100 gene in 6B was in fact a spliced gene containing multiple genes that were expressed. Their paper, published in the American Virology Journal, also found that U100 had one of the highest amounts of difference out of all the genes in the viruses. Additionally, BLAST results from Dominguez et al. and this paper match exactly in glycoproteins L, H, and B. Gompels et al. also said in their paper published in Virology how they found that U100 in HHV 6B is a spliced gene containing multiple glycoproteins. Isegawa et al. also echoed our BLAST findings. In their paper, "Comparison of the Complete DNA Sequences of Human Herpesvirus 6 Variants A and B," Isegawa et al. found exactly the same BLAST results as this paper, citing U100 has having similarity between 70% and 80% and arrived at the conclusion that the B, L, and H glycoproteins are homologous while Q1/Q2 are different and at least partially responsible for the pathogenic disconnect.

Discussion

In this research, it has been found that when comparing HHV-6A and HHV-6B glycoproteins B, L, and H, high levels of similarity are found as opposed to large levels of

variation in regard to glycoprotein Q1/Q2. It has also been shown that the methods of BLAST and DotPlot served in obtaining accurate results in line with other findings.

It has been shown, through the findings in this research regarding glycoproteins B, L, and H, that the percentage differences found during analysis fall exactly in line with results from Dominguez et al. Furthermore, in discussion of glycoprotein Q1/Q2, findings presented in this research are in conjunction with findings presented by Isegawa et al. and figures similar in Ablashi et al. It has also been shown that because of the large variation in percent difference between HHV-6A and HHV-6B glycoprotein Q1/Q2, it is likely glycoprotein Q1/Q2 is at least partially responsible for the pathogenic disconnect observed between HHV-6A and HHV-6B, again falling in line with conclusions reached by

The findings on glycoprotein Q1/Q2 percent differences do exactly align with other findings but are within ranges set or by very marginal percentages. One explanation is that the organization or classification of which genes are included in the glycoprotein Q1/Q2 set may be different due to little current research into glycoprotein Q1/Q2 leading to variation in data content labeled as glycoprotein Q1/Q2 hence explaining the percent differences see from the findings from this research when compared to Isegawa et al. and figures similar in Ablashi et al.

Practical Applications

This research can be used to confirm similar analysis of HHV-6A and HHV-6B glycoproteins. Another application of the research is the identification of glycoprotein Q1/Q2 as a potential gene of interest in determining the causes for HHV-6A and HHV-6B pathogenic differentiation for further analysis of experimentation.

Conclusion

In this project, it has been shown that the gene difference between glycoproteins B, L, and H are minimal and the DotPlot functions support the indication of high levels of similarity between HHV-6A and HHV-6B in regard to the aforementioned glycoproteins. This leads to the findings that glycoproteins B, L, and H are not likely to act in the wide pathogenic disconnect between HHV-6A and HHV-6B. It has been found that large differences in gene similarity occur between HHV-6A and HHV-6B glycoprotein Q1/Q2 potentially caused by the combination of similar gene neighbors. Hence it is concluded through the finding that glycoprotein Q1/Q2 likely bears at least a partial responsibility for the pathogenic disconnect observed between HHV-6A and HHV-6B. Such conclusion is further supported by similar findings in Ablashi et al. in regard to glycoprotein Q1/Q2 as a potential source for pathogenic disconnect.

Future Directions

For future work, it may prove worthwhile to analyze the gene contents for HHV-6A and HHV-6B glycoprotein Q1/Q2 to determine key differential points in the genes for further understanding of causes for pathogenic differentiation between the two HHV-6 strands. Specifically, analyzing the gene contents to determine what was added to allow HHV-6B to have a longer subject length than HHV-6A in relation to glycoprotein Q1/Q2.

Acknowledgments

We would like to express a special thanks of gratitude to Mr. Heintz who gave us the opportunity to learn about the genome data banks through Seeds of Change. We would also like to thank Seeds of Change for creating a wonderful environment to learn and conduct research on this project. A special thanks to Ibrahim Zuniga and Dr. Gabriel Vargas, who were instrumental in assisting with technical difficulties experienced during the process. Finally, we would like to thank our parents who encouraged us throughout the project's time frame.

