special thanks is awared to the Miami University’s CBFG and Virgina tech for providing the FASTA and LALIGN software

A quantitative analysis of the FIV and HIV genome using bioinformatics software

Meredith Horn, Charlotte Siu, and Xiao-Wen Cheng

Department of Microbiology, Miami University, Oxford, OH

abstract

Among viruses, the human immunodeficiency virus (HIV) presented the greatest challenge to human kinds. the HIV and FIV gag genome was sequenced using the Illumina MiSeq Benchtop next-generation sequencer.The DNA sequences obtained were then run through the LALIGN bioinformatics software to compute the E value, bit score, waterman eggert score, percent identity,which are four important indicators of how similar the sequences are. The E value was 3.1×10^{-9} , the percent identity was 54.4 percent and the bit score was 51.9. It was also sensed that base 1600 to 1990 in HIV and base 800 to 910 in FIV have a higher than normal similarity. This reflects that while the DNA sequences of the gag region of both the HIV and FIV genome are rather similar and it is unlikely that this similarity is due to random chance, there are a noticeable amount of differences. A better understanding of the level of similarity and differences in the gag region of the genome sequence would facilitate our understanding of structural and cellular behavioral differences between FIV and HIV, and in the long term it prevides new explanations to differences observed in previous studies, or even facilitate the development of an effective HIV treatment.

Results

The E value of the overall sequences is 3.1×10^{-9} , and a bit score of 51.9. The percent identity of the overall sequences is 54.4 percent (table 2).

Introduction

First discovered in 1983, HIV has infected about 80 million people worldwide so far according to data published by the WHO. In 2020, over 680,000 people died from the virus. Despite its deadliness, there is still no effective and specific treatment for this virus. Three years after the HIV was first sequenced, scientists were able to discover a similar virus in cats which is feline immunodeficiency virus or FIV (Pederson et al,1987). Like many viruses, both HIV and FIV have gag proteins (Coffin et al, 1997). The gag protein is known to play an important role in many stages of the replication cycle of a retrovirus. For example, They play an important role in viral assembly, interact with numerous host cell proteins, regulation of viral gene expression, provide the main driving force for virus intracellular trafficking and budding, and are involved with the pathogenicity (Mullers, 2013). Past studies have suggested that the DNA sequences in both viruses are similar, but it was not clear to what extent are the gag genomes similar. When computed in a DNA analysis and alignment software FASTA, The E value and bit score are good indications of similarity between two sequences. The lower the E value, the more similar the DNA sequences are and the less likely this “match” in DNA sequence is due to a random chance. Generally, an E value below 0.01 is considered as low. A bit score of 50 or above almost always indicates that the match between two DNA sequences is very significant and similar (Pearson, 2013). The percent identity is the percent of nucleotides which match exactly and is adjusted for the length of the DNA sequence. A percent identity of 50 percent or more would mean that a majority of the nucleotides match when adjusted for the length of the DNA sequences. Since previous studies have shown that the HIV and FIV share similar pathogenesis and the gag protein coded by the gag genome plays an important role in pathogenesis (Friedman et al, 2006), it is hypothesized that the gag region of the FIV and HIV genome should be similar in sequence as defined by the aforementioned standards regarding the E value, Bit score and percent identity.

Conclusion

The results of the study support our hypothesis. The gag genome of HIV and FIV, on a whole, are very similar. While they are generally similar, about 46.6 percent of the nucleotide bases do not exactly match.

In addition to calculating the E value, bit score and percent identity,the software alerted us that there is a region in the HIV/FIV genome that special attention is needed, since it shows a striking degree of similarity that can be said as one of the highest in the overall gag genome. It should be somewhere between the base pair 800 to 1200 in FIV and 1500 to 2500 in HIV(Figure 1).

With the software's alert in hand, a FASTA analysis was done. Unlike LALIGN, the FASTA results show a more specific picture of where this “region of concern” is. It turns out the “region of concern” is base 1600 to 1990 in HIV, base 800 to 910 in FIV(Figure 2).The “region of concern” shows a 3.4×10^{-10} E value, and a 55.1 bit score(Figure 2). Both of these are higher than the overall gag genome.

Since DNA, RNA and protein are important parts of the central dogma, the answer to why these structural and cellular behavioral differences shown in gag proteins could lie within the gag genome itself.Building to what this experiment shows subsequent studies could try to identify regions of the two gag genomes that show more differences than usual and what do these regions code for, is there anything that the aforementioned region codes for that we previously did not know. Previous studies have concluded that the FIV Gag is a nuclear shuttling protein that utilizes the CRM1 nuclear export pathway, while HIV-1 Gag is excluded from the nucleus(Kemler et al, 2012). The FIV Matrix is not found as a trimer in the crystal. existing research shows that there is little similarity in the amino acid sequence of the FIV and HIV protein matrix.However,there is evidence that the FIV virus's capsid protein retains the same fundamental structure of the HIV protein(Gonzalez et al, 2018).Since DNA, RNA and protein are important parts of the central dogma, the answer to why these structural and cellular behavioral differences could lie within the DNA itself.Conducting further studies to investigate the aforementioned questions or on the gag genome could give new explanations to why are there such similarities and differences between the HIV and FIV gag as mentioned above, or even help us understand more about how the gag protein works in ways we have never before.As the gag protein is highly involved in virus replication, understanding more about gag could help us develop an effective HIV treatment or prevention in long term.

methods

The DNA sequencing machine used is called Illumina MiSeq Benchtop next-generation sequencer, which follows the next -generation DNA sequencing method. Before the DNA of HIV and FIV could be sequenced, the DNA samples were needed to be amplified by PCR. This could increase the number of copies of the same DNA available for the sequencer to detect(Casbon et al, 2011).

