Analysis of Ebolavirus Variants using VGToolkit

Final Project

Shweta Jones in collaboration with Sahasra Shankar
Biomolecular Engineering and Bioinformatics
University of California, Santa Cruz
shsujone@ucsc.edu

***Abstract*—With the end of the COVID-19 pandemic in the near future, the need for better understanding of past and current epidemic-related viruses has only increased. This paper compares Ebolavirus, one of the most fatal and severe viruses, using vg toolkit to identify what variations exist between its different variants. An understanding of where these variations occur can help for better preparation for future outbreaks of ebola.**

I. Introduction - The Ebola Virus

Ebola, or EVD, is an extremely deadly disease and there have been a total of 5 variants of the virus that have been identified, *Zaire, Sudan, Tai Forest, Reston,* and *Bundibugyo*. Among these variants, three of them - *Zaire, Sudan,* and *Bundibugyo* - have affected the human population. The *Reston variant* has so far only affected nonhuman primates and pigs. The virus has an average mortality rate of 40%, where there have been a total of 28,652 cases of EVD and 11,325 EVD-caused deaths.

The first outbreak of the virus was in the Democratic Republic of Congo in 1976. Since then, there have been a few outbreaks of the virus, predominantly within Africa, with either the same or similar variants of the Zaire variant. The origin of this virus is not exactly known, however, Ebola is thought to be animal-borne, specifically from bats and nonhuman primates. The virus initially originates from direct human contact with animal blood, tissue, or bodily tissue and is then spread among humans through contact with bodily fluids, such as blood.

In regard to symptoms, there are several including:

- Fever
- Aches
- Fatigue
- Gastrointestinal Symptoms
- Hemmoraging, Bleeding, or Bruising

In the later stages of the disease, the patient often struggles with internal bleeding which leads to the vomiting and coughing of blood. There are also long-term side effects to the disease including vision problems, muscle aches, and fatigue. With the proper clinical care and a strong immune system, recovery from EVD is possible even considering its high mortality rate.[1]

Biologically speaking, the virus falls under the genus of viruses in the family of Filoviridae and is classified as a filovirus. Filoviruses are "filamentous virion particles" that are "about 80 nanometres in diameter and are tubular". The "ebolavirus genome is approximately 19 kilobases in length".[13]

Fig. 1. This figure displays what the ebolavirus looks like [14]

II. The Variants

In the following portion, the variants of ebolavirus will be discussed in detail.

A. Zaire Ebolavirus

The *Zaire ebolavirus* has been responsible for a few different outbreaks. The first outbreak was in 1976 in the Democratic Republic of the Congo, and in 1995 there was another outbreak in the Republic of the Congo. Since these initial outbreaks, there have been various outbreaks around the western region of Africa and have affected more than just the human population. During periods of the outbreak, there were mortality rates as high as 80%.[3]

B. Sudan Ebolavirus

The first outbreak of the *Sudan ebolavirus* occurred in 1976 in Sudan and since then there have been several outbreaks near Sudan and Uganda. The mortality rate for this disease averaged around 53% and went to as high as 71%. [4]

C. Tai Forest Ebolavirus

This strain of ebolavirus, also known as *Ebola Côte d'Ivoire*, was first identified in chimpanzees, and "when they were found, blood within their hearts was brown while the blood would be completely frozen within a dozen hours after death"[5]. The host for this virus is still unknown, however, it is thought to be bats. There has only been one human case, of which they recovered fully. This was also the only human case

where the disease appeared outside of the Congo River Basin.[6]

D. Reston Ebolavirus

This strain was first identified in 1989 in cynomolgus monkey while being transported from the Philippines to Virginia. Following the transplant, there was an outbreak in several locations like Virginia, Texas, and the Philippines among monkeys. Since then, it has also been found in pigs as well, however, in both cases, humans have not caught the virus and have instead formed antibodies to the virus, so "it is concluded that Reston ebolavirus has a low pathogenicity in humans". [7]

E. Bundibugyo Ebolavirus

The strain is fairly new and first appeared in 2007 in the Bundibugyo District and had a mortality rate of around 34%. A second outbreak occurred in 2012 where it spread to neighboring regions, such as the Haut-Uélé district in Province Orientale and the Democratic Republic of Congo.[8]

III. Materials

F. Retrieving FASTA and Reference Genomes

FASTA files and reference genomes were necessary for the vgToolkit and were retrieved from the NCBI genomes dataset. The reference sequence was based on the *Zaire ebolavirus* which was downloaded from the NCBI database. The FASTA file for the variants - *Zaire, Sudan, Tai Forest, Reston,* and *Bundibugyo* - were also retrieved from the NCBI database.