Bibliography

Ablashi, Dharam V., et al. “Classification of HHV-6A and HHV-6B as Distinct Viruses.” *Archives of Virology*, vol. 159, no. 5, Springer Science+Business Media, Nov. 2013, pp. 863–70. https://doi.org/10.1007/s00705-013-1902-5.

Caserta, Mary T., et al. “Human Herpesvirus 6.” *Clinical Infectious Diseases*, vol. 33, no. 6, 2001, pp. 829–33. *JSTOR*, http://www.jstor.org/stable/4461706. Accessed 1 Nov. 2023.

Dominguez, Geraldina, et al. “Human Herpesvirus 6B Genome Sequence: Coding Content and Comparison With Human Herpesvirus 6A.” *Journal of Virology*, vol. 73, no. 10, Oct. 1999, pp. 8040–52. https://doi.org/10.1128/jvi.73.10.8040-8052.1999.

Gompels, Ursula A., et al. “The DNA Sequence of Human Herpesvirus-6: Structure, Coding Content, and Genome Evolution.” *Virology*, vol. 209, no. 1, May 1995, pp. 29–51. https://doi.org/10.1006/viro.1995.1228.

Hajim, Yosor Abdul-Yema, et al. “Association of HHV134 Infection with Neuroinflammatory Diseases Patients in Najaf/Iraq: (Cross Sectional Study_.” *HIV Nursing*, vol. 23, no. 1, Jan. 2023, pp. 142–49. *EBSCOhost*, https://doi.org/10.31838/hiv23.01.24

Isegawa, Yuji, et al. “Comparison of the Complete DNA Sequences of Human Herpesvirus 6 Variants A and B.” *Journal of Virology*, vol. 73, no. 10, Oct. 1999, pp. 8053–63. https://doi.org/10.1128/jvi.73.10.8053-8063.1999.

Kasolo, Francis, et al. “Infection With AIDS-related Herpesviruses in Human Immunodeficiency Virus-negative Infants and Endemic Childhood Kaposi’s Sarcoma in Africa.” *Journal of*

General Virology, vol. 78, no. 4, Apr. 1997, pp. 847–55. https://doi.org/10.1099/0022-1317-78-4-847.

Markowitz, Victor M et al. "The integrated microbial genomes (IMG) system." *Nucleic acids research* vol. 34,Database issue (2006): D344-8. doi:10.1093/nar/gkj024

Mori, Yasuko. "Recent Topics Related to Human Herpesvirus 6 Cell Tropism." *Cellular Microbiology*, vol. 11, no. 7, Wiley-Blackwell, July 2009, pp. 1001–06. https://doi.org/10.1111/j.1462-5822.2009.01312.x.

Pantry, Shara N., and Peter G. Medveczky. "Latency, Integration, and Reactivation of Human Herpesvirus-6." *Viruses*, vol. 9, no. 7, Multidisciplinary Digital Publishing Institute, July 2017, p. 194. https://doi.org/10.3390/v9070194.

Pedersen, Simon Metz Mariendal, and Per Höllsberg. "Complexities in Human herpesvirus-6A And -6B Binding to Host Cells." *Virology*, vol. 356, no. 1–2, Elsevier BV, Dec. 2006, pp. 1–3. https://doi.org/10.1016/j.virol.2006.07.028.

Renzette, Nicholas, et al. "Limits and Patterns of Cytomegalovirus Genomic Diversity in Humans." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 30, National Academy of Sciences, July 2015, https://doi.org/10.1073/pnas.1501880112.im

Tang, Huiping, et al. "CD134 Is a Cellular Receptor Specific for Human herpesvirus-6B Entry." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 22, National Academy of Sciences, May 2013, pp. 9096–99. https://doi.org/10.1073/pnas.1305187110.

Tyler, Kenneth L. “Human Herpesvirus 6 and Multiple Sclerosis: The Continuing Conundrum.” *The Journal of Infectious Diseases*, vol. 187, no. 9, 2003, pp. 1360–64. *JSTOR*, http://www.jstor.org/stable/30085410. Accessed 30 Oct. 2023.