Final Project Report - Ebola Virus Genome Sequence Database

Mario Camacho, Janice Le, Arman Meysami

Viruses that come from the family *Filoviridae,* filoviruses*,* get their name from their appearance resembling long filaments and are capable of infecting fish, reptiles, and mammals. Of the three genera of filoviruses what is arguably the most well known is Ebolavirus which are a highly contagious pathogen to humans. [1,4,5] In humans, Ebola virus causes blood vessels to leak by preventing the liver from creating blood clotting proteins that would normally prevent this from happening, causing hemorrhagic fevers. This virus is very dangerous, with some strains having a mortality rate of 50-90%, making them a huge concern for everyone both in and outside of the areas where Ebola virus has previously devastated entire communities and a very important virus for epidemiologists and other experts to research and keep track of. [2,3] Fears of this virus heightened in 2014 which marked the start of an Ebola virus outbreak that lasted until 2016. This outbreak first started in West Africa but spread to countries as far away as the United States and Spain. It spread to over 10 countries between 2014 and 2016, claiming over 11,000 lives according to the World Health Organization. [6]

Our database contains four different Ebola virus genome sequences to aid researchers in understanding which parts of the genome sequence are important for the virus's function and survival. This can be studied by comparing the sequences and analyzing what parts of the sequence have mutated and which parts have remained conserved over the years. Our database contains a sequence that was submitted in September of the year 2000 by the University of Marburg's Institute of Virology [11]. It also includes a sequence from the University of California Santa Cruz Genomics Institute that was submitted in June of 2014, a few months after the Ebola outbreak had begun [12]. The third sequence in our database was submitted in December of 2015 by the J. Craig Venter Institute [13]. The fourth and final sequence we included in our database was submitted by the University of Justus-Liebig-University's Institute for Medical Microbiology towards the tail end of the outbreak in August of 2016 [14]. These sequences in particular were chosen because they can give context for why the virus was such a big problem during the 2014 outbreak. The names of the sequences are 2000EboVirSequence, 2014EboVirSequence, 2015EboVirSequence, 2016EboVirSequence in the database respectively.

Research into the Ebola virus genome sequence has revealed that the sequence contains several genes but in regards to understanding how the virus is so successful at infecting humans the gene for nucleoprotein (NP) was deemed to be particularly important. [1,7] The nucleoprotein has a number of functions but most important of which include its role in allowing the virus to form viral nucleocapsids and helping with viral transcription as well as replication. The nucleocapsids are key to the survival of the virus when inside a host organism, serving as a protein shell for the genome, protecting it from its surroundings and allowing the virus to survive and propagate. The nucleoprotein's role in aiding viral transcription and replication allows the virus to create the proteins it needs to survive and spread inside the host. In our database, the sequences:

2000EboVirSequence, 2015EboVirSequence, and 2016EboVirSequence each have an annotation for the location of the NP gene. Performing a linear synteny view within JBrowse reveals that all four have a start codon at position 470 and a stop codon starting at position 2687 in the same open reading frame. Furthermore, a multiple sequence alignment shows very few mutations between those positions, with only the occasional transversion or transition mutation throughout the sequences rather than many appearing in any particular positions. This all goes to show that the part of the genome sequence relating to the nucleoprotein is highly conserved. Given how important this gene is to the virus's ability to survive within its host and propagate, it is not surprising because mutations at these positions of the sequence could lead to the death of the virus. [8, 9,10, 11, 12, 13, 14]

A second key part of the sequence genome that stood out during our research is the gene for viral protein 35 (VP35). The VP35 protein works with other proteins to aid in transcription and replication, helping the virus spread across different cells once inside a host. More notably, it also helps suppress the host's immune system response by serving as an inhibitor protein, allowing the virus to survive long enough to continue infecting the host. Because of this fact, it should be expected that the part of the sequence containing this gene is very conserved, as VP35 needs to have a very specific structure in order to bind to other proteins to aid with transcription and replication as well as binding to certain binding sites in the cell that would normally help the cell detect Ebola virus as an intruder. When going into JBrowse and performing a linear synteny view, it can be seen that 2000EboVirSequence, 2015EboVirSequence, and 2016EboVirSequence all have an annotation for the VP35 gene with all four sequences having a start codon at 3129 and a stop codon starting at position 4149. Doing a multiple sequence alignment shows that just as with the NP gene, the sequence between these positions is highly conserved. This is very likely due to its importance, although it could also be due to how little time the virus had to mutate [8, 5, 11, 12, 13, 14]

- "Clear documentation of individual team members' contributions (one short sentence per team member)"
  - Mario: Uploaded/helped upload some of the sequences into Jbrowse and wrote the report.
  - Janice: Researched and found the genome assemblies and annotations to upload, did majority of uploading and formatting of the data in the browser and documented steps

    Arman: Procedure and github helped with some debugging with aws

Sources

1. Ebola virus disease
2. Family: Filoviridae | ICTV
3. Filovirus | Ebola, Marburg & Hemorrhagic Fever | Britannica
4. Ebola vs. Hemorrhagic Fever: What's the Difference? | Live Science
5. Interferon Type I - an overview | ScienceDirect Topics
6. Filovirus - PubMed

7. Ebola outbreak 2014-2016 - West Africa
8. Structural and Functional Aspects of Ebola Virus Proteins - PMC
9. Functional mapping of the nucleoprotein of Ebola virus - PubMed
10. The Nucleocapsid Protein of the SARS Coronavirus: Structure, Function and Therapeutic Potential - PMC
11. 2000EboVirSequence: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000848505.1/
12. 2014 EboVirSequence: http://hgdownload.soe.ucsc.edu/downloads.html#ebola_virus
13. 2015EboVirSequence: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_034098425.1/
14. 2016EboVirSequence: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_900094155.1/