G. Retrieving VCF Files

Another piece of data we needed for the toolkit was a VCF file based on the reference genome. The VCF file that was used was based on data gathered from "213 cases evaluated for Ebola virus infection at the Kenema Government Hospital in Sierra Leone between May 25 and June 18, 2014"[9].

H. Generating FASTQ Files

The last portion of the materials list is the FASTQ files of all the variants to draw comparisons and identify variants. There were no accessible FASTQ files, and so SRA files of the variants, *Zaire, Sudan, Tai Forest, Reston,* and *Bundibugyo*, were downloaded from NCBI. These SRA files were then converted into FASTQ files using an SRA toolkit, found on NCBI. The `fasterq-dump` command was used to generate these FASTQ files. However, the FASTQ files for *Zaire Ebolavirus* were publicly available on Dr. Stephen Piccolo's Figshare account.

I. Retrieving GFF Files

The GFF, or General Feature Format, files, to be used in the R processing, were retrieved from NCBI. GFF files are tab-delimited text files used to describe genomic features.

IV. Vg Toolkit - Method, Data, and Analysis

Variation graphs provide in-depth analysis of genetic variation as they help to construct "bidirected DNA sequence graphs that compactly represent genetic variation across a population". Variation graphs are made of:

- Nodes - labeled by sequences and ids
- Edges - connects nodes
- Paths - describes genomes, sequence alignments, and annotations

This is where vg toolkit comes into play as it provides read mapping, variant calling, and visualization tools.

J. Setting Up Vg Toolkit

The vg toolkit was set up on an M1 Macbook Pro through a Docker Desktop. In order to access the files generated by the toolkit, a local directory was mounted to a directory within the Docker. This allowed for the files on the local to be used by the tool and allowed the files generated to be accessible on the local computer.

K. Using Vg ToolKit

Vg toolkit uses 3 files:

- FASTA Reference Genomes
- VCF Files based on the Reference Genome
- FASTQ Files based on the Variants

These files are then used to create a variety of different files, but the one of particular interest was a VCF file. After setting up vg tools with a docker we followed the toolkit documentation. The steps we went through were:

1. Variations Graph Construction

   This builds a graph and aligns it using a reference FASTA file and VCF file.

2. Indexing the Graph

   This stores the graph in the xg/gcsa index pair using vg index.

3. Map Reads to Variation Graph

   Vg map is then used to map reads using maximal exact match algorithm and it is dependent on what the variant FASTQ files look like. In the case of this study, there were 2 FASTQ files with paired-end reads.

4. Embed Variation

   This step uses vg augment to embed variations from alignments back into the graph.

5. Variant Calling

   This computes the read-support from the .gam file using vg pack. In other words, you use the mapped variation graph to compute read support.

6. Snarl Variant Calling

   Vg call was then used to normalize the data to that it can be compared to other outputs. In this step, it calls the variants and snarls from read-support with the same coordinates and reference calls.

7. Merge the variant VCF files

   Used bcftools to merge the variant VCF files.

L. Vg Toolkit Results in IGV

The final output from vg toolkit was VCF files, which were then placed into IGV. IGV is an integrated Genome Browser that acts as a visualization tool that can identify genomic patterns given genomic data sets. These datasets can be "sequence data, gene models, alignments, and data from DNA microarrays"[12]. The reference genome was set to the *Zaire ebolavirus* FASTA file, and then compared the VCF files of the ebolavirus variants:

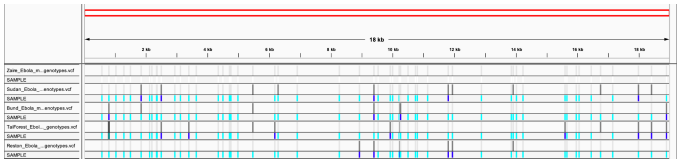

Fig. 2. This is the resulting VCF files displayed in IGV

And a side-by-side comparison looks like this:

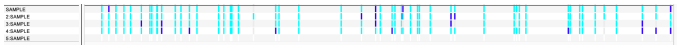

Fig. 3. This is the resulting VCF Files side-by-side in IGV

The graphs display the variations among all the variants, and after further analysis, a table can be developed to show how many differences there are between each of the variants.