In this experiment, the master mix of the PCR consists of 1uL of big dye terminator(table 1), 1.5uL of big dye dilution buffer (table 1), 0.5 uL of the primer (table 1), 4.5 uL of the gag DNA(table 1), 5.5 uL of molecular grade water (table 1). This adds up to a total volume of 10uL. The same recipe was used for both the HIV and FIV gag.After preparing the master mix, we programmed the standard cycle sequencing protocol on the thermocycler. Step one of the cycle lasted for 1 minute at 96 degree celsius. Step 2 of the cycle lasted for 10 seconds at 96 degree celsius. Step 3 of the cycle lasted 5 seconds at 50 degree celsius. Step 4 of the cycle lasted 4 minutes at 60 degree celsius. At the end of step 4, the cycler would go back to step 2 and complete step 3 and 4 again. The cycler did this repeatedly for 24 hours. After completing the PCR run, the master mix was transferred to a 1.5mL eppendorf tube. 1uL of 1.5M NaOAC/ EDTA and 80 uL 95% ethanol were added. After that the master mix was centrifuged for 15 minutes at 12,000g. All supernatants were removed with a pipette, and 100uL of 70 % ethanol was added. The sample was stored at -20 degree celsius. After the purification and PCR process, the sample was handed over to a technician to run the next generation sequencing machine. After a couple of hours, the results were out. The gag genome sequences received were “nucleotide blasted” on NCBI. The results of the blast shows that the sequences were accurate because it shows a 100 percent match to the HIV and FIV gag genome sequences in their records. The HIV and FIV gag sequences obtained were entered into the FASTA software, a software designed to compare DNA sequences. The FASTA software computed the E value, bit score as well as the percent identity of the two sequences. The results were recorded and presented below

bibliography

- 1.World Health Organization. (n.d.). *Number of deaths due to HIV/AIDS*. World Health Organization. Retrieved January 4, 2022, from <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/number-of-deaths-due-to-hiv-aids#:~:text=Situation%20and%20trends%3A%20680%20000,related%20illnesses%20worldwide%20in%202020>
2. Pedersen, N. C., Ho, E. W., Brown, M. L., & Yamamoto, J. K. (1987). Isolation of a T-lymphotropic virus from domestic cats with an immunodeficiency-like syndrome. *Science*, 235 (4790), 790–793. <https://doi.org/10.1126/science.3643650>
3. Cold Spring Harbour. (1997). *Retroviruses*.
4. Müllers, E. (2013). The foamy virus gag proteins: What makes them different? *Viruses*, 5(4), 1023–1041. <https://doi.org/10.3390/v5041023>
5. Miller, C., Abdo, Z., Ericsson, A., Elder, J., & VandeWoude, S. (2018). Applications of the FIV model to study HIV pathogenesis. *Viruses*, 10(4), 206. <https://doi.org/10.3390/v10040206>
6. Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, 42(1). <https://doi.org/10.1002/0471250953.bi0301s42>
7. Sparger, E. E. (n.d.). FIV as a model for HIV: An overview. *In Vivo Models of HIV Disease and Control*, 149–237. https://doi.org/10.1007/0-387-25741-1_7
8. Casbon, J. A., Osborne, R. J., Brenner, S., & Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, 39(12). <https://doi.org/10.1093/nar/gkr217>
9. Kemler, I., Saenz, D., & Poeschla, E. (2012). Feline immunodeficiency virus gag is a nuclear shuttling protein. *Journal of Virology*, 86(16), 8402–8411. <https://doi.org/10.1128/jvi.00692-12>
10. González, S., & Affranchino, J. (2018). Properties and functions of feline immunodeficiency virus gag domains in Virion Assembly and budding. *Viruses*, 10(5), 261. <https://doi.org/10.3390/v10050261>

Figure 1- the software indicates that there is a specific region of the DNA in FIV and HIV that requires attention, since the E value is particularly low which means it is very similar and it is unlikely that it is by chance.

Figure 2- a FASTA analysis of the “region of concern”

Table 1-the ingredients used for preparing the PCR master mix

Ingredient	Quantity (uL)
Big dye terminator	1
Big dye dilution buffer	1.5
HIV/ FIV primer	0.5
HIV/ FIV gag DNA	4.5
Molecular grade water	5.5

Table 2- the results of the LALIGN run

indicator	Value
E value	3.1×10^{-9}
Bit score	51.9
Percent identity	54.4%