	Sudan	Reston	Tai Forest	Bundibugyo
Sudan	----------	8	12	11
Reston	8	----------	16	7
Tai Forest	12	16	----------	11
Bundibugyo	11	7	11	----------

Fig. 4. This is the resulting table that displays the differences between the different VCF variants

These results show that in terms of genetic differences between the variants, the variant with the most variability in bases was the *Sudan Ebolavirus,* then *Reston Ebolavirus,* and finally the last two variants is *Bundibugyo Ebolavirus* and *Tai Forest Ebolavirus.*

M. Vg Toolkit Results in R

The results were also processed using R. For VCF processing, the R package, vcfR was used. This package contains tools that can "set, write, manipulate, and analyze VCF data. For visualization of the VCF data additional data is required:

- Sequence FASTA data
- GFF Annotation Data

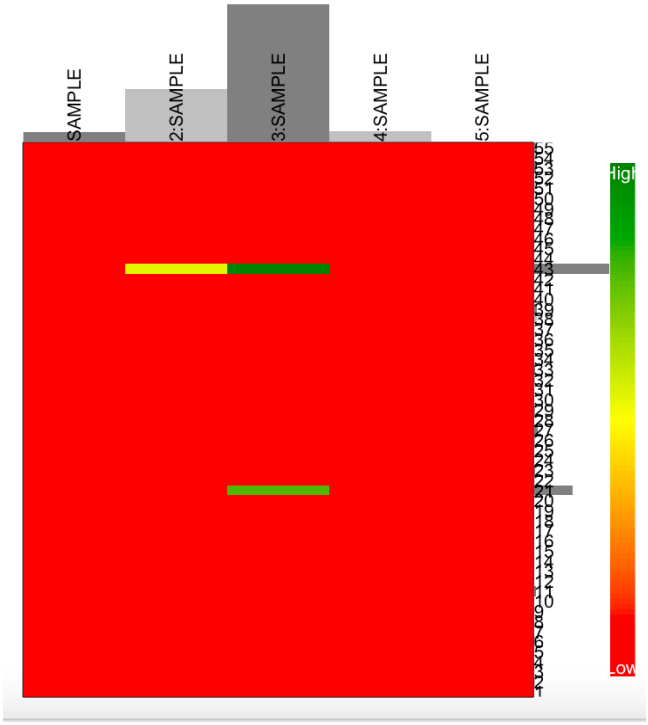

Fig. 5. The resulting plot from vcfR processing

In this diagram, there are 5 samples, each is a different variant. Sample 1 is the *Bundibugyo Ebolavirus,* sample 2 is the *Reston Ebolavirus,* sample 3 is the *Sudan Ebolavirus,* sample 4 is the *Tai Forest Ebolavirus,* and the last sample, sample 5, is the *Zaire Ebolavirus*.

A "high" score in the graph signifies a high sequence quality score which implies a higher probability of a base call being wrong. While a low score would mean a lower probability of a base call being correct since the sequence quality score is low. Looking at this graph, you can see that the *Sudan variant* has two "high" scores which signifies that there were two places where the bases were most likely incorrect. The *Reston variant* has one "high" score which shows that there is one place where the base is most likely incorrect. These conclusions align with the phylogenetic trees that have been constructed on ebolavirus variants. Genetically speaking, *Sudan* has the highest number of differences to the reference,

Zaire ebolavirus, and the second most different is the *Reston variant*.

N. Result Analysis

Based on the data gathered from vg toolkit, IGV, and the vcfR, it can be concluded that there are differences between the variants. Knowing and being able to identify these differences is important for future developments in creating vaccines or even helping to understand similar epidemic-related viruses.

V. Further Processing - CLUSTALW

To confirm our results we used CLUSTALW to create a multiple sequence alignment (MSA), and then used to create a phylogenetic tree using Jalview.

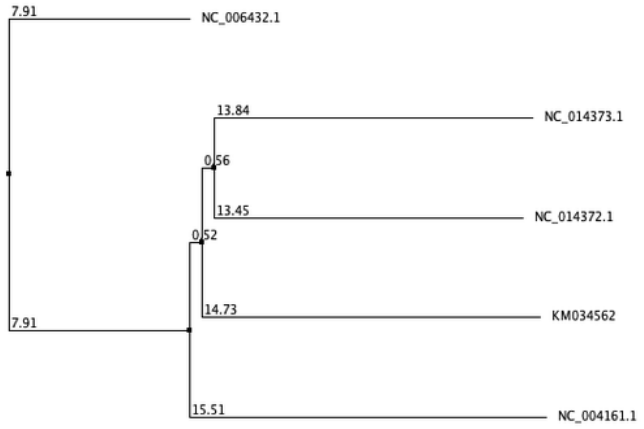

Fig. 6. This is the resulting phylogenetic tree. NC_006432.1 is *Sudan*, NC_014373.1 is *Bundibugyo*, NC_014372.1 is *Tai Forest*, KM034562 is *Zaire*, and NC_004161.1 is *Reston*.

The resulting phylogenetic tree lines exactly with the data gathered from the vg toolkit. *Sudan* had shown the greatest difference, and the *Bundibugyo* and *Tai Forest variant* was the closest to the *Zaire variant*.

Conclusion

The data that has been collected throughout the span of this project shows that these variants are quite different but also have some close relationships to each other. It proves that the use of tools like vg toolkit can help to create accurate phylogenetic trees. Having information like this can help with having a better understanding of EVD, how these variants form, which portions of the genome are the most prone to change, and confirm the evolution of different viral strains.

Information like this can help with the future development of vaccines and be more prepared to understand future variants of the ebolavirus. Especially considering how deadly the virus is, the importance of knowing this information and being able to visualize this information makes research with vg toolkit important. Recent studies have shown that the ebolavirus does have can have a long hibernation rate, and so there is a high possibility that future outbreaks of this virus will occur. While the mutation rate for ebola is not quite as high as the coronavirus or influenza virus, the fatality and severity of this virus prove the importance of further research into this virus and its variants.

Acknowledgment

We would like to thank the authors of all the publishing we referenced throughout the paper and their in-depth explanation of certain topics. we would also like to thank all those that created the vg toolkit since it allowed us to develop the final VCF variant files. In addition, we would like to thank Jorn Eizenger for helping us figure out how to use the vg toolkit and Prof. Benedict Paten for his continual assistance in the creation of this final paper and for helping us determine which tools to use for our paper.

References

[1] Centers for Disease Control and Prevention. (2021, April 27). What is ebola virus disease? Centers for Disease Control and Prevention. Retrieved March 16, 2022, from https://www.cdc.gov/vhf/ebola/about.html

[2] Authors Rachel Tanner, Authors Meirav Leibman-Markus, Authors Kirsten Maertens, & Authors Marie Neunez. (n.d.). *Vaccines and antibodies: Weapons in the fight against ebola virus*. Frontiers for Young Minds. Retrieved March 16, 2022, from https://kids.frontiersin.org/articles/10.3389/frym.2021.593713

[3] *Zaire ebolavirus - sino biological*. (n.d.). Retrieved March 16, 2022, from https://www.sinobiological.com/research/virus/zaire-ebolavirus

[4] *Sudan ebolavirus - sino biological*. (n.d.). Retrieved March 16, 2022, from https://www.sinobiological.com/research/virus/sudan-ebolavirus

[5] *Tai forest ebolavirus - Sino Biological*. (n.d.). Retrieved March 16, 2022, from https://www.sinobiological.com/research/virus/tai-forest-ebolavirus

[6] *Ebola Taï Forest: A unique emergence, and the dawn of the modern age of ebola*. O'Neill. (2021, July 19). Retrieved March 16, 2022, from https://oneill.law.georgetown.edu/ebola-tai-forest-a-unique-emergence-and-the-dawn-of-the-modern-age-of-ebola/

[7] *Reston ebolavirus - sinobiological.com*. (n.d.). Retrieved March 16, 2022, from https://www.sinobiological.com/research/virus/reston-ebolavirus

[8] *Bundibugyo ebolavirus - Sino biological*. (n.d.). Retrieved March 16, 2022, from https://www.sinobiological.com/research/virus/bundibugyo-ebolavirus

[9] Colubri, A. (2015, January 26). *Ebola: Dataset v1.4*. Zenodo. Retrieved March 16, 2022, from https://zenodo.org/record/14565#.YjIKnBBKhfV

[10] Vgteam. (n.d.). *Vgteam/VG: Tools for working with genome variation graphs*. GitHub. Retrieved March 16, 2022, from https://github.com/vgteam/vg

[11] Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018, October 1). *Variation Graph Toolkit improves read*

mapping by representing genetic variation in the reference. Nature News. Retrieved March 16, 2022, from https://www.nature.com/articles/nbt.4227#:~:text=Variation%20graphs%20are%20bidirected%20DNA,as%20inversions%20and%20duplications1

[12] *Home*. Home | Integrative Genomics Viewer. (n.d.). Retrieved March 16, 2022, from https://software.broadinstitute.org/software/igv/

[13] Encyclopædia Britannica, inc. (n.d.). *Ebolavirus*. Encyclopædia Britannica. Retrieved March 16, 2022, from https://www.britannica.com/science/ebolavirus

[14] *Ebola*. Diversey. (2021, November 29). Retrieved March 16, 2022, from https://www.solutionsdesignedforhealthcare.com/ebola/

[15] Visualizing VCF data 2. (n.d.). Retrieved March 16, 2022, from https://knausb.github.io/vcfR_documentation/visualization_2.html

[16] Piccolo, S. (2016, March 14). Ebolavirus genome sequencing data for one sample from 2014 outbreak in Zaire. figshare. Retrieved March 16, 2022, from https://figshare.com/articles/dataset/Ebolavirus_genome_sequencing_data_for_one_sample_from_2014_outbreak_in_Zaire/3114454